Intelligent Data Analysis in bioMedicine and Pharmacology

Workshop Notes

Lucia Sacchi, Tomaz Curk

AIME 2009, Verona, Italy
July 19, 2009

International Medical Informatics Association
Intelligent Data Analysis and Data Mining Workgroup

American Medical Informatics Association
Knowledge Discovery & Data Mining SIG

IDAMAP 2009 Intelligent Data Analysis in bioMedicine and Pharmacology

Artificial Intelligence in Medicine 2009, Verona, Italy

Lucia Sacchi[1], Tomaz Curk[2]

[1] Department of Computer Engineering and System Science, University of Pavia, Italy
lucia.sacchi@unipv.it

[2] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
tomaz.curk@fri.uni-lj.si

Welcome

Welcome to IDAMAP 2009, the fourteenth workshop on Intelligent Data Analysis in bioMedicine and Pharmacology, held this year in conjunction with the 12th Conference on Artificial Intelligence in Medicine in Verona, Italy.

IDAMAP is devoted to computational methods for data analysis in medicine, biology and pharmacology that somehow exploit knowledge of the problem domain at different stages of the data-analysis and model-building process and present results of analysis in the form communicable to domain experts. Typical methods include data visualization and exploration, machine learning, and data mining.

Gathering in an informal setting, researchers and practitioners have the opportunity to meet and discuss selected technical topics in an atmosphere which fosters an active exchange of ideas. The workshop is intended to be a genuinely interactive event, thus ample time is allotted for general discussion of papers, tools and problem descriptions. Problem description is a new category of workshop contribution that we introduced with the hope to draw new research ideas and applications.

Congratulations to Lan Umek, Minca Mramor and Blaz Zupan (supervisor) for winning the student challenge.

Program

This year's selection of eight long papers, four short papers, three tool presentations and two problem descriptions cover the following broad topics:

- Biological Data Mining and Data Interfaces
- Networks and Visualization
- Classification
- Machine Learning in Medical Applications

Acknowledgments and Program Committee

IDAMAP is organized in collaboration with Intelligent Data Analysis and Data Mining Working Group of International Medical Informatics Association, and Knowledge Discovery & Data Mining Working Group of American Medical Informatics Association. We thank AMIA KDDM Working Group for the financial support.

We would like to thank our invited speaker, Hendrik Blockeel, for presenting his interesting work on mining relational databases. We would like to acknowledge all authors that submitted their work to the workshop and thank those who will be presenting at the workshop.

Many thanks to John H. Holmes, Allan Tucker and Niels Peek for their helpful suggestions on the organization of the workshop.

We are grateful to the AIME'09 Organizing committee who greatly facilitated the organization of this workshop: Carlo Combi, Yuval Shahar, Barbara Oliboni and Roberto Posenato.

Finally, many thanks to our colleagues who served on the program committee:

- Ameen Abu-Hanna, Academic Medical Center, Amsterdam, The Netherlands
- Riccardo Bellazzi, University of Pavia, Italy
- Michael Berthold, University of Konstanz, Germany
- Carlo Combi, University of Verona, Italy
- Tomaz Curk, University of Ljubljana, Slovenia (chair)
- Janez Demsar, University of Ljubljana, Slovenia
- Dragan Gamberger, Rudjer Boskovic Institute, Croatia
- John H. Holmes, University of Pennsylvania School of Medicine, USA (chair)
- Matjaz Kukar, University of Ljubljana, Slovenia
- Nada Lavrac, J. Stefan Institute, Slovenia
- Xiaohui Liu, Brunel University, UK
- Oscar Luaces, Universidad de Oviedo at Gijon, Spain
- Marco Masseroli, Politecnico of Milan, Italy
- Silvia Miksch, Vienna University of Technology, Austria
- Lucila Ohno-Machado, Harvard Medical School and M.I.T., Boston, USA
- Niels Peek, Academic Medical Center, Amsterdam, The Netherlands

- François Portet, Laboratoire d'Informatique de Grenoble, France
- Mollie Poynton, College of Nursing, Salt Lake City, USA
- Jan Ramon, Katholieke Universiteit Leuven, Belgium
- Lucia Sacchi, University of Pavia, Italy (chair)
- Paola Sebastiani, Boston University School of Public Health, USA
- Yuval Shahar, Ben-Gurion University of the Negev, Israel
- Niel Smalheiser, UIC Psychiatric Institute, Chicago, USA
- Stephen Swift, Brunel University, UK
- Allan Tucker, Brunel University, UK
- Blaz Zupan, University of Ljubljana, Slovenia

IDAMAP 2009 Program
Sunday, July 19, 2009

Morning Session

Afternoon Session

Timing of presentations (presentation + discussion)

Long presentation (★★★): 20 + 5 minutes
Short presentation (★): 10 + 5 minutes
Tool demonstration (🖱): 8 + 2 minutes
Problem description (✎): 10 + 10 minutes

Mining relational databases

Hendrik Blockeel

Katholieke Universiteit Leuven, Department of Computer Science, Belgium
Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

Standard methods for data mining and statistical analysis typically assume data to be in a relatively simple "attribute-value" format. In practice, data are often stored in relational databases with a relatively complex structure: relevant information for an individual may be spread over multiple tables in the database, may consist of complex objects such as time series, etc. It is an open question how such diverse kinds of data can be integrated in an optimal way in the data mining process. In this talk I will discuss some of the challenges and how they can be addressed. The concepts will be illustrated through an application of data mining in intensive care.

Does replication groups scoring reduce false positive rate in SNP interaction discovery?

Marko Toplak[1], **Tomaž Curk**[1], **Janez Demšar**[1], **Blaž Zupan**[1,2]

(1) Faculty of Computer and Information Science, University of Ljubljana, Slovenia

(2) Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Abstract

Inference of single nucleotide polymorphism (SNP) interactions from experimental data can in theory help us to reveal biological mechanisms to non-Mendelian diseases. However, practical aspects of such analysis face many problems. SNP arrays typically record data on hundreds of thousands of SNPs, with only hundreds of samples being recorded in a typical experimental study. Computational interaction analysis would typically test over a billion hypotheses on a small number of samples, possibly leading to very high false positive rates. Recently, a group of authors proposed that instead of scoring interactions on entire data set, scoring them on subsets of data samples (so-called replication groups) and then considering the lowest obtained score would lead to reduced false positive rate. In this paper, we investigate if this is so. The results on synthetic and false interactions-imputed experimental data sets indicate that the use of replication groups does not reduce false positive rate, and in this respect behaves similar or worse to the standard interaction scoring that uses the entire data set directly.

1 Introduction

The onset of many common chronic diseases is governed by genetic factors that do not follow "Mendelian" or "single gene" patterns [Smith *et al.*, 2005]. Such diseases include hypertension, diabetes, various cancers, Alzheimer's disease, heart disease, and Parkinson's disease. Genetics governing the susceptibility of these diseases remains largely unknown. The onset of complex diseases may be triggered by polymorphisms across the genome whose effects do not simply (linearly) sum up but instead interact in complex, non-linear way. The field that studies such interplay of gene interactions is called epistasis analysis [Moore, 2005].

A number of computational and scoring methods that can facilitate the detection of gene-gene interactions from data on single nucleotide polymorphism (SNP) have recently been proposed. These methods operate on a set of typically several hundreds to several thousands of patients (with their diagnosis) and controls, each characterized by whole-genome SNP profile consisting of measurements of several hundred thousands SNPs. The classification problem is often binary, studying the susceptibility of a particular disease or stage of the disease versus the control group. Computational methods that can analyze such data and report on SNP interactions include Multifactor Dimensionality Reduction (MDR) [Hahn *et al.*, 2003], estimation of interactions by logistic regression modeling [Park and Hastie, 2008], and methods that stem from information theory and that measure the interaction gain [Jakulin and Bratko, 2003], also known as synergy [Anastassiou, 2007].

Finding interacting combinations of SNPs requires an evaluation of all candidate combinations, as – by definition – each SNP from the interacting set may hold little or no information about the outcome. This leads to a combinatorial explosion in hypothesis formation and testing. Searching for interacting SNP pairs using the data from SNP chips with a million probes means scoring and ranking of about $5 \cdot 10^{11}$ hypotheses – one for each candidate pair. Because of a limited number of samples in such genomic studies, the number of spurious results can be overwhelming. As consequent experimental validation of hypotheses is both labor intensive and expensive, we would gladly exchange some statistical power for a lower type I error.

We can approach the above problem with stricter significance thresholds, perhaps by adopting false discovery rate (FDR) controlling approaches [Reiner *et al.*, 2003]. FDR offers a solution to multiple testing problem by controlling the proportion of erroneous significant results. It is more relaxed than family-wise type I error rate. A recently proposed scoring method called Hypothesis Free Clinical Cloning (HFCC) [Gayan *et al.*, 2008] approaches the same problem differently. It uses replication groups for feature scoring. The core idea is to partition the available samples into non-overlapping subsets and aggregating the results by proposing only those interacting SNPs that have been consistently found across the subsets. Gayan *et al.* hypothesize that this approach may allow identification of frequent and consistent epistatic effects at the expense of lower test power, and claim to improve the filtering of false positive results at the expense of increasing false negative samples. In the HFCC paper, the method is verified in a practical study and while it shows promise in terms of predictions found, it was not compared to any alternative method.

If HFCC indeed works as its authors proposed, it should

have stirred up the community, and the core idea should spread well beyond interaction analysis. Entire field of whole-genome analysis, using SNP, transcriptome, proteome or metabolome data suffers from the course of dimensionality and low samples to features ratio. Introduction and use of replication groups could change the way we approach hypothesis formation and ranking for any feature scoring method, regardless of it being a standard feature selection or one involving feature synthesis by constructive induction.

In the study reported here, we wanted to verify if the utility of replication groups indeed has the intended effect, that is, it improves the performance of constructed models by the reduction of false positive predictions. To investigate this, we have compared HFCC with more traditional, straightforward approaches to interaction analysis. Our goal was to assess to what degree HFCC avoids the discovery of false positive interactions, and how it compares in this respect to other techniques. We performed our study on a set of simulated data and on several SNP data sets from Gene Expression Omnibus [Barrett *et al.*, 2007].

With our experimental results we were not able to confirm the hypothesis that replication groups efficiently filter false positive results. A simpler technique not using replication groups is performing at least as well as replication group-based analysis. Quite surprisingly, using replication groups we consistently obtained more false positives regardless of the cutoff point.

2 Methods

2.1 Feature scoring based on replication groups

The utility of replication groups requires the partitioning of available samples into non-overlapping subgroups. Ideally, groups should be of equal size and have the same class distribution. Candidate features are scored for each subgroup independently and only interactions whose score on all subgroups exceeds a certain threshold are considered as relevant. There are virtually no limitations to types of features and scoring functions used. A feature can be either a single gene or any combination of them. The only assumption about the scoring function is that better features are scored higher. Formally, scoring a feature a using replication proceeds as follows (Figure 1):

1. A set of data samples S is split into p independent sets S_i, where $0 \leq i < p$ and $S_i \cap S_j = \emptyset$.
2. Scores for feature a on each of the subsets S_i are computed, yielding f_i^a.
3. Given a threshold T, feature a is relevant if $f_i^a > T$ for all $0 \leq i < p$ or equivalently, feature a is relevant if $\min_{0 \leq i < p} f_i^a > T$.

The term

$$\min_{0 \leq i < p} f_i^a$$

defines a feature score for feature a using p replication groups. We can regard the procedure above as a newly constructed feature scoring method and can therefore evaluate it as such.

The basic idea of replication groups is reminiscent to statistical meta-analysis, as using p replication groups is the same as using Wilkinson's method of case p [Birnbaum, 1954]. Combining multiple independent tests of significance, the core idea of replication groups, is believed to be sensible only as a last resort – if underlying statistics or data are unavailable [Birnbaum, 1954]. Replication groups oppose this principle, since they combine multiple significance tests on data subsets when the whole data could be used for analysis. Authors of [Gayan *et al.*, 2008] seem to be aware of this, since they acknowledge the replication groups' loss of power.

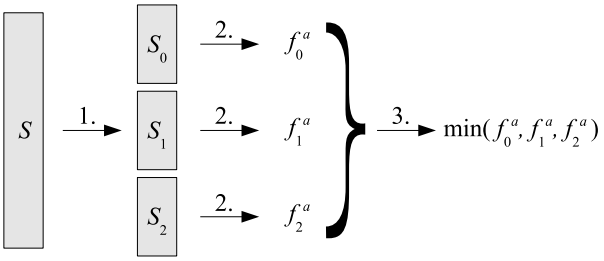


Figure 1: Replication groups-based feature scoring involves three steps: (1) data partitioning, (2) assessment of score f_i^a for given feature a on each of individual data partitions S_i, and (3) computation of the final score f^a, which is the minimal score obtained across different data partitions.

2.2 Experimental methodology

The usefulness of the replication group scoring was assessed by analyzing the differences in distributions of false positives. SNP interaction discovery results obtained by using the replication group scoring technique were compared to those obtained by scoring that did not use replication groups and assessed the interactions score directly, from the entire training data set. Two interaction scoring techniques, HFCC and interaction gain, were used.

For replication group scoring, the data with 541 samples was partitioned to three non-overlapping subgroups in the showcase practical study [Gayan *et al.*, 2008]. Since our data sets are smaller, including from 200 to 500 samples, we have also tested the partitioning to two subgroups. The partitioning was class-stratified and the sizes of resulting groups differed for at most one sample.

Let us assume for the moment true interactions are known. The scoring technique assigns interaction scores to the pairs of SNPs. The quality of discrimination between interacting and non-interacting pairs throughout a whole set of scores can be measured as the area under ROC curve (AUC) [Provost and Fawcett, 2001]. AUC score can be interpreted as the probability that the method correctly discriminates between interacting and non-interacting pairs [Hanley and McNeil, 1982]. Although AUC scores provide valuable insight into the scoring function's quality, they are in essence averaged across scores for all gene pairs.

SNP interaction studies are supposed to report on a small number of most promising interaction pairs. Either pairs of SNPs with scores exceeding a certain threshold or a fixed number of top scored pairs are reported. Since such pairs are usually in minority, they may not affect AUC scores significantly. To complement AUC-based evaluation, we

also counted number of false positives in a list of n best candidate interactions, where n ranged from 1 to the number of all interactions on the data set. We report the results graphically, depicting the dependency of false positives on the number of considered best-scored SNP interactions.

False positives are clearly identifiable in our synthetic data sets. For real SNP experimental data sets, however, the true interactions are not known, but due to selective changes we have made in the data, some of the false interactions are known. We have based our analysis on this particular data set, and counted an algorithm-proposed interaction as a false one only if it belong to this particular set.

All reported evaluation scores are averages over 50 repetitions (bootstrap sampling) for experimental SNP data sets and 100 repetitions (generation of the entire data set using different random seeds) for synthesized data sets.

2.3 Interaction scoring

Hypothesis Free Clinical Cloning

Hypothesis Free Clinical Cloning (HFCC) [Gayan *et al.*, 2008] is a complex suite of epistasis detection methods. It enables scoring of individual genes and gene combinations according to the selected genetic disease models [Li and Reich, 2000] specifying high and low-risk locus combinations. It also features the use of replication groups and different post-hoc filters. Scoring given pairs is simple. Samples are split to two genotypic classes according to the selected disease models. The difference between frequencies of cases and controls in each of two genotypic classes is evaluated with a Wald test Z statistic or a chi-squared statistic with one degree of freedom, obtaining a score for a gene pair and a disease model.

We used the binary version of HFCC software provided as a supplement to [Gayan *et al.*, 2008] with the same configuration as described in HFCC article, performing exhaustive two-locus searches with nine simple disease models of types M1, M2 and M16 [Li and Reich, 2000] – models with only one high-risk two-locus combination. The score of a given pair of genes was the highest score across all disease models for that pair. All post-hoc filters were disabled since any further post-processing could obscure the effects of replication group technique. The software is limited to 299 samples in groups with the same phenotype, therefore it was not possible to analyze data sets with more than 598 samples. Also, the size of the file holding intermediate results can not exceed 2 gigabytes, limiting us to data sets with at most 2000 genes.

Interaction gain

Jakulin and Bratko [2003] proposed interaction gain, a measure for strength of attribute interactions based on information theory principles. The same measure also described as bivariate synergy [Anastassiou, 2007]. Two genes and a class variable are in interaction (scored highly), if those two genes used simultaneously provide more evidence about the class variable as opposed to when used separately. Formally, interaction gain of attributes X and Y with respect to class C is defined as $\mathrm{IntGain}_C(X,Y) = \mathrm{Gain}_C(X \times Y) - \mathrm{Gain}_C(X) - \mathrm{Gain}_C(Y)$, where $\mathrm{Gain}_C(X)$ denotes information gain of attribute X with respect to class C and $X \times Y$ a Cartesian product of attributes X and Y.

	BB	**Bb**	**bb**
AA	0	.10	0
Aa	.10	0	.10
aa	0	.10	0

	BB	**Bb**	**bb**
AA	0	0	.10
Aa	0	.05	0
aa	.10	0	0

Figure 2: Two disease penetrance models. First model specifies that 10% of patients with gene combinations AABb, AaBB, Aabb or aaBb and none of the patients with remaining genotypes have the disease.

We used Orange data mining suite [Demšar *et al.*, 2004] to compute interactions gains. Prior to computation of interaction gain of a pair of genes, samples with either gene's value missing were removed.

3 Data sets

To measure the distribution of false positives, we needed "gold standard" data sets where we know which are the true interactions. Two kinds of data sets were used in our analysis: synthetic data sets and SNP data from actual experimental studies. As synthetic data was generated using known interaction models, the true and false interactions were known. For the SNP data from experimental studies, neither true nor false interactions are known. To enable their use, we here propose a technique which is based on data permutation resulting in a set of SNP pairs for which we know that there should not be any interactions.

3.1 Synthetic data sets

The synthetic data sets were generated according to six two-SNP epistasis models and procedures to simulate four types of noise (5% genotyping error, 5% missing data, 50% phenocopy, 50% genetic heterogeneity) as proposed in [Ritchie *et al.*, 2003]. The models specify disease penetrances for all combinations of polymorphisms for a pair of genes. Two of them are shown on Figure 2. We used the same models to simulate SNP interactions. In contrast to [Ritchie *et al.*, 2003], where there was only one interactions in each data set, out data sets included multiple interactions.

Each interaction in synthetic data sets was defined by a pair of SNPs. Each SNP was involved in at most one interaction, thereby limiting the number of interactions in the data set to half of the number of SNPs. In case of genetic heterogeneity an interaction was defined by two pairs of SNPs, as it comprises two weaker interactions, also limiting the maximum number of interactions to a quarter of the number of SNPs. Genes not involved in an interaction were generated to follow Hardy-Weinberg equilibrium with major allele frequencies randomly chosen between 0.5, 0.75 and 0.9.

We generated two types of data sets, each comprising 200 samples in each (case/control) class:

1. 100 genes with four times all six types of epistasis models, resulting in 24 interactions (model1).
2. 500 genes with ten times all six types of epistatic models, resulting in 60 interactions (model2).

A brief description of noise types follows. For missing data noise (mN), simply 5% randomly chosen values for each SNP were set to missing. Genotyping error (gN) was simulated using directed-error model [Akey *et al.*, 2001], modeling genotyping errors as consistent overrepresentation of one allele. For phenocopies (pN), 50% of affected individuals had a genotype consistent with low disease risk, simulating disease occurrences caused by environmental factors. Genetic heterogeneity (gN) was modeled with two two-locus combinations, where half of the affected individuals had one high risk combination, whereas the other half had another high risk combination.

We have generated six data sets of each type (model1 or model2), with different noise combinations: one without noise (WN), four with each of four noise types (one of mN, gN, pN or gN) and one including noise of all types (AN). In total, twelve data sets were synthesized in this way. The name of each data set name is composed out of the data set type and noise used. For example, synthetic data set on 500 genes with 5% genotyping error is named model2gN.

In synthetic data sets we are always certain whether a pair of genes interacts for data sets generated as described. Since SNP interactions were created independently, these are the only true interactions in the data sets. Still, due to noise, high number of candidate pairs and relatively low number of samples, interaction scoring methods can score other non-interacting SNP pairs highly, leading to false positive discoveries.

3.2 SNP data from Gene Expression Omnibus

We have downloaded a set of SNP data sets from Gene Expression Omnibus [Barrett *et al.*, 2007], for which we requested to contain more than 200 samples, which could be split into two similarly sized groups based on sample subgroup annotations (class). We found five such data sets with following accession numbers:

- GSE6754 [Consortium *et al.*, 2007] describing families with two individuals affected by autism spectrum disorders. Sample's status (affected / unaffected) was used as a class. Due to the computational limitations of HFCC software only first 2000 SNPs were considered and a (stratified) sample of 500 reading was used, including 292 affected and 208 unaffected samples.
- GSE8054 [Tan *et al.*, 2008] comprising 901 SNPs for each of the 121 cancerous and 87 control samples.
- GSE8055 [Tan *et al.*, 2008] comprising 1189 SNPs for each of the 141 cancerous and 89 control samples.
- GSE7226 [Friedman *et al.*, 2006] with platform designation GPL2004, comprising 102 samples from mentally retarded children and 213 samples from their unaffected siblings or parents. Only the first 2000 SNPs were considered.
- GSE7226 [Friedman *et al.*, 2006] with platform designation GPL2005, comprising 103 samples from mentally retarded children and 210 samples from their unaffected siblings or parents. Only the first 2000 SNPs were considered.

True SNP interactions that could be inferred from these data sets are now known. To enable their use in our study, we deliberately destroyed some interactions by random data permutations. For this, one half of the SNPs were randomly chosen and their values were randomly permuted between samples. SNP pairs including at least one SNP with permuted values were afterwords regarded as interactions that were not supposed to be inferred from the data, and constituted the set of potential false positive discoveries.

4 Results and discussion

Our goal was to find out whether scoring by replication groups reduces the discovery of false positive SNP interactions as inferred from twelve synthetic and five experimental SNP data sets. We report AUC scores for each combination of data set and scoring method, where interaction scoring was done on entire data set (no replication groups used) or was estimated using either two or three replication groups (Table 1). With both interaction gain and HFCC the AUC scores decreased with increasing number of replication groups. The sole exception is the gse8055 data set scored with interaction gain, but the difference is slight. Notice that AUC scores for non-synthetic data sets are very low due to a low number of interactions among the non-permuted genes.

Each graph on Figure 3 depicts a single base scoring method used either on the entire data set directly or with 2 or 3 replication groups. The count of false positives with respect to a number of selected best candidates is shown, lower counts being better. The use of replication groups increased the proportion of false positive results for all combinations of tested data sets and base scoring methods at any given cutoff. Also, two replication groups consistently performed better than three, hinting that increasing their number would not lead to better results. We were not able to find a single data set where replication groups improved the filtering of false positive results. In fact, using replication groups seems to increase the false positive rate.

The use of replication groups affected results based on interaction gain more than those based on HFCC scoring, both in terms of AUC and counts of false positives. For synthetic data, AUC scores using interaction gain on the whole data set directly are consistently better than those of HFCC scoring. Since replication groups decrease AUC scores of interaction gain scoring more than those of HFCC, with three replication groups the role is reversed – HFCC performed better. This might indicate that interaction gain is more sensitive to smaller numbers of samples. Interestingly, the HFCC achieved better AUC scores on experimental data regardless of the number of subgroups used. The reasons for bad performance of interaction gain on data from GEO are unclear.

5 Conclusion

Finding possible SNP interactions from data sets that include from several thousands to several hundred of thousands SNPs and only several hundreds of cases seems elusive. The approach which tests all pairs of SNPs is computationally feasible, but generates vast number of features and can, due to noise and vast number of hypothesis tested,

Table 1: AUC scores on both synthetic and data sets from GEO for scoring from entire data set (EDS) or replication groups scoring using two (2RG) or three groups (3RG).

interaction gain	EDS	2RG	3RG
model1WN	1.000	0.993	0.967
model1gN	0.998	0.980	0.941
model1mN	1.000	0.990	0.955
model1pN	0.934	0.820	0.726
model1hN	0.938	0.845	0.775
model1AN	0.690	0.618	0.584
model2WN	1.000	0.993	0.965
model2gN	0.998	0.978	0.929
model2mN	0.999	0.988	0.951
model2pN	0.932	0.804	0.702
model2hN	0.933	0.818	0.725
model2AN	0.673	0.562	0.505
gse7226-gpl2004	0.558	0.517	0.502
gse7226-gpl2005	0.587	0.530	0.509
gse8054	0.495	0.493	0.493
gse8055	0.493	0.497	0.497
gse6754	0.679	0.593	0.553

HFCC	EDS	2 RG	3 RG
model1WN	0.998	0.990	0.973
model1gN	0.993	0.972	0.942
model1mN	0.996	0.985	0.964
model1pN	0.904	0.812	0.749
model1hN	0.911	0.827	0.774
model1AN	0.661	0.611	0.581
model2WN	0.997	0.989	0.971
model2gN	0.992	0.970	0.935
model2mN	0.995	0.984	0.959
model2pN	0.899	0.801	0.735
model2hN	0.902	0.812	0.745
model2AN	0.637	0.571	0.538
gse7226-gpl2004	0.619	0.571	0.547
gse7226-gpl2005	0.637	0.584	0.556
gse8054	0.604	0.571	0.551
gse8055	0.584	0.557	0.541
gse6754	0.702	0.642	0.607

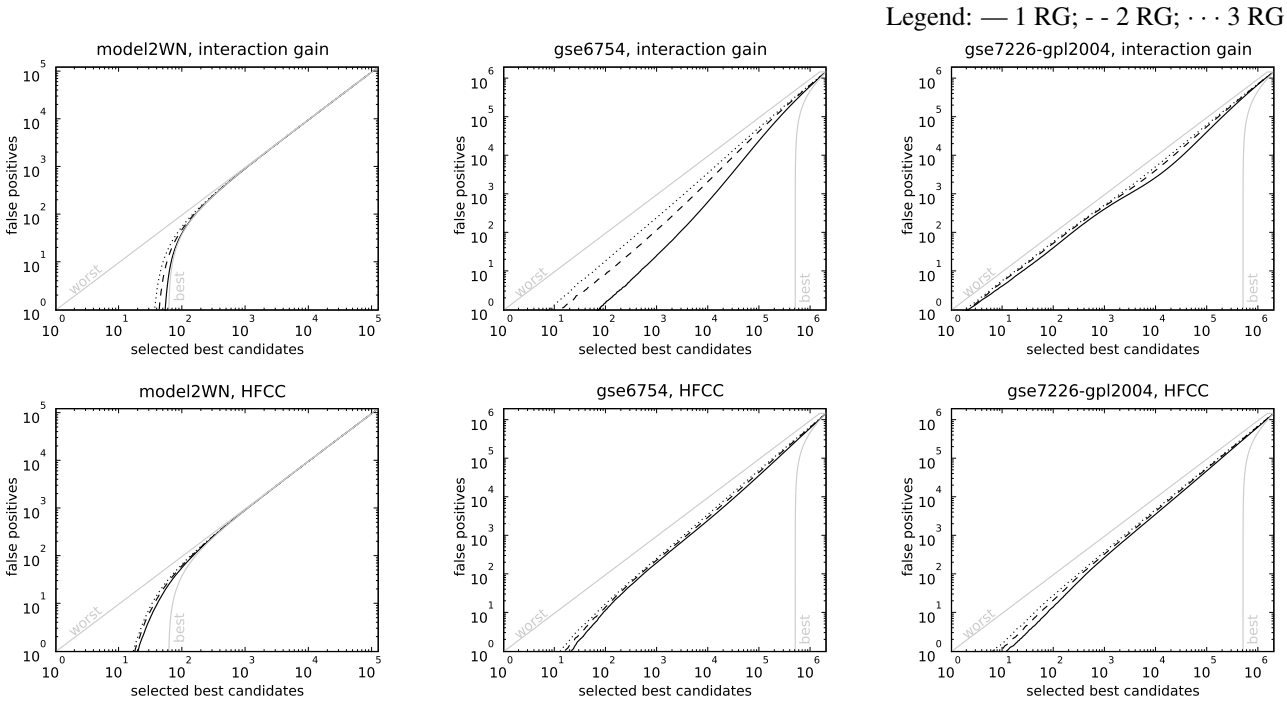


Figure 3: Number of false positives in a given number of selected best candidates. Solid curves show results without the use of replication groups (RG), dashed curves for 2 replication groups and dotted curves for 3 replication groups. Lower curves represent better scoring. Both axes are in logarithmic scale to emphasize results for smaller numbers of best candidates. The (theoretically) best and worst possible scorings are shown in light gray. The results for other data sets (not shown here) were similar and exhibited similar rankings of scoring techniques.

result in many false positive observations. A recently proposed scoring technique used in the method called Hypothesis Free Clinical Cloning that uses replication groups raised hopes in scoring the interactions in attempt to decrease false positives. However, the experiments presented in this paper indicate that the method does not improve upon the standard scoring which does not use replication groups. Estimating interactions scores directly from the entire data set performed better.

At this point, it looks like computational approaches cannot overcome lack of experimental data (samples) and very low sample-to-feature ratio. The venue toward more robust scoring of SNP interactions is therefore an increase of sample size or use of background knowledge, like information on protein-to-protein interactions, knowledge on pathways or functional gene labels [Pattin and Moore, 2008].

Acknowledgments

This work was supported by grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

References

[Akey *et al.*, 2001] Joshua M. Akey, Kun Zhang, Momiao Xiong, Peter Doris, and Li Jin. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *The American Journal of Human Genetics*, 68(6):1447–1456, 2001.

[Anastassiou, 2007] Dimitris Anastassiou. Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol*, 3, February 2007.

[Barrett *et al.*, 2007] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucl. Acids Res.*, 35:D760–765, 2007.

[Birnbaum, 1954] Allan Birnbaum. Combining Independent Tests of Significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

[Consortium *et al.*, 2007] Autism Genome Project Consortium, P Szatmari, AD Paterson, L Zwaigenbaum, et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics*, 39(3):319–328, 2007.

[Demšar *et al.*, 2004] J. Demšar, B. Zupan, and G. Leban. Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper, 2004.

[Friedman *et al.*, 2006] J.M. Friedman, Agnes Baross, Allen D. Delaney, Adrian Ally, et al. *The American Journal of Human Genetics*, 79(3):500–513, 2006.

[Gayan *et al.*, 2008] Javier Gayan, Antonio Gonzalez-Perez, Fernando Bermudo, Maria Saez, Jose Royo, Antonio Quintas, Jose Galan, Francisco Moron, Reposo Ramirez-Lorca, Luis Real, and Agustin Ruiz. A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics*, 9(1):360, 2008.

[Hahn *et al.*, 2003] Lance W. Hahn, Marylyn D. Ritchie, and Jason H. Moore. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19(3):376–382, 2003.

[Hanley and McNeil, 1982] JA Hanley and BJ McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[Jakulin and Bratko, 2003] Aleks Jakulin and Ivan Bratko. Analyzing attribute dependencies. In *PKDD 2003, volume 2838 of LNAI*, pages 229–240. Springer-Verlag, 2003.

[Li and Reich, 2000] W Li and J Reich. A complete enumeration and classification of two-locus disease models. *Hum Hered*, 50:334–349, 2000.

[Moore, 2005] Jason H. Moore. A global view of epistasis. *Nature Genetics*, 37(1), 2005.

[Park and Hastie, 2008] Mee Young Park and Trevor Hastie. Penalized logistic regression for detecting gene interactions. *Biostat*, 9(1):30–50, 2008.

[Pattin and Moore, 2008] Kristine Pattin and Jason Moore. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Human Genetics*, 124(1):19–29, August 2008.

[Provost and Fawcett, 2001] Foster J. Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[Reiner *et al.*, 2003] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, February 2003.

[Ritchie *et al.*, 2003] Marylyn D. Ritchie, Lance W. Hahn, and Jason H. Moore. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiology*, 24(2):150–157, 2003.

[Smith *et al.*, 2005] G Davey Smith, S Ebrahim, S Lewis, AL Hansell, LJ Palmer, and PR Burton. Genetic epidemiology and public health: hope, hype, and future prospects . *The Lancet*, 366(9495):1484–1498, 2005.

[Tan *et al.*, 2008] Aik Choon Tan, Jian-Bing Fan, Collins Karikari, Marina Bibikova, et al. Allele-specific expression in the germline of patients with familial pancreatic cancer: an unbiased approach to cancer gene discovery. *Cancer biology & therapy*, 7(1):135–144, 2008.

Symbolic Representations for Reasoning about Temporal Gene Profiles

Marco Falda
Dept. of Information Engineering
Via Gradenigo, 6 - 35131 Padova - Italy
marco.falda@unipd.it

Abstract

Understanding the mechanism of gene regulation in a living cell is very important to predict the behavior of cell in response to interacting factors. In fact such prediction capability can lead to the development of improved diagnostic tests and therapeutics. In this work, it is proposed to represent symbolically gene expression time series and to adapt sub-string matching algorithms (such as the Longest Common Subsequence) for deciding the regulation direction. As a preliminary validation test, the approach is applied to four biological datasets composed of Yeast cell-cycle regulated genes under different synchronization methods.

1 Introduction

Microarrays have allowed the interrogation of gene expression data in a massive way, and have become the most popular platform in Systems Biology. Temporal gene profiles coming from Microarrays experiments capture expression of genes at discrete time points in a cellular process; numerous time-series Microarray experiments have been performed to study such biological processes, for example the biological rhythms or circadian clock of *Arabidopsis,* or its flowering time [Laule *et al.*, 2003; Carpita *et al.*, 2001]. A significant challenge in dealing with genomic data comes from the enormous number of genes involved in biological systems (for example the human Genome has 30.000 genes); furthermore, the unavoidable presence of noise enhances the difficulty in distinguishing real from random patterns and increases the potential of misleading analyses. To overcome these problems, some studies proposed to identify symbolic features of the series; examples include temporal abstraction-based methods that define trends (i.e., increasing, decreasing and steady) over subintervals [Sacchi *et al.*, 2005], or a difference-based method that uses the first and second order differences in expression values to detect the direction and rate of change of the temporal expressions for clustering [Kim and Kim, 2007].

Symbolic methods have also the benefit to reduce the noise in the original data to some degree when decreasing the dimension of the time-series data, thus making the subsequent analysis more robust to noise. This was demonstrated by Sacchi et al. [Sacchi *et al.*, 2005] with their adaptation of the Temporal Abstractions (TA)-clustering method from the field of Artificial Intelligence to gene expression analysis.

In this paper a recently started study about symbolic representations for gene temporal profiles will be presented. Tests performed on a simple fragment of a real biological regulatory network seem to show that such qualitative representations could be useful for finding the correct regulation directions, since they have the further advantage to be able to abstract delays among genes, and therefore be less penalized by the diverse temporal scales typical of biological systems.

2 Symbolic representations

Interactions among genes can be formalized as a directed graph $\langle G, A\rangle$ where G represents the set of genes and A the set of relations between genes; the graph can be weighted by associating a number to each arc $a_{ij} \in A$, but in a simpler scenario each arc a_{ij} will assume the value 1 or 0 depending on the fact that gene i influences gene j or not. The temporal evolution of a single gene in a regulatory network, that is its *time series*, is usually represented as a sequence of K samples $\mathcal{V} = \{v_k, k \in \{1, \ldots, K\}\}$, where $k \in \mathbb{N}^+$ is the index of the discrete sampling time and $v_k \in \mathbb{R}$ its value at index k.

2.1 Preprocessing of data

When one is measuring a variable that is both slowly varying and also corrupted by random noise, as in the case of gene temporal profiles, it can sometimes be useful to replace each data point by some kind of local average of surrounding data points. Since nearby points measure very nearly the same underlying value, averaging can reduce the level of noise without (much) biasing the value obtained.

A particular type of low-pass filter, well-adapted for data smoothing is Savitzky-Golay filters family, initially developed to render visible the relative widths and heights of spectral lines in noisy spectrometric data [Savitzky and Golay, 1964]. The simplest type of digital filter replaces each data value $v_k \in \mathcal{V}$ by a linear combination of itself and some number of nearby neighbors:

$$\widetilde{v}_k = \sum_{n=-n_L}^{n_R} c_n \cdot v_{k+n}$$

Here n_L is the number of points used "to the left" of a data point k, i.e., earlier than it, while n_R is the number used to the right, i.e., later.

The algorithm of Savitzky-Golay applies the least-squares principle to determine an improved set of kernel coefficients c_n for use in a digital convolution; these improved coefficients are determined using polynomials rather than, as for the case of simple averaging, a constant value determined from a sub-range of data. Indeed, the Savitzky-Golay method could be seen as a generalization of averaging data, since averaging a sub-range of data corresponds to a Savitzky-Golay polynomial of degree zero. The idea of this kind of filtering is to find coefficients that preserve higher moments.

2.2 Features

To reason about the temporal evolution of each gene, a symbolic representation can be developed starting from quantitative data and applying simple Discrete Calculus definitions; in this way, it is possible to describe a time series $\mathcal{V}$ as sequence of symbols $\mathcal{S}_\mathcal{V}$ representing significant features. The features that have been considered are: maxima, minima, inflection points and points where the series becomes stationary, zero or saturates:

Definition 1 (Symbolic features)**.** *The significant features of a time series are defined over the set* $\mathcal{F} = \{M, m, f, s, z, S\}$.

Definition 2 (Symbolic representation)**.** *A time series* $\mathcal{V}$ *can be represented as a sequence of symbols* $\mathcal{S}_\mathcal{V} = \{\sigma_j, j \in \{1, \ldots, J\}\}$ *where each symbol* σ_j *belongs to the set of features* $\mathcal{F}$.

To maintain a link with the original series a mapping function $m_\mathcal{S}$ between $\mathcal{S}_\mathcal{V}$ and $\mathcal{V}$ is defined:

Definition 3 (Mapping function)**.** *Given a symbolic representation* $\mathcal{S}_\mathcal{V}$ *and its original time series* $\mathcal{V}$, $m_\mathcal{S} : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ *is a function that maps the index* j *of a symbol* $\sigma_j \in \mathcal{S}_\mathcal{V}$ *in the index* k *of the corresponding time series element* $v_k \in \mathcal{V}$.

2.3 Enriching the symbolic representation

In the symbolic sequence it is possible to add further information, namely the intensity, both relative and absolute, of the time series at a given point, and the width of the feature itself. To do this, it is necessary to define how this kind of information will be represented, and a natural way is to express it in terms of time series parameters.

Definition 4 (Range of a time series)**.** *The range of a time series* $\mathcal{V} = \{v_k, k \in \{1, \ldots, K\}\}$ *is provided by the function* $ext : \mathbb{R}^K \longrightarrow \mathbb{R}^+$ *defined as* $ext(\mathcal{V}) = |\max_k(v_k) - \min_k(v_k)|$.

Definition 5 (Range of a set of time series)**.** *The range of a set of time series* $\mathcal{W} = \{\mathcal{V}_h, h \in \{1, \ldots, H\}\}$ *is defined as* $set\_ext : (\mathbb{R}^K)^H \longrightarrow \mathbb{R}^+, set\_ext(\mathcal{W}) = |\max(v_k) - \min(v_k)|, v_k \in \mathcal{W}$.

Definition 6 (Length of a time series)**.** *The length of a time series is the cardinality of the set* $\mathcal{V}$ *and it will be written as* $|\mathcal{V}|$.

Given these basic parameters which allow to have a reference w.r.t. a specific time series and w.r.t. the whole set of time series, it is possible to describe more intuitively the properties of the features identified.

Definition 7 (Absolute height of a feature)**.** *Given a set of time series* $\mathcal{W} = \{\mathcal{V}_h, h \in \{1, \ldots, H\}\}$ *and a symbolic sequence* $\mathcal{S}_{\mathcal{V}h}$*, the absolute height of the feature represented by the symbol* $\sigma_j \in \mathcal{S}_{\mathcal{V}h}$ *is defined by the function* $ha_\mathcal{S} : \mathbb{N}^+ \longrightarrow \mathbb{R}^+$

$$ha_\mathcal{S}(j) = \left| \frac{v_{m_\mathcal{S}(j)}}{set\_ext(\mathcal{W})} \right|$$

Definition 8 (Relative height of a feature)**.** *Given time series* $\mathcal{V} = \{v_k\}$ *and its symbolic sequence* $\mathcal{S}_\mathcal{V}$*, the relative height of the feature represented by the symbol* $\sigma_j \in \mathcal{S}_\mathcal{V}$ *is defined by the function* $hr_\mathcal{S} : \mathbb{N}^+ \longrightarrow \mathbb{R}^+$

$$hr_\mathcal{S}(j) = \left| \frac{v_{m_\mathcal{S}(j)} - v_{m_\mathcal{S}(j-1)}}{ext(\mathcal{V})} \right|$$

Definition 9 (Width of a feature)**.** *Given time series* $\mathcal{V}$ *and its symbolic sequence* $\mathcal{S}_\mathcal{V}$*, the width of the feature represented by the symbol* $\sigma_j \in \mathcal{S}_\mathcal{V}$ *is defined by the function* $wr_\mathcal{S} : \mathbb{N}^+ \longrightarrow \mathbb{R}^+$

$$wr_\mathcal{S}(j) = \left| \frac{m_\mathcal{S}(j) - m_\mathcal{S}(j-1)}{|V|} \right|$$

These functions can be associated to the symbols of a sequence $\mathcal{S}$ by means of a function $q_\mathcal{S}$ that describes the properties of a feature.

Definition 10 (Properties of a symbol)**.** *Given a symbolic sequence* $\mathcal{S}_\mathcal{V}$*, the properties of a symbol* $\sigma_j \in \mathcal{S}_\mathcal{V}$ *are defined by the function* $q_\mathcal{S} : \mathbb{N}^+ \longrightarrow \langle \mathbb{R}^+, \mathbb{R}^+, \mathbb{R}^+ \rangle$

$$q_\mathcal{S}(j) = \langle ha_\mathcal{S}(j), hr_\mathcal{S}(j), wr_\mathcal{S}(j) \rangle$$

Example 1. *The series* $\mathcal{V}$ *in Figure 1 can be represented by the sequence* $\mathcal{S}_\mathcal{V} = \{m, f, M, f, m, \ldots\}$*, and the properties of its symbols are* $q_\mathcal{S}(1) = \langle 0.63, 0, 0 \rangle$*,* $q_\mathcal{S}(2) = \langle 0.12, 0.51, 0.06 \rangle$*,* $q_\mathcal{S}(3) = \langle 0.33, 0.45, 0.09 \rangle$*,* $q_\mathcal{S}(4) = \langle 0.08, 0.25, 0.07 \rangle$*,* $q_\mathcal{S}(5) = \langle 0.17, 0.25, 0.09 \rangle$ *et c.* .

3 Reasoning about regulation directions

The symbolic representation of time series allows reasoning about strings in which each symbol representing a feature is linked to a point of the real series (through an index given by the function $m_\mathcal{S}$). In the following five methods have been considered. In all cases the hypothesis that in a causal process the cause always precedes its consequence is assumed and exploited. By now just the basic symbolic representations have been used.

Reverse Engineering of a gene regulatory network means inferring relations among genes starting from experimental data, in this specific case from time series data. It can be solved by providing a "similarity measure" function $f : \mathbb{N}^{|G|} \longrightarrow \mathbb{R}$ from a set of indices, which identify the genes, to a real number; $|G|$ represents the cardinality of the set G. Since the focus of this work is the symbolic processing of time series, the domain of the measures will

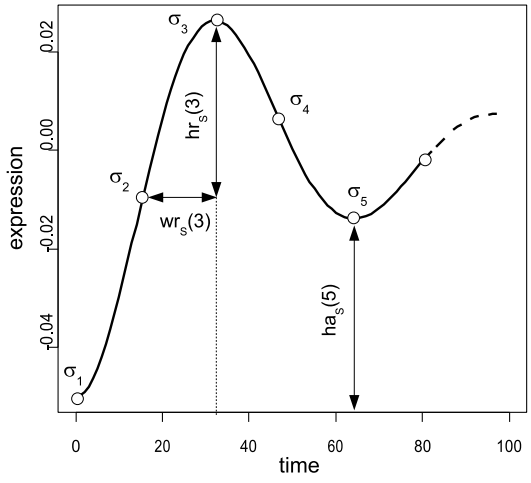


Figure 1: Example of a numerical time series and its symbolic representation.

be $\mathcal{F}^J \times \mathcal{F}^J$, that is pairs of symbolic sequences whose length is J. In this work these measures will be used just to establish whether two genes are correlated or not, so the resulting real number will be eventually compared with a given threshold to obtain a Boolean value.

3.1 Shifted Correlation (sC)

The simplest metric that can be applied on the symbol sequences is a Pearson correlation. The aim is to identify directions, so this measure has been made asymmetric by shifting the series by one temporal sample (the cause precedes the effect); it will be called "Shifted Correlation" (sC). The correlation is applied to the original time series points identified by the mapping function $m_{\mathcal{S}}$.

3.2 Matching between maxima and minima with temporal precedences (tMM)

A second easy idea is to find a one-to-one correspondence between maxima and minima, direct or inverse, in the symbolic representation taking into account the fact that the regulator gene should always precede the regulated one, and to evaluate the relative length of the matching features with respect to the shorter sequence:

$$tMM(\mathcal{S}_1, \mathcal{S}_2) = \frac{\max(|\mathcal{M}_{1,2}|, |\mathcal{M}^-_{1,2}|)}{\min(|\mathcal{M}_1|, |\mathcal{M}_2|)}$$

$\mathcal{M}_1$ and $\mathcal{M}_2$ are sub-sequences containing only maxima and minima of the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$, $\mathcal{M}_{1,2} = \{\sigma_{j1} \in \mathcal{S}_1 : \exists \sigma_{j2} \in \mathcal{S}_2 \wedge \sigma_{j1} = \sigma_{j2} \wedge m_{\mathcal{S}_1}(j_1) < m_{\mathcal{S}_1}(j_2)\}$ and $\mathcal{M}^-_{1,2}$ is defined as $\mathcal{M}_{1,2}$ but matching in an inverse fashion (e.g.: maxima with minima and vice versa).

3.3 Temporal Longest Common Substring ($tLCStr$)

A further step is to notice that noise could alter the series, therefore it could be the case that just some segments of the temporal expressions match, therefore looking for the longest segment should help. The longest segment shared between two symbolic sequences can be found using the Longest Common Substring algorithm, which exploits Dynamic Programming techniques and has a $O(J^2)$ asymptotic complexity in the worst case [Cormen *et al.*, 2005]. As for the precedence criterion, the algorithm matches only the features of the regulator gene which precede the corresponding features of the regulated one (the "t" in the name $tLCStr$). The formula is:

$$tLCStr(\mathcal{S}_1, \mathcal{S}_2) = \frac{\max(|tLCStr_{1,2}|, |tLCStr^-_{1,2}|)}{\min(|\mathcal{S}_1|, |\mathcal{S}_2|)}$$

where $tLCStr^-_{1,2}$ is the Longest Common Substring matching in an inverse fashion.

3.4 Temporal Longest Common Subsequence ($tLCS$)

It is possible to hypothesize that the effects of a gene could be hidden by saturation effects, and therefore trying to identify the longest non-contiguous subsequence shared between two symbolic sequences could be useful. Also in this case there exists a $O(J^2)$ algorithm based on Dynamic Programming techniques [Cormen *et al.*, 2005]; the formula is analogous to the previous one and so it has not been reported here.

The precedence criterion has been added as in the previous case:

$$tLCS(\mathcal{S}_1, \mathcal{S}_2) = \frac{\max(|tLCS_{1,2}|, |tLCS^-_{1,2}|)}{\min(|\mathcal{S}_1|, |\mathcal{S}_2|)}$$

where $tLCS^-_{1,2}$ is the Longest Common Subsequence matching in an inverse fashion.

3.5 Directional Dynamic Time Warping ($dDTW$)

The last algorithm, adapted to take into account the asymmetry of the time arrow, is the Dynamic Time Warping, a procedure coming from the Speech Recognition field [Sakoe and Chiba, 1978]; it is a "elastic" alignment that allows similar shapes to match even if they are out of phase in the time axis; the algorithm complexity is again $O(J^2)$. The precedence criterion has been added by matching features of regulated genes with preceding features of the regulator ones (Figure 2). The computations are performed on the original time series points identified by the mapping function $m_{\mathcal{S}}$.

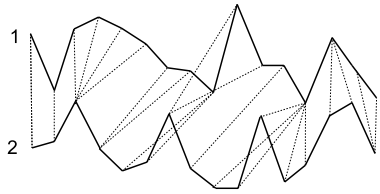
3.6 Adding qualitative properties

In the symbolic sequence it is possible to add further information, namely the intensity, both relative and absolute, of the time series at a given point, and the relative width of the feature itself; the definitions are not reported in this paper, we will simply postulate the existence of the functions $hr_{\mathcal{S}}(j)$, $ha_{\mathcal{S}}(j)$ and $wr_{\mathcal{S}}(j)$ respectively.

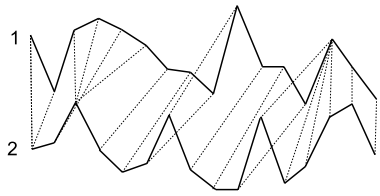
These functions can be associated to the symbols of a sequence $\mathcal{S}$ by means of a function $q_{\mathcal{S}}$ that describes the properties of a feature.

Definition 11 (Qualitative properties)**.** *Given a symbolic sequence $\mathcal{S}_{\mathcal{V}}$, the properties of a symbol $s_j \in \mathcal{S}_{\mathcal{V}}$ are given by the function $q_{\mathcal{S}} : \mathbb{N}^+ \longrightarrow \langle \mathbb{R}^+, \mathbb{R}^+, \mathbb{R}^+ \rangle$*

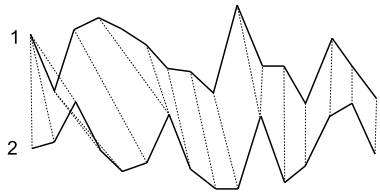
$$q_{\mathcal{S}}(j) = \langle ha_{\mathcal{S}}(j), hr_{\mathcal{S}}(j), wr_{\mathcal{S}}(j) \rangle$$



(a) original DTW;



(b) dDTW directed towards the past;



(c) dDTW directed towards the future;

Figure 2: original vs. directional DTW.

Functions $ha_{\mathcal{S}}(j)$, $hr_{\mathcal{S}}(j)$ and $wr_{\mathcal{S}}(j)$ has been quantized in a fixed number n of levels:

Definition 12 (Quantized functions)**.** *Given a number $n \in \mathbb{N}$, the quantized version of a function $f : \mathbb{N}^+ \longrightarrow \mathbb{R}^+$ is a function $\varphi_n : (\lambda : \mathbb{N}^+ \longrightarrow \mathbb{R}^+) \longrightarrow \mathbb{N}^+$*

$$\varphi_n[f] = \left\lceil \frac{f}{|\max[f]|} \cdot n \right\rceil$$

In this way the properties can be fuzzified in n levels and compared in an approximated way using a function $\widetilde{eq}_{\mathcal{S}_1,\mathcal{S}_2}$ defined as follows.

Definition 13 (Approximately equal)**.** *Given two symbol sequences $\mathcal{S}_1$ and $\mathcal{S}_2$ the function $\widetilde{eq}_{\mathcal{S}_1,\mathcal{S}_2} : \mathbb{N}^+ \times \mathbb{N}^+ \rightarrow \{0,1\}$ is defined as*

$$\begin{aligned}\widetilde{eq}_{\mathcal{S}_1,\mathcal{S}_2}(j_1,j_2) \quad = \quad & g((\varphi_n[ha_{\mathcal{S}1}](j_1)) = \varphi_n[ha_{\mathcal{S}2}](j_2)),\\ & (\varphi_n[hr_{\mathcal{S}1}](j_1)) = \varphi_n[hr_{\mathcal{S}2}](j_2)),\\ & (\varphi_n[wr_{\mathcal{S}1}](j_1)) = \varphi_n[wr_{\mathcal{S}2}](j_2)))\end{aligned}$$

where $g : \{0,1\}^3 \rightarrow \{0,1\}$ is a function that weights the relevance of each qualitative property and can be defined using heuristics.

4 Results

To test the five measures discussed above, time series coming from the Yeast cell cycle under four different synchronization conditions [Spellman *et al.*, 1998] have been considered; each series has 26 time samples. To validate the results the simplified Yeast network topology from [Li *et al.*, 2004], which represents interactions among 29 genes, has been chosen (Figure 3).

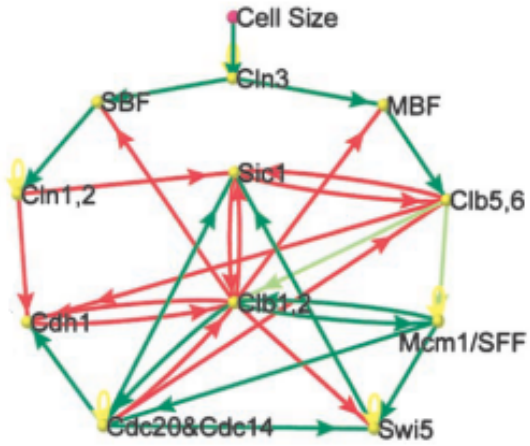


Figure 3: Simplified cell-cycle network with only one checkpoint [Li *et al.*, 2004].

Also the algorithms which exploit the qualitative properties of the features have been implemented but, by now, no extensive tests have yet been done.

As a performance criterion, the precision of the above algorithms in recognizing the regulation directions has been taken into account in the hypothesis that another algorithm gave the correct undirected pairs (for example the state-of-the-art ARACNe algorithm [Margolin *et al.*, 2006] has good performances but it does not compute directions). Let a_{ij} be the arc between two genes i and j in the graph G and $f(i,j)$ be a function that estimates how much they are correlated, then the definitions for *true positives* (TP), *false positives* (FP) and *false negatives* (FN) become:

$$TP \Leftarrow (a_{ij} = 1) \wedge f(i,j) > \vartheta$$

$$FP \Leftarrow (a_{ij} = 0 \wedge a_{ji} = 1) \wedge f(i,j) > \vartheta$$

$$FN \Leftarrow (a_{ij} = 1) \wedge f(i,j) \leq \vartheta$$

where ϑ is a threshold, in this work set to zero.

In particular, two common indices have been calculated[1]: the positive predictive value (PPV), called also precision, which refers to the fraction of returned true positives that are really positives:

$$PPV = \frac{TP}{TP + FP}$$

and the sensitivity (also known as recall), which gives the proportion of real positives which are correctly identified:

$$Sensitivity = \frac{TP}{TP + FN}$$

In order to have an idea of the performances obtained, a software based on Dynamic Bayesian Networks (Banjo [Yu *et al.*, 2004]) has been applied with default parameters

[1] the performances of ARACNe algorithm on the considered dataset for the problem of identifying *undirected* pairs are: PPV = 65.2 % and Sensitivity = 13.9 %.

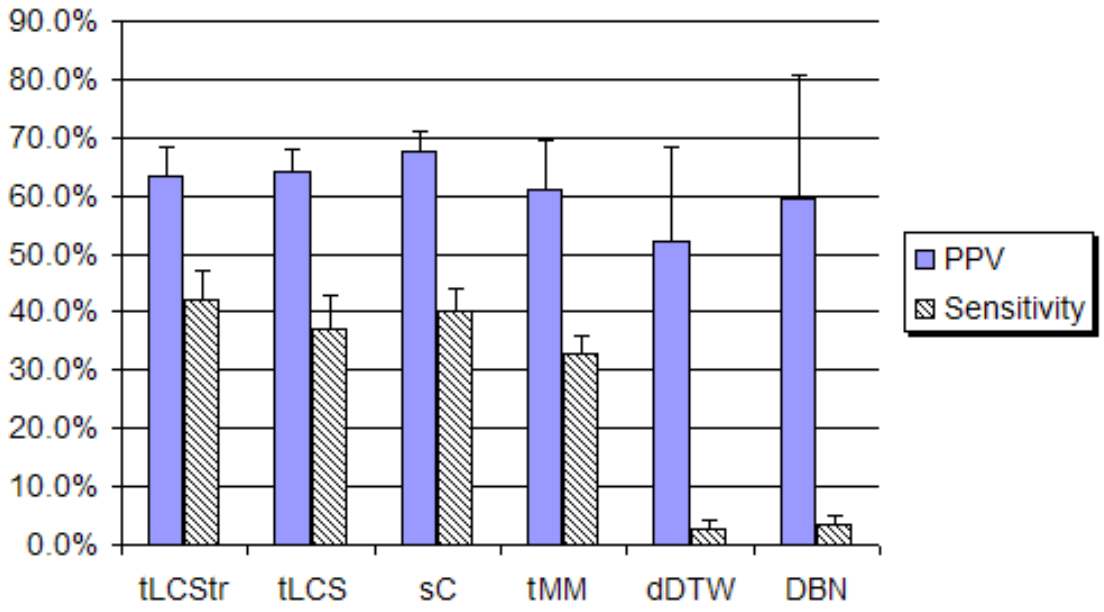


Figure 4: Positive Predictive Value and Sensitivity of the four algorithms proposed in the paper compared with those of Dynamic Bayesian Networks.

and 7 quantization levels to the same datasets: it seems to be precise but not very sensitive; the upper bound time for computation has been set to 10 minutes, a reasonable time, considering that the other four algorithms take seconds.

The mean performances of the five measures over the four different synchronization experiments have been reported in Table 1; for sC and $dDTW$ there are also numerical versions, computed on the original time series (their scores have been reported in parentheses). In Figure 4 the results of the algorithms operating on symbolic data have been plotted with their standard deviation as error bar.

Table 1: Mean PPV and sensitivity values for the measures discussed in the paper over different synchronization experiments (in parentheses the performances on the numerical series).

	PPV	**Sensitivity**
sC	67.5 % (64.6 %)	40.0 % (38.5 %)
tMM	61.1 %	32.8 %
$lcstr$	63.4 %	42.0 %
lcs	64.0 %	32.1 %
$dDTW$	52.2 % (51.3 %)	2.6 % (2.2 %)
Banjo (DBN)	59.3 %	3.6 %

5 Discussion and conclusions

It is possible to notice that the symbolic versions of sC and $dDTW$ both improve with respect to their numerical counterparts. Besides, all the measures provide a PPV above the 50% threshold; this means that they could be useful for deciding regulation directions.

The next step will be to perform extended tests, possibly on larger datasets, with series enriched by qualitative properties of the features estimated using fuzzy quantization levels; hopefully, this should enhance the recall index, that are still under the threshold of a random choice, in particular the recall of $dDTW$, the most recently studied among the five measures proposed.

Real gene networks present more complex patterns than simple direct pairwise regulations, for example the so-called "network motifs" [Alon, 2007]; another research direction concerns the combination of more causes to identify relations among several genes and, hopefully, improve the overall performances when considering more general problems, and not just the regulation directions.

References

[Alon, 2007] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8:450–461, 2007.

[Carpita *et al.*, 2001] N. Carpita, M. Tierney, and M. Campbell. Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. *Plant. Mol. Biol.*, 47:1–5, 2001.

[Cormen *et al.*, 2005] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. McGraw-Hill, 2nd edition, 2005.

[Kim and Kim, 2007] J. Kim and J. H. Kim. Difference-based clustering of short time-course microarray data with replicates. *BMC Bioinformatics*, 8:253, 2007.

[Laule *et al.*, 2003] O. Laule, A. Fürholz, H. S. Chang, T. Zhu, X. Wang, P. B. Heifetz, W. Gruissem, and M. Lange. Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. *PNAS USA*, 100:6866–6871, 2003.

[Li *et al.*, 2004] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101:4781–4786, 2004.

[Margolin *et al.*, 2006] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 2006.

[Sacchi *et al.*, 2005] L. Sacchi, R. Bellazzi, C. Larizza, P. Magni, T. Curk, U. Petrovic, and B. Zupan. TA-clustering: Cluster analysis of gene expression profiles through temporal abstractions. *Int. J. Med. Inform.*, 74:505–517, 2005.

[Sakoe and Chiba, 1978] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acous. Speech Signal Process.*, 26(1):43–49, 1978.

[Savitzky and Golay, 1964] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[Spellman *et al.*, 1998] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. of the Cell*, 9:3273–3297, 1998.

[Yu *et al.*, 2004] J. Yu, V. Smith, P. Wang, A. Hartemink, and E. Jarvis. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20:3594–3603, 2004.

dictyExpress: An explorative web-based interface to *Dictyostelium discoideum* gene expression database

Gregor Rot[2,*], Anup Parikh[1,*], Tomaz Curk[2], Adam Kuspa[1,3], Gad Shaulsky[1], Blaz Zupan[1,2,#]

1 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
2 Faculty of Computer and Information Science, University of Ljubljana, Slovenia
3 Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA

Abstract

We present dictyExpress, an Internet application offering explorative analysis over a collection of gene expression data on social amoeba *D. discoideum*. With using the newest engineering techniques resulting in visualization and interaction-rich interface, it could be considered as a precursor for the next generation systems biology software applications.

1 Introduction and motivation

Public databases of results from high-throughput experiments are abundant and important but most biologists lack the training and the computer power to interact with the data. A solution to this problem is afforded by recent developments in information technology, which facilitate the development of web-based systems that support interaction and explorative data analysis. These systems require only basic web-surfing skills and modest computer power, but deliver powerful data analysis capabilities to the biologist's fingertips.

Gene expression data from microarray experiments in the social amoeba *Dictyostelium discoideum* are deposited in several databases and some expression profiles were available as static graphs on dictyBase (http://www.dictybase.org). Curious biologists could view the data but cannot interact with them in any way. dictyExpress provides the community with a comprehensive database of all the expression data published by the functional genomics project at Baylor College of Medicine, as well as a web-based interface that can query the database and perform rather sophisticated data mining tasks. The interface can be easily expanded to include additional tools and adapted to the analysis of other public databases.

2 Results

dictyExpress (http://www.ailab.si/dictyexpress) includes over 3,600 expression array experiments from the BCM functional genomics group. It includes expression data collected from wild type and 14 mutant strains during normal growth and development and from cells subjected to various treatments. The web interface includes components for data retrieval, selection of individual genes or groups of genes, graphical display of gene expression time courses, gene ontology term enrichment, co-expression network construction and hierarchical clustering. The user can interact with the data in several ways. For example, entering the names of several genes of interest and the application returns their expression trajectories, Gene Ontology annotations and a clustering dendrogram with heat map. Alternatively, the user can select a gene ontology term and query the database for genes that match it. One can even hand-draw an approximate gene expression profile requesting dictyExpress to find genes that match it.

3 Conclusion

dictyExpress provides a public database of gene expression data for a popular model organism and original, yet simple means of data exploration. The system has been launched in June 2008 and a tight linkage with dictyBase has been implemented in April 2009. dictyExpress has an enormous value to the entire community. Currently, it allows researchers to explore a large database of expression arrays without having knowledge in data mining or access to powerful computers and expensive software. Eventually, we anticipate that other research communities will adopt this approach so that public databases of gene expression arrays would become readily available through easy to use, explorative interfaces.

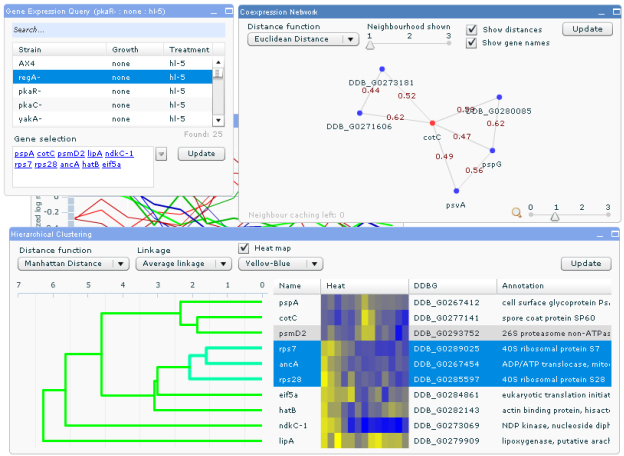


Figure 1: A screenshot of dictyExpress (experiment and gene selection, gene co-expression network, and clustering)

* denotes equal contribution
# Contact: e-mail: blaz.zupan@fri.uni-lj.si, phone: +386 4768 402, fax: +386 1 4264 647

Visualization of fragmented networks

Miha Stajdohar, Minca Mramor, Blaz Zupan, Janez Demsar
Faculty of Computer and Information Science
University of Ljubljana, Slovenia
miha.stajdohar@fri.uni-lj.si

Abstract

Biomedical data analysis often resorts to visualizations of observed items (genes, SNPs, proteins, ...) and their similarities. Two popular techniques include networks and multidimensional scaling (MDS). Networks require a threshold on similarity score and its optimization often fail in global placement of disconnected subnetworks. MDS can consider entire similarity matrices, does not require a similarity threshold, but its optimization is computationally expensive. We here propose a hybrid method, relying on network optimization for network components, which are then globally arranged by MDS. The proposed method is fast and can be used in optimization of very large networks.

1 Introduction

In modern biology we often deal with huge numbers of objects, such as genes, SNPs, data sets, genesets and proteins. To organize them into a structure with some discernible meaning, we often proceed by defining a measure of similarity between these objects. For instance, the similarity between genes can be defined by correlations in their expressions over various conditions, and the similarity between genesets can be based on the number of common genes they include.

Various methods can help us to organize such objects according to the measured similarities. From the onset of high-throughput genomics, clustering was a technique of choice for this task. But some clustering approaches, most notably k-means [Hartigan and Wong, 1979], are difficult to visualize and may offer no other information than merely lists of objects belonging to each of the inferred groups. Another deficiency of most clustering techniques is that two objects from different clusters can be more similar to each other than to most of those from their corresponding clusters.

Another popular approach for exploration of relations between large number of objects is based on visualization of networks. The objects of interest are shown as vertices, and similar objects are connected by edges. The "intelligent" part of the data exploration is left to the human expert who makes hypotheses from the layout of the network's vertices and through interactive analysis of its components. For this purpose, a number of graph layout optimization techniques are available within specialized tools and general data mining environments [Pavlopoulos *et al.*, 2008].

Networks may suffer from a similar problem as clustering: in most cases two objects are connected to each other if they are sufficiently similar with respect to the defined measure of similarity. The shape of the network depends on the chosen threshold, which is undesirable. Moreover, even if a pair of objects is just below the threshold of being sufficiently similar, it is treated as if it was completely dissimilar. Finally, unconnected components of the network are arranged randomly, disregarding any similarities which fell short of the threshold.

There are techniques which can elegantly avoid having to define appropriate thresholds. Multidimensional scaling (MDS) [Cox and Cox, 1994], for one, is a visualization method whose optimization algorithm is quite similar to that of typical graph drawing algorithms. MDS finds a projection of objects into a two dimensional space such that the distances between them correspond to the given similarities as closely as possible. The advantage of MDS over network visualization is that it does not require setting any thresholds: all pairwise distances are taken into account. At the same time, this is its disadvantage: since the time needed to compute the projection increases quadratically with the number of objects being plotted, the MDS becomes prohibitively slow when the number of objects goes into tens of thousands.

In this work we propose a new data analysis and visualization approach which combines networks and MDS. The proposed procedure breaks a possibly large network into a set of smaller components by increasing the required similarity threshold in the network construction. We call such networks *fragmented networks*. The layout of vertices in each network component is optimized separately using the Fruchterman-Rheingold [1991] algorithm, and the network components are then globally positioned using the MDS algorithm. To help the expert recognize the individual components, we use a simple text mining algorithm to find suitable labels for naming the individual components.

We start with a brief description of Fruchterman-Reingold's algorithm for the graph layout of optimization, and continue with description of the MDS algorithm. We then describe the proposed method, and conclude the paper with a case study on the leukemia gene expression data set.

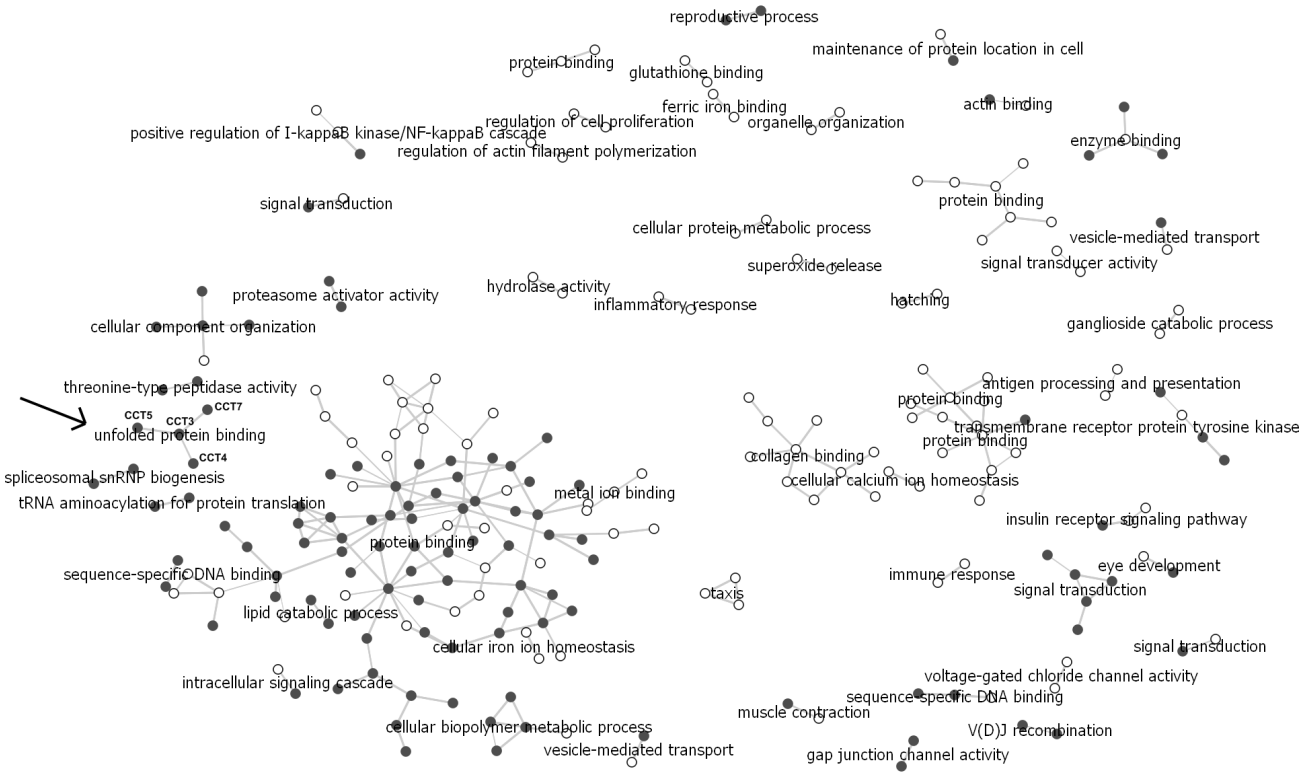


Figure 1: The network of the most differentially expressed genes from the leukemia data set. The similarity matrix of the chosen genes was taken from the recently published work of Huttenhower et al. The genes represented with solid circles were significantly over-expressed in acute lymphoblastic leukemia and the genes shown as empty circles had higher expression in acute myeloid leukemia.

2 Related work

2.1 Fruchterman-Reingold algorithm

Fruchterman-Reingold algorithm can find the optimal placements of vertices of a possibly large undirected graph. Each edge can also be assigned a weight. The algorithm can be described with a metaphor from physics: each vertex is a particle which repels other particles. The particles which correspond to connected vertices are, however, connected with a (non-linear) spring whose length is proportional to the weight of the edge. The algorithm simulates a physical process in which the particles are initially positioned randomly and then moved around to the configuration with the minimal potential energy.

In order to avoid ending up in a local minimum, the optimization uses simulated annealing [Kirkpatrick *et al.*, 1983] in which the particle can sometimes move into a "wrong direction", which increases the tension. As the system gradually cools down, the particles reach the stable configuration and the graph is typically well laid out.

The time complexity of the FR algorithm is linear with the number of edges, the number of vertices and the number of steps made.

2.2 Multidimensional scaling

The input for the MDS algorithm is a matrix of distances between pairs of objects. The algorithm finds the configuration of objects in a two-dimensional plot, such that the distances between them match the given distances as closely as possible.

The function which is usually optimized in the MDS is the so called "stress" of the system, expressed as

$$W = \sum_{i<j\leq n} w_{ij}(d_{ij} - \delta_{ij})^2 \qquad (1)$$

where d_{ij} is the actual distance and δ_{ij} is the desired distance between points v_i and v_j, and w_{ij} is the weight of the pair. For the sake of simplicity we usually assume equivalent weights for all pairs, *i.e.* $2w_{ij} = 1$.

Again, we can use concepts from classical physics to derive the MDS optimization procedure. Consider a single pair of points, v_i and v_j. The derivative of the corresponding term $W_{ij} = w_{ij}(d_{ij} - \delta_{ij})^2$ in the above sum,

$$F_{ij} = \frac{\partial W_{ij}}{\partial d_{ij}} = 2w_{ij}(d_{ij} - \delta_{ij}), \qquad (2)$$

is the force of the spring with a spring constant $k = 2w_{ij}$ displaced by $d_{ij} - \delta_{ij}$.

A simple steepest descent optimization procedure for MDS would mimic the physical system in which each vertex v_i would move by $\sum_{j\neq i} F_{ij}$ at each step (assuming that the system has infinite friction so that the acceleration cannot accumulate). There exist, however, better approaches to MDS, such as the SMACOF algorithm [de Leeuw and Mair, 2008].

The time complexity of the MDS algorithm is quadratic with the number of vertices. To plot a layout for 10.000 genes, each step of optimization would require computing 100 million distances and corresponding forces.

In comparison with the FR algorithm, the distances between objects in MDS more closely represent the actual distances. In both methods, a pair of similar points can be incidentally drawn far from each other if the system gets stuck in a local minimum or if the data cannot be accurately presented in two dimensions.

3 Methods

The input for the proposed method is a set of objects and a matrix of their mutual (dis)similarities. We would like to plot the objects in a two-dimensional space such that similar objects would be positioned close to each other. The outline of the procedure we propose is:

1. Construct a fragmented graph based on the matrix of similarities. Individual fragments – components – should (ideally) share some concepts in problem domain, that is, their composition should be meaningful to the domain expert.
2. Use the Fruchterman-Reingold or another similar algorithm to optimize the layout of the objects within each individual component.
3. Use multidimensional scaling to globally arrange (position, orientation) the components according to the provided matrix of similarities.

The first step is rather trivial and the second step is based on standard graph drawing algorithms. We therefore concentrate on the third one. The third step treats the components as rigid objects and its task is to move them around and rotate them so that the distances between vertices belonging to different components match the given distances as closely as possible. For a physical metaphor, imagine each component as a board with vertices as pegs. Pegs from different components are connected to each other with springs of different lengths. The nature (or, in our case, a computer simulation) finds the lowest energy configuration of the boards (components).

Formally, we assume that we are given a graph $\mathcal{G} = (V, E)$ constituted of p disjunct components $V = \bigcup_{k=1}^{p} V_k$. The task is to find the placement and the orientation of components, that is, their mass centers $\mathbf{c_k}$, and the orientation of their coordinate systems $\boldsymbol{\phi_k}$.

We will describe two algorithms. The first is based on the exact simulation of the physical system, and the second sacrifices exactness for a substantial speed improvement. To simplify the notation, we shall use indices i and j to denote quantities related to vertices (*e.g.* $\mathbf{F_i}$) and k and l to denote quantities related to entire components (*e.g.* $\mathbf{F_k}$).

3.1 Exact simulation

We start the simulation with random arrangement of the components, described by the position of their mass centers, $\boldsymbol{c_k}$ and orientations of coordinate systems, $\boldsymbol{\phi_k}$. At each step, the algorithm computes the forces acting on each vertex. It then computes the sum of forces and torques for each component and moves and rotates it accordingly.

The mass of each component and its moment of inertia do not change during the simulation, so we can compute them in advance. Assuming that all points have equivalent mass m, we have

$$m_k = |V_k| m \tag{3}$$

and

$$I_k = m \sum_{v_i \in V_k} ||\mathbf{v_i} - \boldsymbol{c_k}||^2, \tag{4}$$

where $\mathbf{v_i}$ is the position of vertex v_i and $\mathbf{c_k}$ is the position of the component's mass center.

Instead of computing the force by definition (2), we execute a single step of the more efficient SMACOF-based MDS optimization, which gives us a set of new positions of vertices, $\mathbf{v'_i}$. For each point we then compute the force which would move it by the same distance and direction, $\mathbf{v'_i} - \mathbf{v_i}$, as SMACOF had.[1]

$$\mathbf{F_i} = \frac{2m(\mathbf{v'_i} - \mathbf{v_i})}{t^2} \tag{5}$$

The component (that is, its mass center) moves according to the sum of the forces acting on its vertices

$$\boldsymbol{\Delta c_k} = \frac{\mathbf{F_k} t^2}{2m_k} = \frac{t^2 \sum_{v_i \in V_k} \mathbf{F_i}}{2|V_k|m} = \frac{\sum_{v_i \in V_k} (\mathbf{v'_i} - \mathbf{v_i})}{|V_k|} \tag{6}$$

To compute the rotation of the component, we need to compute the sum of torques. The torque for each vertex v_i equals $\boldsymbol{\tau_i} = \mathbf{F_i} \times \mathbf{v_i}$, so the total torque acting on the component is $\boldsymbol{\tau_k} = \sum_{v_i \in V_k} \mathbf{F_i} \times \mathbf{v_i}$. The rotation of the component V_k equals[2]

$$\boldsymbol{\Delta\phi_k} = \frac{\boldsymbol{\tau_k} t^2}{2I_k} = \frac{t^2 \sum_{v_i \in V_k} \mathbf{F_i} \times \mathbf{v_i}}{2m \sum_{v_i \in V_k} ||\mathbf{v_i}||^2} = \frac{\sum_{v_i \in V_k} \mathbf{v'_i} \times \mathbf{v_i}}{\sum_{v_i \in V_k} ||\mathbf{v_i}||^2} \tag{7}$$

The exact simulation algorithm repeatedly computes the translation ($\boldsymbol{\Delta c_k}$) and rotation ($\boldsymbol{\Delta\phi_k}$) of all components, and moves and rotates them accordingly, until reaching the (local) minimum ($\boldsymbol{\Delta c_k} \doteq 0$ and $\boldsymbol{\Delta\phi_k} \doteq 0$ for all k).

The time complexity of each step of the algorithm is dominated by that of the MDS, that is, $O(|V|^2)$.

3.2 Fast approximation

For a more efficient algorithm, we break the optimization into two phases: we first find the optimal positions of components in their initial orientations and in the second phase we optimize their orientations in these positions.

To speed-up the first phase we compute the MDS for positions of components, not of vertices as in the exact simulation. The desired distances between the components are the average distances between the corresponding vertices.

$$\delta_{kl} = \frac{1}{|V_k||V_l|} \sum_{\substack{v_i \in V_k \\ v_j \in V_l}} \delta_{ij} \tag{8}$$

[1] The derivation is based on elementary physics, $F = ma$ and $s = at^2/2$, hence $F = 2ms/t^2$ or $s = Ft^2/2m$. The chosen mass m and time t are inconsequential since they get canceled out. The model digresses from the actual physics in that objects stop when the force applied to them ceases, so the system stops at the local minimum instead of oscillating around it.

[2] From rotational dynamics, $\boldsymbol{\tau} = I\boldsymbol{\alpha}$ and $\boldsymbol{\Delta\phi} = \boldsymbol{\alpha}t^2/2$, so $\boldsymbol{\Delta\phi} = \boldsymbol{\tau}t^2/2I$. Note also that $(\mathbf{v'_i} - \mathbf{v_i}) \times \mathbf{v_i} = \mathbf{v'_i} \times \mathbf{v_i}$

This way of computing distances between components is inspired by the average linkage in hierarchical clustering analysis [Sokal and Michener, 1958].

In the second phase we apply the same procedure as in the exact simulation, except that we only compute the rotation without the translation. We noted that the optimization algorithm is quite likely to end up in local minima, so we use simulated annealing where the component can also rotate in the "wrong direction", with the probability of doing so decreasing with time.

This algorithm's time complexity is formally again quadratic in the number of vertices. The first step with such complexity is computation of average distances between components, and the second is the rotation phase of the optimization. In practice, these two operations are rather fast, so the algorithm is essentially quadratic only in the number of components, $O(p^2)$, due to the MDS-based optimization of their position in the first phase.

4 Case study

We have tested our hybrid algorithm for the visualization of fragmented networks on a leukemia gene expression data set [Golub *et al.*, 1999]. Our goal was to obtain clear visual representations of the most important genes and their biological functions for two major types of acute leukemia, yielding insight and valuable clues about the disrupted biological processes and pathways in leukemic cells.

4.1 Data

The leukemia gene expression data set [Golub *et al.*, 1999] includes the information on 7074 genes whose expression was measured using DNA microarrays in 72 tissue samples belonging to two distinct classes of acute leukemia (48 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples).

We built different networks from the most differentially expressed genes between the two classes of leukemia. Approximately 1000 genes were used that had the Student t-statistic significantly smaller or larger (p-value < 0.01) with respect to the null distribution of the t-statistic. The null distribution was obtained by randomly permuting the class labels and calculating the t-statistic for all the genes. In all the figures, the genes represented with solid circles were significantly over-expressed in the ALL samples and the genes shown as empty circles had higher expression in the AML samples.

4.2 Similarity scores

Two in principle different views were used to assess the similarity between the genes. In the first one, the similarity of the genes relates to their biological functions. It is calculated based on their membership in canonical biological pathways using the Jaccard index. The canonical pathways part of the C2 collection of the MsigDB [Subramanian *et al.*, 2005] was used where genes with roles in the same biological pathways are grouped into gene sets. Figure 2 represents a network of leukemia specific genes where the threshold of "biological function similarity" was set to 0.3 and only the vertices with at least one edge are shown.

The similarity between the genes can also be assessed based on their expression across different tissues, physiological states and diseases. The recently published work of Huttenhower et al. [2009] exploits this information to calculate the similarity matrix of all human genes. In their work, the similarity scores are calculated using the information on all publicly available gene expression and protein interaction data, combined with prior knowledge from the Gene Ontology, KEGG, HPRD and other biological data bases. We used their "gene expression similarity score" to build the network of the selected genes shown in Figure 1 and Figure 3.

4.3 Labels of network components

The graph components are named based on gene ontology terms [Ashburner *et al.*, 2000]. Each component of the graph is analyzed with respect to the biological process and molecular function of gene ontology. The ontological terms are scored in such a way that the coverage of the genes in the component, the p-value of the term enrichment and the number of all the genes in the ontology term are considered. The name of the component, presented in the graph, is chosen as the term with the highest score.

For example, in Figure 1 we can observe that the component marked with an arrow is comprised of genes CCT3, CCT4, CCT5, and CCT7. These genes are the gamma, delta, epsilon, and eta subunits of the chaperonin containing TCP1 complex (CCT), also known as the TCP1 ring complex (TRiC). Chaperonins are key components of the cellular chaperone machinery that is essential for the folding of the vast majority of cellular proteins. The TCP-1 ring complex is indispensable for cell survival because the folding of an essential subset of cytosolic proteins requires TRiC [Spiess *et al.*, 2004]. The component on the graph, comprised of these 4 genes is named "unfolded protein binding". Since all four genes from the component bind to unfolded proteins, out of all together 106 genes that have this molecular function, the p-value for the enrichment of this ontological term is very small. This ontological term covers all the genes from the component (4), has a small number of genes assigned to it (106), and a very small corresponding p-value of enrichment. Therefore our algorithm for naming components scores it with the highest score and the component is named after it.

4.4 Results and discussion

Figure 1 and Figure 2 represent two networks of the most differentially expressed genes from the leukemia data set, where different similarity scores of the genes (described in 4.2) are used to construct them. Both networks are fragmented into smaller components that are named according to the function of the genes comprising them. One can observe that most of the graph components connect genes that are over-expressed in one of the two investigated kinds of leukemia (all genes in the component are the same color), demonstrating the well known phenomenon that not only individual genes, but whole processes and pathways are disrupted in cancer cells [Hanahan and Weinberg, 2000].

To allow for the exploration of the biological processes connected to acute myeloid and acute lymhoblastic leukemia on different levels, from specific to more general,

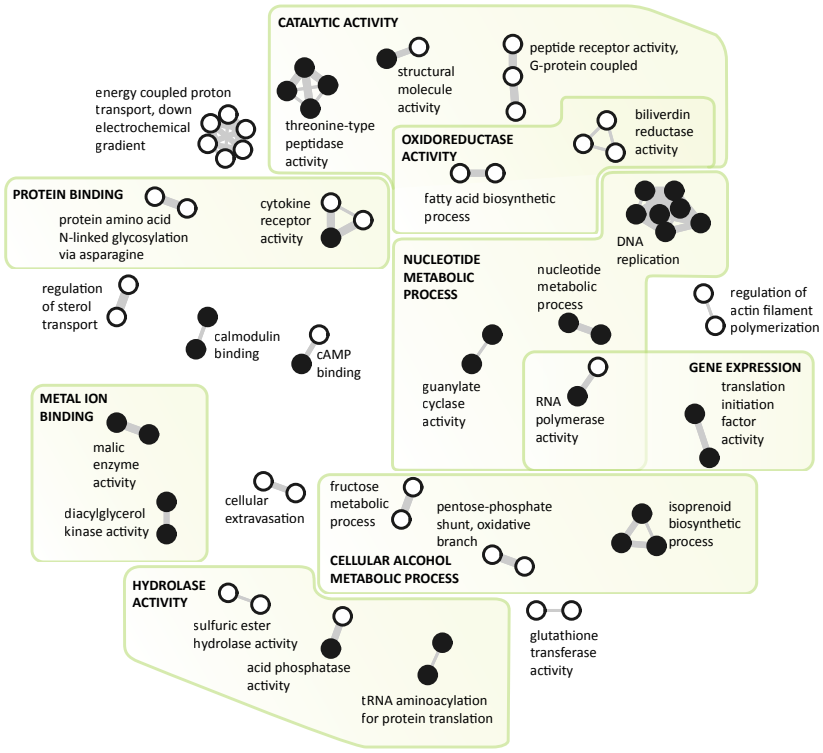


Figure 2: The network of the significantly differentially expressed genes form the leukemia data set, where the distance between the chosen genes was calculated based on their membership in biological pathways. Only the genes connected into components with a similarity score greater than 0.7 are shown. The genes represented with solid circles were significantly over-expressed in the ALL samples and the genes shown as empty circles had higher expression in the AML samples. The individual graph components and clusters of graph components are named with gene ontology terms as described in the text.

additional gene ontology terms describing larger parts of the graph are shown in Figure 2. These ontological terms apply to all the genes in the marked areas and are significantly enriched with a p-value < 0.01. One can observe how the components of the graph that are near each other have similar biological and/or molecular functions according to gene ontology.

Interestingly, the "guanylate cyclase activity", "nucleotide metabolic process", "RNA polymerase activity", and "DNA replication" components all connect genes significantly over-expressed in acute lympoblastic leukemia. All of these genes have a function in nucleotide metabolism and DNA biosynthesis. It is well known that lymphoblastic cells typically have severalfold higher activity of enzymes responsible for nucleotide metabolism enabling excessive proliferation of transformed cells [Scholar and Calabresi, 1973]. Moreover, some of the pathways active in nucleotide metabolism, for example de novo purine synthesis (DNPS), have been recognized as important targets of antileukemic agents (eg. methotrexate, mercaptopurine). In combination with other therapeutical agents, these drugs have improved survival of children with ALL to an overall cure rate of approximately 80 percent [Pui and Evans, 2006]. The network shown in Figure 2 clearly demonstrates this characteristic of acute lymphoblastic leukemia.

Figure 3 shows the same network as Figure 1, where also the unconnected vertices (genes not connected into graph components) are added, to observe the similarity of all the 1000 selected genes. It demonstrates the ability of our algorithm to combine network and MDS characteristics to visualize connected components and single genes simultaneously. One can observe that the genes significantly differentially expressed in the two investigated leukemias cluster together. The empty circles (AML) are clustered in the right part of the graph and the solid ones (ALL) in the left part, again demonstrating that expression changes in cancer tissues are disrupted on the level of pathways and processes.

5 Conclusion

Recently, we have witnessed the emergence of large repositories of biomedical research and clinical data. Methods to sip through these data sets that would allow domain experts to gather information, reason on the hidden patterns and form plausible hypothesis to be tested in specialized studies are needed. Here, visualization combined with visual data analytics plays a major role, as it can nicely reveal the data patterns and allow the experts to explore the data.

Visualizations require algorithms that craft the proper placement of the object under consideration, and explorative data analysis requests these to be fast to be able to construct responsive interfaces. We have developed a hybrid network layout optimization technique that addresses these requirements, and show its utility in the analysis of the microarray data set. Initial exploration of the methods were positive and encouraged us to embed it within the network optimization and visualization components of Orange [Demsar *et al.*, 2004], an open-source data mining and visual analytics framework available at http://www.ailab.si/orange.

Acknowledgments

This work was supported by grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

Figure 3: The same network as in Figure 1 where also the unconnected vertices (genes) are shown.

References

[Ashburner *et al.*, 2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[Cox and Cox, 1994] Trevor F. Cox and Michael A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, har/dis edition, January 1994.

[de Leeuw and Mair, 2008] Jan de Leeuw and Patrick Mair. Multidimensional scaling using majorization: Smacof in r. *Department of Statistics, UCLA. Department of Statistics Papers*, 2008.

[Demsar *et al.*, 2004] J. Demsar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining. Technical report, Faculty of Computer and Information Science, University of Ljubljana, 2004.

[Fruchterman and Reingold, 1991] T M J Fruchterman and E M Reingold. Graph drawing by force-directed placement. software practice and experience, 1991.

[Golub *et al.*, 1999] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

[Hanahan and Weinberg, 2000] Douglas Hanahan and Robert A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57 – 70, 2000.

[Hartigan and Wong, 1979] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.

[Huttenhower *et al.*, 2009] Curtis Huttenhower, Erin M. Haley, Matthew A. Hibbs, Vanessa Dumeaux, Daniel R. Barrett, Hilary A. Coller, and Olga G. Troyanskaya. Exploring the human genome with functional maps. *Genome Research*, pages –, 2009.

[Kirkpatrick *et al.*, 1983] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[Pavlopoulos *et al.*, 2008] Georgios Pavlopoulos, Anna-Lynn Wegener, and Reinhard Schneider. A survey of visualization tools for biological network analysis. *BioData Mining*, 1(1):12, 2008.

[Pui and Evans, 2006] Ching-Hon Pui and William E. Evans. Treatment of Acute Lymphoblastic Leukemia. *N Engl J Med*, 354(2):166–178, 2006.

[Scholar and Calabresi, 1973] E. M. Scholar and P. Calabresi. Identification of the Enzymatic Pathways of Nucleotide Metabolism in Human Lymphocytes and Leukemia Cells. *Cancer Res*, 33(1):94–103, 1973.

[Sokal and Michener, 1958] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

[Spiess *et al.*, 2004] Christoph Spiess, Anne S. Meyer, Stefanie Reissmann, and Judith Frydman. Mechanism of the eukaryotic chaperonin: protein folding in the chamber of secrets. *Trends in Cell Biology*, 14(11):598 – 604, 2004.

[Subramanian *et al.*, 2005] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. From the cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, October 2005.

CNET: an algorithm for the inference of gene regulatory interactions from gene expression time series

Francesco Sambo*, **Barbara Di Camillo, Marco Falda, Gianna Toffolo and Silvana Badaloni**
Department of Information Engineering
University of Padova, Italy

Abstract

The process of inferring regulatory interactions among genes from DNA microarray experiments is known as Reverse Engineering. We present a novel reverse engineering algorithm, CNET, which exploits the principles of Shannon Entropy and Mutual Information through a heuristic scoring function.

This function is designed to discover causal relations even if gene temporal profiles exhibit non ideal behaviours, such as noise, quantization errors and variable regulatory delays.

Experimental results, both on simulated and on real datasets, show that CNET achieves performance comparable to the state of the art methods in reverse engineering and doubles its performance when inferring on regulatory effects directly dependent on a perturbed target.

Introduction

One of the most important discoveries of the last century in biology is that all the information necessary for an organism to live is coded in the genes of its DNA. On the other hand, the certainty emerged that almost every biological function is carried over by proteins. DNA molecules are *transcribed* into mRNA molecules, which, in turn, direct chemical machinery which *translates* the nucleic acid message into a protein [Hunter, 2004].

Some proteins, called *transcription factors*, have the role, possibly in combination with each other, to activate or inhibit the transcription of genes and to control the translation of mRNA into new proteins; the process by which genes, through the proteins they code, control the expression (i. e. the mRNA transcription rate) of other genes is known as *genetic regulation*.

In the past few years, the study of genetic regulation was drastically improved by the discovery of the new technology of DNA microarray [Molla *et al.*, 2004], which allows researchers to monitor the expression of the whole genome under various genetic, chemical and environmental perturbations. One of the goals of microarray experiments is to understand the mechanism of genetic regulation, which can be modelled as a *gene regulatory network*, a graph in which nodes represent genes or proteins and two or more nodes are connected if a regulatory relation exists between them.

Different approaches have been adopted in the literature to model and infer Gene Regulatory Networks from DNA microarray experiments (for a survey see [de Jong, 2002]) and some assessment papers have recently appeared, to compare the performance of various algorithms under different experimental conditions [Bansal *et al.*, 2007] [Corradin *et al.*, 2008] [Soranzo *et al.*, 2007].

A widely used approach to infer regulatory relations is the analysis of the Shannon Entropy and Mutual information of gene expression signals, proposed by Liang *et al.* in the REVEAL algorithm [Liang *et al.*, 1998]. In this work, we propose an extension of the REVEAL approach, introducing a regulatory network model, *Causal Networks*, and a scoring function for regulatory relations. The scoring function allows our approach to account for inconsistencies in gene expression time series caused by variable regulatory delays, measurement noise and quantization errors. We compare our algorithm, CNET, with the original REVEAL algorithm and with Dynamic Bayesian Networks (DBNs) [Ferrazzi *et al.*, 2007], an approach which was identified as promising in the three assessment papers [Bansal *et al.*, 2007], [Corradin *et al.*, 2008] and [Soranzo *et al.*, 2007].

Comparisons are carried over on a realistically simulated data set, obtained with the simulator described in [Di Camillo *et al.*, 2009]: simulated data are useful when comparing different algorithms, because of the lack of information on real regulatory networks, which prevents one from obtaining both reliable test beds from real biological experiments and multiple cases to evaluate average performance [Soranzo *et al.*, 2007]. We also run our algorithm on a real dataset [Whitfield *et al.*, 2002], consisting on the observation during cell cycle of 9 human genes.

The rest of this paper is organized as follows: Section 1 presents REVEAL algorithm and related works, Section 2 describes the Causal Network model and CNET algorithm, Section 3 shows experimental results

*Corresponding author: francesco.sambo@dei.unipd.it

and Section 4 presents conclusions.

1 REVEAL and related works

In the Boolean model, adopted by the REVEAL algorithm [Liang *et al.*, 1998], regulatory relations are organized in a directed graph, in which nodes represent genes and edges represent boolean relations among genes. Nodes can be in two possible discrete states $\{0, 1\}$ and gene interactions are synchronous: the whole state of the genome at time t is completely determined by its state at time $t - \Delta$, where Δ is a constant time step.

A boolean relation between a set of regulators $(x_1, \ldots, x_K)$ and a regulated gene x_0 is inferred if the Mutual Information between the regulators and x_0 is equal to the Shannon Entropy of x_0 or, equivalently, if the joint Shannon Entropies of the sets $(x_1, \ldots, x_K)$ and $(x_1, \ldots, x_K, x_0)$ are the same. This condition, however, tends to be impractical when applied to real data sets, in which behaviours of gene profiles are usually far from ideal: data contains noise, genes have variable regulatory delays and quantization introduces errors.

Two approaches to extend Mutual Information to continuous gene expression values, thus avoiding the loss of information related to quantization, have been proposed by Margolin *et al.* [Margolin *et al.*, 2006] and Daub *et al.* [Daub *et al.*, 2004]. Both, however, are defined as pairwise measures between gene profiles and then lose the ability of inferring complex N to 1 relations. Moreover, they are usually adopted as measures between synchronous observations, rather than time series of gene expression, because they do not specifically consider the ordering of the samples.

2 The Causal Network model and CNET algorithm

2.1 Model

The Causal Network model is a generalization of the Boolean model: nodes represent genes and edges represent causal regulatory interactions among genes. Nodes can be in three possible discrete states $\{1, 0, -1\}$, corresponding to high, medium and low (or increasing, steady and decreasing) expression level.

To adapt our algorithm to the behaviour of real gene profiles, which comprehends noise and variable regulatory delays, and to limit the effect of the error introduced by the quantization, we decided to map the REVEAL condition on Shannon Entropies to the domain of *consistent pairs*: for signals with a finite set of possible values (*quantized signals*), the pair $\langle regulators, regulated\ signal \rangle$ is said to be consistent if and only if each combination of values for the regulators univocally correponds to a particular value for the regulated signal after Δ time steps. In Appendix A, we prove mathematically that the REVEAL condition is verified if and only if $\langle regulators, regulated\ signal \rangle$ is a consistent pair.

x_1	1	1	1	0	0	0	0	0	1	1	1	1	1	1
x_2	-1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	-1	-1
x_0	0	0	0	0	0	-1	-1	-1	-1	-1	0	0	0	0
Block Length	BL_1= 3			BL_2= 5					BL_3= 6					
Steady State Length	SSL_1= 3			SSL_2=3					SSL_3=4					

Figure 1: Example of a causal relation $(x_1, x_2) \Rightarrow x_0$. Profiles are already aligned properly.

This switch of domains allowed us to design a novel heuristic function for regulatory pairs: each term of the function is ment to capture a particular aspect of gene expression profiles and the ensemble of terms gives an indication of how far the regulatory pair is from being consistent, thus inducing an ordering among regulatory pairs. For a given regulated signal, then, the particular combination of regulators that maximizes the scoring function can be searched.

2.2 Scoring function

For every gene x_0, the algorithm searches extensively for the best set of k regulators $(x_1 \ldots x_k)$ that maximizes a *scoring function* f:

$$f = w_e \frac{1}{1+e} + w_s s + w_c \frac{c}{3^k} \quad \text{with} \quad w_e + w_s + w_c = 1 \ . \tag{1}$$

f ranges in (0,1] and combines the contribution of an error term e, a shape term s and a completeness term c, by weighting them with parameters w_e, w_s and w_c. Each term is explained in what follows.

Error Term

To illustrate the meaning of $1/(1+e)$, we refer, without loss of generality, to the case with two regulators x_1 and x_2 for a regulated gene x_0: $(x_1, x_2) \Rightarrow x_0$. The error term e is defined as:

$$e = \sum_{i=-1}^{1} \sum_{j=-1}^{1} e_{ij} = \sum_{i=-1}^{1} \sum_{j=-1}^{1} \frac{Lp_{ij} - Mp_{ij}}{Lp_{ij}} \ , \tag{2}$$

where Lp_{ij} is the number of occurrences of the input pattern $\langle i, j \rangle$ (*e.g.* $\langle 1, -1 \rangle$ in Figure 1 occurs 9 times, then $Lp_{1,-1} = 9$), and Mp_{ij} is the value of gene x_0 that is most frequent as output in correspondence to the input $\langle i, j \rangle$ (*e.g.* in Figure 1, for the input pattern $\langle 1, -1 \rangle$, the value $x_0 = 0$, occurring 7 times, is more frequent than -1, occuring twice, thus $Mp_{1,-1} = 7$).

If there is a univocal correspondence between input and output profiles (and then the pair $\langle (x_1, x_2); x_0 \rangle$ is consistent), e_{ij} are equal to zero, thus $1/(1 + e) = 1$, otherwise $1/(1 + e) < 1$. This allows to relax the consistency condition and tolerate a certain amount of noise and quantization errors in the data.

Shape term

The shape term is calculated based on data organization into blocks of equal input combinations, such as

the three blocks identified by vertical bars in Figure 1. The shape term s is computed as

$$s = \frac{1}{\#blocks} \sum_{i=1}^{\#blocks} s_i = \frac{1}{\#blocks} \sum_{i=1}^{\#blocks} \frac{SSL_i}{BL_i} \quad (3)$$

where BL_i is the length of the i-th block and SSL_i is the length of the rightmost substring of identical characters in the i-th block of x_0.

For example, in Figure 1, the shape terms for the three blocks are 1 (3/3), 0.6 (3/5) and 0.67 (4/6), thus leading to an average shape term of 0.76.

The shape term s is similar to e, *i.e.* it relaxes the consistency condition, but it assigns lighter penalties to output inconsistencies occurring right after a change of state in regulators; it then rewards situations in which the output signal, after a change in the input, shows a transient state followed by a steady state. This terms helps the algorithm to capture regulatory relations even in the presence of regulatory delays variable from gene to gene and longer than the fixed value Δ.

Completeness term

The completeness term $c/3^k$ is the normalized number of different combinations of values for regulators present in data ($1 \leq c \leq 3^k$, if k is the size of the set of regulators). For example, in Figure 1 two combination of values, $\langle 1, -1 \rangle$ and $\langle 0, 1 \rangle$, are present in input, then $c/3^k = 2/9$. This term induces the algorithm to prefer simpler solutions, *i.e.* solutions with less regulators, the other two terms being equal.

2.3 Algorithm

Pseudocode for CNET algorithm is as follows:

```
CNET(data, max_causes)
 1  for i ← 0 to n_genes
 2  do max_fitness[i] ← 0
 3     C[i] ← ∅
 4     for k ← 1 to max_causes
 5        do for causes in combinations(k, n_genes)
 6           do f ← fitness(i, causes)
 7              if f = max_fitness[i]
 8              then
 9                   C[i] ← C[i] ∪ causes
10              if f > max_fitness[i]
11              then
12                   C[i] ← causes
13                   max_fitness[i] ← f
14     if C[i] > 1
15     then weight each cause proportionally to
            the number of times it appears in C[i]
16  return C
```

For each gene, CNET algorithm searches extensively for all the possible combinations of regulators, from one to a maximum user defined number (max_causes), and keeps track of the best scoring combinations. If more than one set of regulators for the same gene achieve the best score, the weight of each regulator is set proportional to the number of times it appears among the best scoring sets.

3 Experimental Results

We tested CNET first on a simulated dataset, to compare its performance with the REVEAL algorithm and with Dynamic Bayesian Networks (DBNs), one of the best approaches for reverse engineering gene regulatory networks from time series data. DBNs are inferred with the greedy search K2 algorithm as described in [Ferrazzi *et al.*, 2007], to which we refer the reader for futher information on the particular implementation. We then validate performance on a real microarray dataset of 9 human genes. Datasets and performance measures are explained in what follows.

3.1 Experimental data set

Simulated data consist of 60 networks of 10 genes and 60 networks of 20 genes, generated with the simulator recently presented in [Di Camillo *et al.*, 2009]: network topology is generated according to the current knowledge of biological network organization, including scale-free distribution of the connectivity and clustering coefficient independent of the number of nodes in the network.

The simulator explicitly represents interactions among the regulators of each gene and, by using differential equations, accounts for saturation in the response to regulation, transcription activation thresholds and shows robustness to perturbations. It implements the possibility to observe gene dynamics by either letting the system free to evolve from opportunely chosen initial conditions (*free evolution*) or exciting it by external stimuli acting on chosen nodes.

For each network, 4 different time series of 50 samples were generated: the first time series is obtained observing free evolution from random initial conditions, the other three time series are obtained stimulating the network with a sinusoid, a ramp and a step signal respectively. Networks are stimulated at their *hub*, *i.e.* the node with the highest out degree, to excite the highest number of nodes in the network.

The real dataset, on the other hand, consists of 9 genes involved in human cell cycle, for which samples were taken every hour for 47 hours (approximately three complete cell cycles). To measure the performance of CNET, we compared the output of the algorithm with the interactions documented in the BioGRID database (www.thebiogrid.org). The dataset has already been used by Sacchi *et al.* [Sacchi *et al.*, 2007] to test the accuracy of a method for extracting temporal relationships between genes.

3.2 Performance Measures

To quantify the overall performance of the algorithms, we adopted two widely used measures, Precision (P) and Recall (R), defined as: $P = tp/(tp + fp)$ and $R = tp/(tp + fn)$, where tp is the number of true positives, *i.e.* the number of causal relations correctly identified by the algorithms, fp is the number of false positives, *i.e.* the number of relations identified by the algorithms which are not correct, and fn is the number of false negatives, *i.e.* the number of relations present

Algorithm	Measure	Sin	Step	Ramp	No stimulus
CNET	P	0.25 ± 0.08	0.20 ± 0.07	0.20 ± 0.08	0.17 ± 0.07
	R	0.38 ± 0.11	0.33 ± 0.13	0.34 ± 0.15	0.26 ± 0.11
REVEAL	P	0.20 ± 0.16	0.15 ± 0.23	0.19 ± 0.24	0.17 ± 0.22
	R	0.07 ± 0.06	0.04 ± 0.05	0.05 ± 0.05	0.05 ± 0.06
DBNs	P	0.19 ± 0.07	0.20 ± 0.07	0.19 ± 0.07	0.17 ± 0.08
	R	0.34 ± 0.13	0.36 ± 0.13	0.35 ± 0.14	0.31 ± 0.14

Algorithm	Measure	Sin	Step	Ramp	No stimulus
CNET	P	0.14 ± 0.05	0.10 ± 0.05	0.10 ± 0.04	0.09 ± 0.05
	R	0.21 ± 0.07	0.14 ± 0.08	0.14 ± 0.07	0.12 ± 0.07
REVEAL	P	0.10 ± 0.10	0.08 ± 0.11	0.09 ± 0.13	0.06 ± 0.08
	R	0.03 ± 0.03	0.02 ± 0.03	0.03 ± 0.03	0.02 ± 0.02
DBNs	P	0.13 ± 0.04	0.13 ± 0.04	0.12 ± 0.05	0.11 ± 0.04
	R	0.28 ± 0.10	0.28 ± 0.10	0.28 ± 0.10	0.24 ± 0.09

Table 1: Average Precision and Recall for REVEAL, CNET and DBNs on 60 networks of 10 genes (upper table) and 60 networks of 20 genes (lower table). Values are reported for each of the three external stimuli and for the case without stimulus, as mean ± standard deviation.

in the simulated networks but not identified. Both measures range in the interval $[0, 1]$.

Scores were compared using exact Wilcoxon two-sample tests: we considered as significant differences corresponding to a p-value < 0.05.

3.3 Results

Results of the comparison between REVEAL, CNET and DBNs are shown in Table 1. CNET significantly outperforms REVEAL both in terms of Precision an Recall, and its performance is comparable to DBNs: on networks of 10 genes, there is no significant difference between the two algorithms, whereas on 20 genes networks DBNs exhibit a significantly higher Recall. Precision and Recall are in line with the literature, being in the same range of values reported by [Corradin *et al.*, 2008] and [Soranzo *et al.*, 2007] for similar experiments. CNET seems to be more sensitive than DBNs to the kind of input signal, showing significantly higher performance when the network is stimulated with the sinusoidal signal. Both algorithms, however, show equal or higher performance when an external stimulus is present.

In microarray experiments, when a particular gene is externally stimulated, the interest usually focuses mostly on genes directly dependent from the target: for this reason, we analysed also the performance of CNET and DBNs on the inference of regulatory relations directly outgoing from the externally stimulated gene. Results are shown in Table 2. Performance on this subset of relations is more than doubled, and scales best when network size increases, both for CNET and DBNs. The sensitivity of CNET to a particular external stimulus is even more evident here: in the presence of a sinusoidal signal, CNET exhibits an average Recall of 0.80 on networks of 10 genes and of 0.65 on network of 20 genes.

Results of the CNET algorithm on the real dataset are shown in Figure 2. On the left side of the figure known relations between genes are plotted, while

Algorithm	Measure	Sin	Step	Ramp
CNET	P	0.50 ± 0.16	0.48 ± 0.29	0.52 ± 0.32
	R	0.80 ± 0.20	0.46 ± 0.29	0.48 ± 0.33
DBNs	P	0.60 ± 0.19	0.53 ± 0.25	0.55 ± 0.20
	R	0.52 ± 0.27	0.46 ± 0.27	0.49 ± 0.26

Algorithm	Measure	Sin	Step	Ramp
CNET	P	0.42 ± 0.18	0.33 ± 0.27	0.36 ± 0.33
	R	0.65 ± 0.23	0.29 ± 0.25	0.32 ± 0.29
DBNs	P	0.61 ± 0.22	0.47 ± 0.23	0.48 ± 0.26
	R	0.46 ± 0.25	0.38 ± 0.24	0.37 ± 0.25

Table 2: Precision and Recall for CNET and DBNs, on edges directly outgoing from the stimulated gene, for networks of 10 genes (upper table) and 20 genes (lower table).

on the right side the reconstructed network is represented: four of the original edges, depicted in bold, were correctly identified, and two other edges, $CCNE1 \rightarrow CDC2$ and $PCNA \rightarrow CDC2$, were identified reversed, thus leading to Precision and Recall of (0.36,0.44) on the oriented network and (0.54,0.67) on the unoriented network, which are in line with the results on simulated data. Precision and Recall of DBNs on the same dataset are (0.19,0.33) on the oriented network and (0.31,0.56) on the unoriented network.

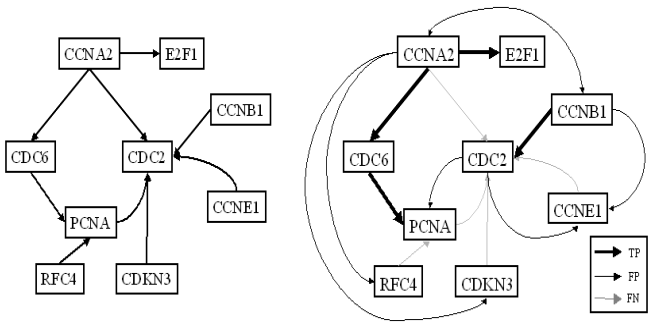


Figure 2: Network of known interactions (left) and reconstructed network (right) for the real dataset of 9 human genes related to cell cycle.

4 Conclusions

In this paper we presented a novel algorithm, CNET, for Reverse Engineering of Gene Regulatory Networks from time series data of DNA microarray experiments. Tests on simulated data showed that CNET outperforms the REVEAL algorithm, of which it can be considered an improvement, and exhibits performance comparable to a state-of-art approach, Dynamic Bayesian Networks, in reconstructing both entire networks and subsets of them close to external stimulation. Similar performance was exhibited on a real dataset.

Appendix A

In this appendix we prove mathematically the equivalence between the REVEAL condition for causal relations, based on Shannon Entropy, and our condition, on which the scoring function is based.

Definition 1 - The pair $\langle (X_1 \dots X_K); X_0 \rangle$ is *consistent* if, every time a given combination of values $(\overline{x}_1 \dots \overline{x}_K)$ appears for $(X_1 \dots X_K)$, the value of X_0 after Δ time steps is always the same.

Definition 2 - The *Shannon Entropy* for a sequence X of symbols x_i from an alphabet of size b is

$$H(X) = -\sum_{i=1}^{b} p(x_i) \log_b p(x_i)$$

where $p(x_i)$ is the probability of observing the particular symbol x_i.

Definition 3 - The *Joint Shannon Entropy* for the sequences $(X_1 \dots X_K)$ of symbols from an alphabet of size b, is

$$H(X_1, ..., X_K) = -\underbrace{\sum_{i=1}^{b} \dots \sum_{i=1}^{b}}_{K\ times} p(x_{1i}, \dots, x_{Ki}) \log_b p(x_{1i}, \dots, x_{Ki})$$

where $p(x_{1i}, \dots, x_{Ki})$ is the probability of observing simultaneously the particular combination of symbols $x_{1i}, \dots, x_{Ki}$ in sequences $X_1, \dots, X_K$.

Proposition 1 - The causal relation

$$X_1 \dots X_K \Rightarrow X_0 \tag{4}$$

holds if and only if

$$H(X_1 \dots X_K) = H(X_0, X_1 \dots X_K)$$

after a proper shift of sequence X_0, to account for the fixed delay Δ in the causal relation.

Proof - Proven in [Liang *et al.*, 1998].

Theorem 1 - The causal relation (4) holds if and only if the pair $\langle (X_1 \dots X_K); X_0 \rangle$ is consistent.

Proof - From Proposition 1, equation (4) holds if and only if

$$H(X_1 \dots X_K) = H(X_0, X_1 \dots X_K) \tag{5}$$

but then, from Definition 3

$$H(X_1, ..., X_K) = -\underbrace{\sum_{i=1}^{b} \dots \sum_{i=1}^{b}}_{K\ times} p(x_{1i}, \dots, x_{Ki}) \log_b p(x_{1i}, \dots, x_{Ki})$$

and, supposing w. l. o. g. that $X_i \in \{i, s, d\}$[1]

$$\begin{aligned} H(X_0, X_1 \dots X_K) = & \qquad (6) \\ & -\underbrace{\sum_{i=1}^{b} \dots \sum_{i=1}^{b}}_{K+1\ times} p(x_{0i}, \boldsymbol{x_i}) \log_b p(x_{0i}, \boldsymbol{x_i}) = \\ & -\underbrace{\sum_{i=1}^{b} \dots \sum_{i=1}^{b}}_{K\ times} p((x_0 = i) \wedge \boldsymbol{x_i}) \log p((x_0 = i) \wedge \boldsymbol{x_i}) \\ & -\underbrace{\sum_{i=1}^{b} \dots \sum_{i=1}^{b}}_{K\ times} p((x_0 = s) \wedge \boldsymbol{x_i}) \log p((x_0 = s) \wedge \boldsymbol{x_i}) \\ & -\underbrace{\sum_{i=1}^{b} \dots \sum_{i=1}^{b}}_{K\ times} p((x_0 = d) \wedge \boldsymbol{x_i}) \log p((x_0 = d) \wedge \boldsymbol{x_i}) \\ & \text{where } \boldsymbol{x_i} = (x_{1i}, \dots, x_{Ki}) \end{aligned}$$

If the pair $\langle (X_1 \dots X_K); X_0 \rangle$ is consistent, every time a given combination $(x_{1i} \dots x_{Ki})$ appears for $(X_1 \dots X_K)$, x_0 after Δ time steps has always the same value $\overline{x}$, being it *increasing*, *steady* or *decreasing*. Then,

$$p(x_{0i}, x_{1i}, \dots, x_{Ki}) = \begin{cases} p(x_{1i}, \dots, x_{Ki}) & \text{if } x_{0i} = \overline{x}; \\ 0 & \text{otherwise.} \end{cases}$$

And then, in equation (6), for every combination of values $(x_{1i} \dots x_{Ki})$ there is exactly one term in one of the three summations which is different from zero. And then equation (5) holds.

Acknowledgements We wish to thank Fulvia Ferrazzi, Lucia Sacchi and Riccardo Bellazzi from the Bio-Medical Informatics research group of the University of Pavia, for their kindly provision of the code of the Dynamic Bayesian Networks algorithm and of the Human Cell Cycle dataset.

References

[Bansal *et al.*, 2007] M. Bansal, V. Belcastro, Ambesi-Impiombato, and D. A., Di Bernardo. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 78(3), 2007.

[Corradin *et al.*, 2008] A. Corradin, Di Camillo, Toffolo B., and C G., Cobelli. In silico assessment of four reverse engineering algorithms: role of network complexity and multi-experiment design in network reconstruction and hub detection. In *ENFIN - DREAM Conference Assessment of Computational Methods in Systems Biology, April 28 - 29, 2008, Madrid*, 2008.

[Daub *et al.*, 2004] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska. Estimating mutual information

[1] Which stand for *increasing*, *steady* and *decreasing*.

using b-spline functions - an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(118), 2004.

[de Jong, 2002] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[Di Camillo *et al.*, 2009] B. Di Camillo, G. Toffolo, and C. Cobelli. A gene network simulator to assess reverse engineering algorithms. *Annals of the New York Academy of Sciences*, 1158:125–142, 2009.

[Ferrazzi *et al.*, 2007] F. Ferrazzi, P. Sebastiani, M. F. Ramoni, and R. Bellazzi. Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear gaussian networks. *BMC Bioinformatics*, 8 Suppl 5, 2007.

[Hunter, 2004] Lawrence Hunter. Life and its molecules: A brief introduction. *AI Magazine - Special issue on AI and Bioinformatics*, 25(1):9–22, 2004.

[Liang *et al.*, 1998] S. Liang, S. Fuhrman, and R. Somogyi. Reveal: a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, pages 18–29, 1998.

[Margolin *et al.*, 2006] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1, 2006.

[Molla *et al.*, 2004] M. Molla, M. Waddell, D. Page, and J. Shavlik. Using machine learning to design and interpret gene-expression microarrays. *AI Magazine - Special issue on AI and Bioinformatics*, 25(1):23–44, 2004.

[Sacchi *et al.*, 2007] L. Sacchi, C. Larizza, P. Magni, and R. Bellazzi. Precedence temporal networks to represent temporal relationships in gene expression data. *Journal of Biomedical Informatics*, 40(6):761–774, 2007.

[Soranzo *et al.*, 2007] Nicola Soranzo, Ginestra Bianconi, and Claudio Altafini. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, 23(13):1640–1647, July 2007.

[Whitfield *et al.*, 2002] M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 2002.

Data representation and mining using multi-layer networks

Lan Žagar Miha Štajdohar Janez Demšar Blaž Zupan
Faculty of Computer and Information Science
University of Ljubljana, Slovenia
{lan.zagar,miha.stajdohar,janez.demsar,blaz.zupan}@fri.uni-lj.si

Abstract

Standard network visualization techniques, often used in biomedical data analysis, usually display objects of a single type. Many biomedical applications consider objects of various types, and relations between them are often semantically different. To accommodate for such a diversity, we propose a data presentation using a multi-layer network, each layer including objects of the same type. We propose an algorithm to optimize the layout of such networks. We have used this approach to visualize the relations between genes and their publication-based MeSH term annotations. The resulting visualizations were found to be informative and could convey insights about the data that would not be easily extracted from standard, single-layer networks.

1 Introduction

A popular technique for organizing and presenting a set of objects and their relations are networks. Often, however, studying more than a single type of objects is desired. We here propose a data exploration and visualization approach based on construction, optimization and visualization of a multi-layer network. Each layer of such a network includes objects of one type. Objects are related to those from the same layer, and to objects at adjacent layers. An example of such a hypothetical two-layer network is shown on Fig. 1. In the paper, we propose a layout optimization algorithm for multi-layer networks, and present a case study, where the proposed method was used to relate genes and MeSH terms of their corresponding publications.

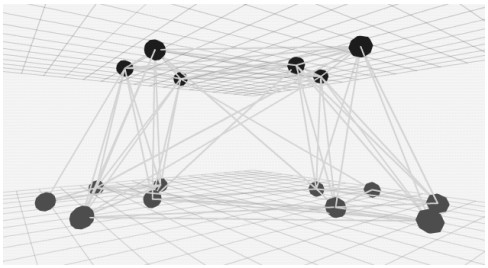


Figure 1: 3D visualization of a two-layer network.

2 Methods

We have developed a method for multi-layer network layout optimization based on an extension of a state-of-the-art network layout algorithm. Finding an appropriate network layout is critical for effective visualizations, since it helps to reveal inherent network structure. If vertices were to be positioned randomly, the resulting visualization would be incomprehensible for all but the simplest networks. We also wish to identify vertices, which are in the particular layout of the network well associated to their neighbors, and provide means to emphasize them in the visualization.

For optimization of multi-layer networks, we have adapted the Fruchterman-Reingold algorithm [Fruchterman and Reingold, 1991], one of the best known representatives from the force-directed family of methods for graph drawing [Tollis *et al.*, 1998]. In this algorithm, the connected vertices attract each other, and all vertices — connected or unconnected — repel each other. The algorithm starts with an arbitrary placement of vertices and iteratively moves them based on the sum of forces acting on them. The degree of relative movements of the vertices is decreased with time, using the simulated annealing approach [Davidson and Harel, 1996].

Our modification (Eq 1–3) for multi-layer networks retains the general structure of the Fruchterman-Reingold algorithm, but alters the definitions of attractive (f_a) and repulsive (f_r) forces as well as of the coefficient k, that controls the equilibrium distance between vertices. A new parameter λ is introduced in the definition of the latter, through which the influence of forces between layers can be adjusted with respect to forces within layers. In the equations $d(v,u)$ refers to the Euclidean distance between drawn vertices v and u, while $area$ refers to the area of the whole drawing surface.

$$k_{i,j} = \begin{cases} \sqrt{\frac{area}{|V_i|}} & \text{if } i = j \\ \lambda * \frac{k_{i,i}+k_{j,j}}{2} & \text{if } i \neq j \end{cases} \quad (1)$$

$$f_a(v,u) = \frac{d(v,u)^2}{k_{i,j}} \quad v \in V_i,\ u \in V_j \quad (2)$$

$$f_r(v,u) = \begin{cases} -\frac{k_{i,i}^2}{d(v,u)} & \text{if } v,u \in V_i \\ 0 & \text{if } v \in V_i,\ u \in V_j,\ i \neq j \end{cases} \quad (3)$$

In a multi-layer network visualization, we are interested how concepts from one layer associate to concepts in the

adjacent layer. To identify objects where such association is strong, we score the placement of vertex $v \in V_i$ from layer i with respect to its neighbors in layer j. We assess the proximity of the latter by computing the median Euclidean distance to them. This distance is then normalized with the median distance between vertices from layers i and j. In our network visualizations, the score of vertices is denoted through the size of the points.

3 Experimental study

We here present a simple demonstration of utility of the proposed procedure. In an experimental study, we have considered genes from a social amoeba *Dictyostelium discoideum*. The network consisted of two layers, one holding the genes and the other their associated Medical Subject Headings (MeSH) descriptors annotated to the papers, where these genes have been cited.

Data sets and network structure

A central web site dedicated to *D. discoideum* is called dictyBase [1] and among other maintains a data file with a list of references (PubMed identifiers) associated to each of the genes of this model organism. For each publication included in this set we have queried PubMed [2] for a list of associated MeSH annotations. MeSH is a controlled vocabulary used in indexing of biomedical documents and provides us with a way to summarize articles with a small number of informative descriptors. We next associated each gene in our set to its respective MeSH profile, which consisted of MeSH terms and a corresponding number of papers citing the gene and including a particular MeSH term. This profile is equivalent to a bag (i.e. a multi-set) of descriptors, a data representation format frequently used in text mining. Using standard text mining methods on MeSH profiles we computed: weights measuring the importance of individual MeSH terms for each gene, and the similarities between gene pairs. These were then used to define edges between gene–MeSH term and gene–gene pairs, respectively. For the MeSH network layer, we associated each MeSH term with its publication profile, a set containing all *D. discoideum* publications that are annotated with this particular term. Again, a similarity function was computed and used to define the MeSH term network.

Results

We used the methods from section 2 to optimize the constructed multi-layer network. Visualizations were rendered in a network visualization and explorative analysis component in the Orange data mining framework [Demsar *et al.*, 2004]. In these visualizations the two layers have been overlaid resulting in a 2D image.

The obtained gene–annotation network contained several interesting substructures and enabled many insights to be extracted from the visualization. It could be noticed for example that several MeSH terms are highly related to some clusters of genes. Placed close together they help elucidate these groups.

Another effect can be observed in Fig. 2 showing a section of the network. What would be one homogeneous cluster of genes is divided into two distinct groups by MeSH terms. Two of the terms are common to genes from both groups (*Phylogeny* and *Receptors, Cyclic AMP*), while *Green Fluorescent Proteins* and *GTP-Binding Proteins* are associated with two disjoint groups inside the cluster, thereby distinguishing two subsets of the whole component. We can also see that *Phylogeny*, *Green Fluorescent Proteins* and *Receptors, Cyclic AMP* have been placed very close together, although they are not actually connected in the MeSH term network. This is because they share common neighbors on the gene layer.

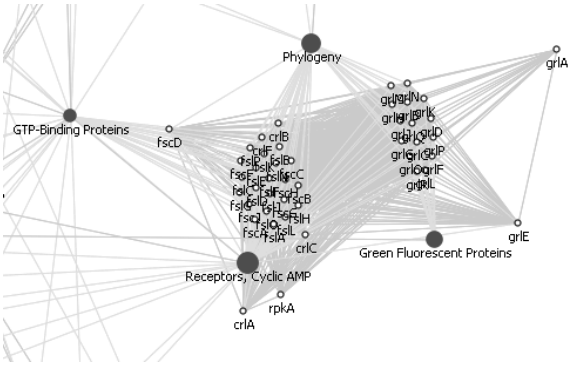


Figure 2: Two MeSH terms are pulling apart distinct groups of an otherwise highly connected cluster of genes.

Many more interesting properties and insights can be obtained by observing various substructures that emerge from this multi-layer network visualization. The meaning of some was clear even to a non-expert, while others have yet to be inspected by researchers studying this interesting model organism.

Acknowledgments

This work was supported by grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

References

[Davidson and Harel, 1996] Ron Davidson and David Harel. Drawing graphs nicely using simulated annealing. *ACM Trans. Graph.*, 15(4):301–331, 1996.

[Demsar *et al.*, 2004] J. Demsar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining. White paper (`www.ailab.si/orange`), Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, 2004.

[Fruchterman and Reingold, 1991] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11):1129–1164, 1991.

[Tollis *et al.*, 1998] Ioannis G. Tollis, Giuseppe Di Battista, Peter Eades, and Roberto Tamassia. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, July 1998.

[1] `http://dictybase.org`

[2] `www.ncbi.nlm.nih.gov/pubmed`

Bayesian Network Wizard: a user-friendly software to learn Bayesian networks

Fulvia Ferrazzi, Antonio De Donno, Riccardo Bellazzi
Department of Computer Engineering and Systems Science, University of Pavia, Italy
fulvia.ferrazzi@unipv.it

Abstract

We present Bayesian Network Wizard, a software to learn static and dynamic Bayesian networks, both for discrete and continuous nodes. The user is guided through all steps of network learning by a user-friendly wizard. The software is freely available to nonprofit users upon request.

1 Introduction

Bayesian networks (BNs) are probabilistic graphical models for the representation and analysis of models involving uncertainty. They are widely used in various fields, such as data mining, diagnostic systems, decision support systems, and bioinformatics. In bioinformatics BNs are becoming increasingly used to learn cellular networks from gene expression data [Friedman, 2004]. Their probabilistic framework makes them able to represent the intrinsic variability of biological systems and naturally take into account the unavoidable noise in the data. Furthermore, BNs have proven to be effective tools for genetic association studies, where they allow learning complex multivariate models involving SNPs and phenotypic traits [Sebastiani *et al.*, 2005].

Different Bayesian network learning algorithms have been proposed in the literature and software tools developed, among which the Bayes Net Toolbox (`http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html`), Bayesware Discoverer (`http://bayesware.com/`), Hugin (`http://www.hugin.com/`), and Banjo (`http://www.cs.duke.edu/~amink/software/banjo/`). However, there are only a few freely available software tools that implement algorithms for both static and dynamic networks. Moreover their use generally requires learning at least the basic syntax of the programming language used by the authors.

We have developed Bayesian Network Wizard (BNW), a software to learn different types of Bayesian networks (static/dynamic) with continuous or discrete variables. The software has a user-friendly interface that guides the user through all phases of network learning, from data loading to the choice of the network type and the parameters required by the algorithm, to graphical visualization of the learned network. BNW is freely available to nonprofit users upon request to the authors.

2 Implementation

Bayesian Network Wizard provides learning algorithms both for static and dynamic BNs with variables that are either all discrete or all continuous.

For discrete variables, BNW employs Matlab functions contained in the Bayes Net Toolbox by K. Murphy. In the case of static networks the learning algorithm proposed by Cooper and Herskovits [Cooper and Herskovits, 1992] is implemented; in the case of dynamic networks the algorithm proposed by Friedman *et al.* is used [Friedman *et al.*, 1998].

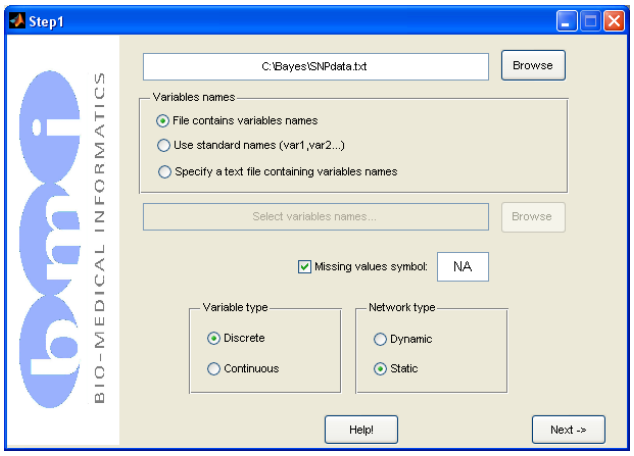
For continuous variables BNW relies on a learning algorithm proposed by us, which assumes that variables are conditionally Gaussian with respect to the parents [Ferrazzi *et al.*, 2007].

BNW has been written in Matlab (version R2007a) and the graphical user interface has been developed using the Matlab GUIDE tool. The program has been compiled with the Matlab Compiler (v. 4.6) for Windows (98/NT/2000/XP) platforms. The stand-alone executable file is distributed together with the runtime Matlab libraries and does not require any license.

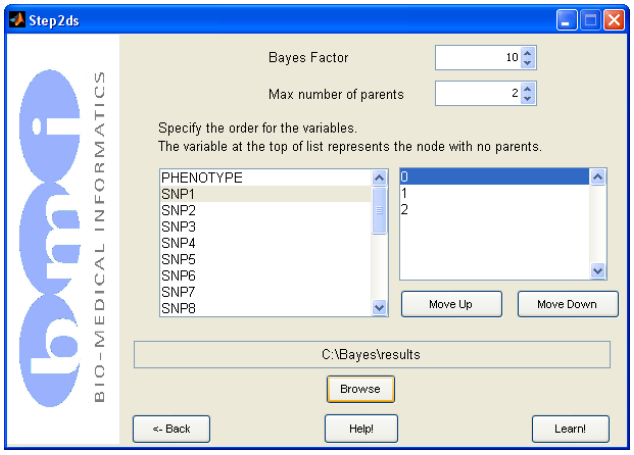
The graphical interface appears as a wizard that guides the user during all steps of the network learning process. At the beginning data are loaded and the user has to choose which type of network he wants to learn (static/dynamic) and the variables type (continuous/discrete). Then another window appears, which allows the user to set the values for the parameters required by the learning algorithm. The appearance of this window varies according to the network type selected at the first step.

After the user has chosen the parameters learning starts and, as soon as it is concluded, a window containing the learned graph opens. The graphical visualization exploits the open source Java libraries Jung 1.7.6 (`http://jung.sourceforge.net`). A Menu bar allows the user to choose between different graph layouts, personalize node and arc color, zoom in/out, and save the graph in different file formats. Furthermore, the graph is saved also as a GML file, a text-based graph file format that can be visualized using yEd, a powerful graph editor (`http://www.yworks.com/en/products_yed_about.html`).

Figure 1 presents a screenshot of the program windows in the case in which a static network is learned from discrete data, while Figure 2 shows a graphical representation of the learned network.



(a)



(b)

Figure 1: **Screenshot of the program windows.** (a) The first window allows the user to load the data and choose the network type; (b) The second window allows the selection of the necessary parameters for the chosen network type.

3 Conclusions

We have presented BNW, a software to learn Bayesian networks. BNW provides learning algorithms for different network types and guides the user by means of a wizard. BNW can be applied in a variety of application settings, including data mining, bioinformatics, and biomedical systems modeling.

Its flexibility and easiness of use allow also inexperienced users to quickly obtain results on different types of data, thus making it a useful instrument for research and especially suitable in teaching settings.

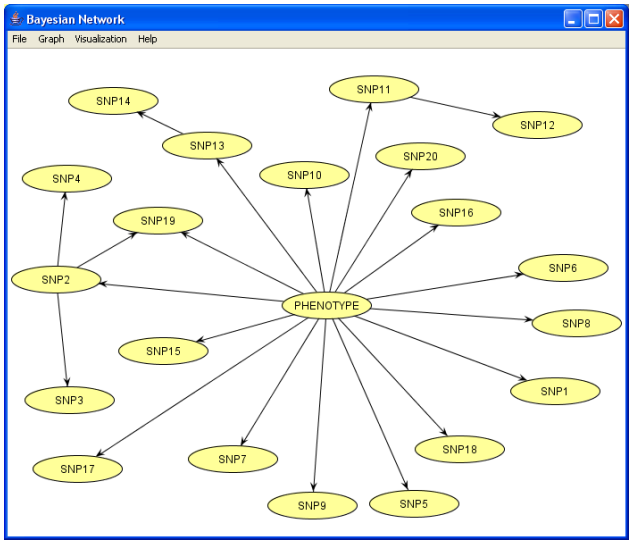


Figure 2: **Graphical representation of the learned network.** Once learning is completed, a window containing a graphical representation of the network opens.

Acknowledgments

This work is part of the FIRB project "ITALBIONET-Rete Italiana di Bioinformatica". F. Ferrazzi is supported by an Investigator Fellowship from Collegio Ghislieri, Pavia, Italy.

References

[Cooper and Herskovits, 1992] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[Ferrazzi *et al.*, 2007] F Ferrazzi, P Sebastiani, M F Ramoni, and R Bellazzi. Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear gaussian networks. *BMC Bioinformatics*, 8 Suppl 5, 2007.

[Friedman *et al.*, 1998] N. Friedman, K. Murphy, and S. Russel. Learning the structure of dynamic probabilistic networks. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 139–147, 1998.

[Friedman, 2004] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[Sebastiani *et al.*, 2005] P. Sebastiani, M. F. Ramoni, V. Nolan, C. T. Baldwin, and M. H. Steinberg. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet*, 37(4):435–40, 2005.

SNP2Net: a tool for gene-based predictive modeling in genome-wide association studies

João V. Duarte[1], Angelo Nuzzo[2], Alberto Malovini[1,3], Annibale A. Puca[3], Riccardo Bellazzi[1]
[1]Department of Computer Science and Systems and [2]Centre for Tissue Engineering,
University of Pavia, Italy
[3]IRCCS Multimedica, Milano, Italy
riccardo.bellazzi@unipv.it

Abstract

In this paper we present an integrated tool to define gene-based predictive models from genome-wide SNPs datasets. Compared to the standard SNP-based models, our approach leads to more parsimonious models without losing predictive ability. We developed an integrated framework to automate the overall complex analysis strategy, in order to perform a robust validation of our approach and to make it applicable to different datasets for further analysis. We tested the tool on a dataset coming from a real genome-wide association study, comparing SNP and gene-based models performances. Results show that our proposed method is more parsimonious and therefore less prone to over fitting than standard SNP-based approaches, while preserving prediction performances.

1 Introduction

Genome-wide association studies (GWAS) represent a powerful approach to identify disease susceptibility genes, offering the promise of discovering novel targets for therapeutic treatments. However, these studies are affected by the non-trivial problem multi-hypotheses testing significance [Balding, 2006]. To overcome this problem, non parametric approaches may offer a suitable methodological solution. Bayesian networks (BNs) represent an appropriate approach to perform multivariate analyses, as they are able to represent complex associations between phenotype, genetic and environmental factors using a small number of parameters [Sebastiani *et al*., 2005]. However, learning a BN from a SNP-based GWAS dataset is often non-trivial due to the high number of variables to be taken into account in the model (hundreds of thousands), with respect to the instances of the dataset (dozens to few thousands). Therefore, it becomes very interesting to use an abstraction of the variable space that suitably reduces its dimensionality without losing information. As the final aim of genetic association studies is to identify how genes influence the phenotype, we showed how representing the information of the set of SNPs mapping to the same gene as a new meta-attribute may be a good choice to build a more parsimonious model [Malovini *et al.*, 2009]. However, the complexity of that analysis required the use of different software tools for each step, making it hard to be done automatically and affecting its reproducibility. In this paper we describe how we overcome this problem with the development of a tool performing an overall automation of the strategy, so that: i) a more robust evaluation of the performance of the proposed approach is provided, and ii) an integrated framework is available to be used for further analysis on different datasets.

2 Methods

We developed *SNP2Net*, a MATLAB tool that aims at automatically perform the analysis strategy, which is made of two main steps: i) generation of meta-attributes (each one representing a gene), by applying a classification tree based method, ii) learning of a BN in which the nodes represent the meta-attributes and the phenotypic trait of interest. Moreover, the tool allows an easy setting of the many parameters required by using a graphical user interface (GUI).

The meta-attribute definition is made by grouping SNPs according to their gene annotation mapping. After mapping the 40 top-associated SNPs (i.e. with a p-value $< 10^{-4}$ coming from standard GW allelic association tests), the tools selects genes represented by at least two SNPs (24 genes). Then each meta-attribute is given a value ("state") by learning a classification tree (as described in [Malovini *et al.*, 2009]). There are several parameter affecting both the classification tree generation and the logic of states assignment, which may be setup by the user in the GUI (see result section). Finally, a dataset containing meta-attributes states is generated applying the rules learned by the classification tree on the original SNPs dataset.

Once the meta-attributes are defined, both SNPs and meta-attributes datasets are used to build a BN for phenotype prediction. BN structure and probability distribution learning are performed including the BN toolbox [Murphy, 2002], which implements the K2 search algorithm for the structure learning phase [Cooper and Herskovits, 1992]. The algorithm requires the specification of the variables searching order, which can be chosen from the GUI between gain ratio and information gain. Once the BN is learned, it can be used to infer the posterior probability of a node status given the evidence of the other nodes: this allows to predict the phenotype status given the other variables values.

In order to assess the network prediction performance and how it will generalize when applied to a new independent dataset, we made: i) an internal validation, using a K-fold cross-validation (CV) strategy (properly generating folds with respect to the class values distribution), and ii) an external validation, consisting in performing the K-fold cross-validation thousands times, which gives a measure of the stability of the performance with respect to the fold sampling.

3 Results and Discussion

We tested our method using data coming from a genome-wide scan involving 570 35-55 years old patients affected by arterial hypertension (AH) and a control population of 664 individuals without an AH historical. We removed examples with missing values (17,4% of the sample), as a suitable method to deal with missing data is now under development. The classification tree algorithm has been setup with the following parameters: *m-estimate* = 8 for the Minimal Error Pruning; number of different states for each meta-attributes = 5; minimum number of instances in one leave of the tree to be considered as a state = 13 (1% of total instances); K2 ordering criteria = gain ratio.
Performing a 10-fold CV, we obtained a classification accuracy of 58,18% for the SNP-based network and 62,80% for the gene-based net. Finally, performing a 150-times replication of the 10-fold CV, we obtained two different accuracy values distribution summarized in the box-plots shown in Figure.1.

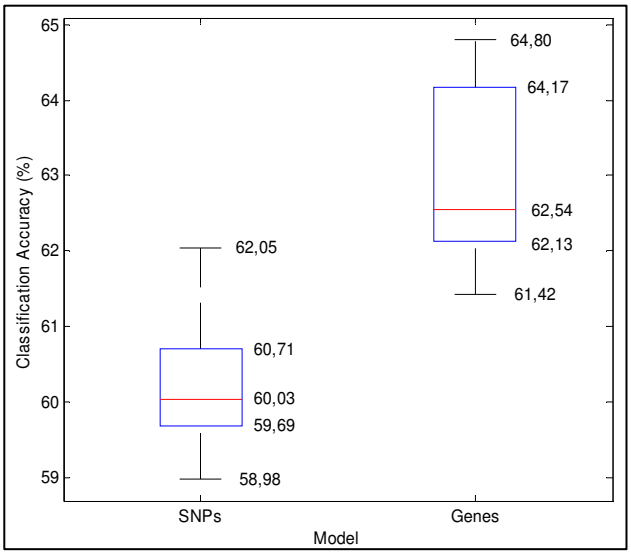


Figure 1. Box plot of the classification accuracy distribution obtained from a 150-times replication of 10-fold CV, on both SNPs and genes (meta-attributes) nets. Each box has lines at the lower quartile, median, and upper quartile values. SNP-based model has a distribution around the median accuracy of 60,03%, while the gene-based model has a distribution around a higher median value, 62,54%.

Thanks to the availability of *SNP2Net* (a demo version is available at http://bioinfo.unipv.it) it was possible to perform a thorough evaluation of the performances of the gene-based prediction method herein presented. In particular it was possible to show that:

- Our proposed method is more parsimonious than standard SNP-based approaches, while preserving prediction performances; moreover, since our proposed method has less variables than the SNP-based one, it is also less prone to over fitting.
- It is rather interesting to note that our proposed method seem to have less variance than the SNP-based one, as shown by the box plot diagram.
- The proposed model is a suitable alternative to haplotypes, which are on the contrary frequently used also as prediction factors.

It is needless to say that the implemented tool may easily allow further generalization. For example, it will be possible for the user to define lists which contains SNPs that may be related, even if they do not belong to specific gene. In this case, it will be possible to implement a customized two step prediction strategies that may allow using different kinds of knowledge and user-provided information.

Acknowledgments

This work is a part of the project "Bioinformatics for Tissue Engineering: Creation of an International Research Group", funded by the "Fondazione Cariplo", the "ITALBIONET - Rete Italiana di Bioinformatica" and the RBIN04X9XE FIRB projects funded by MIUR.

References

[Balding, 2006] David J. Balding. A tutorial on statistical methods for population association studies. *Nature Review Genetics* 7(10):781-91, Oct 2006.

[Cooper and Herskovits, 1992] Gregory F. Cooper, Edward Herskovits: A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 1992, 9:309-347.

[Malovini *et al.*, 2009] Alberto Malovini, Angelo Nuzzo, Fulvia Ferrazzi, Annibale A. Puca, Riccardo Bellazzi: Phenotype forecasting with SNPs data through gene-based Bayesian networks. *BMC Bioinformatics*, 5-10 Suppl 2:S7, Feb 2009.

[Murphy, 2002] Kevin Murphy: Bayes Net Toolbox for Matlab,1997-2002. http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html

[Sebastiani *et al.*, 2005] Paola Sebastiani, Marco F. Ramoni, Vikki Nolan, Clinton T. Baldwin, Martin H. Steinberg: Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nature Genetics*, 37(4):435-440, 2005.

Classification of ICU Patients via Temporal Abstraction and Temporal Patterns Mining

Robert Moskovitch[1], Niels Peek[2], Yuval Shahar[1]
[1]Medical Informatics Research Center, Department of Information Systems Engineering,
Ben Gurion University, P.O.B. 653, Beer Sheva 84105, Israel
[2]Dept of Medical Informatics, [3]Dept of Intensive Care Medicine, Academic Medical Center,
University of Amsterdam, P.O.B. 22700, 1100 DE Amsterdam, The Netherlands
{robertmo, yshahar}@bgu.ac.il, {n.b.peek }@amc.uva.nl

Abstract

The use of unsupervised mining, such as association rules and sequential mining was used in the past as a preprocessing step to the classification task. The recent growth in the use of temporal abstraction for temporal knowledge discovery from multivariate temporal data, through time intervals mining, is proposed to be employed for the classification of multivariate temporal data. We present KarmaLego, an efficient time intervals mining which discovers non-ambiguous patterns based on a flexible version of Allen's relations. Later the discovered patterns are used as features which represent the relations of the multiple variables. Several settings for the suggested approach are examined through a rigorous evaluation.

1 Introduction

With the increase in patient data logged overtime and the reduction of storage costs, classification of multivariate temporal data is very useful in many tasks. Temporal data provides more detailed description of the patient situation which is expected to lead to better diagnosis and decisions, however, analyzing multivariate temporal data, especially for classification is challenging, often made without consideration of the relations among the multivariate time series. Commonly the representation of time stamped data is made using time windowing, in which features such as the mean, minimal or maximal value is extracted. More sophisticated approaches represent the time window with discrete values, such as qualitative mean (e.g., low, medium, high) [20]. Other approaches transform the time series to the frequency domain by Furrier transform [2] for example.

However, determining the right time window size is commonly problematic. Then extracting features from the time series within a given time window, such as minimal value, or transformations such as wavelets or Fourier transform, do not allow an explicit temporal analysis. Additionally, these approaches do not allow expressing the relations among the multi variables along time.

In this paper we present an approach, in which the time series are abstracted into time intervals series [18] and frequent temporal patterns of the multivariate time interval series are discovered. Then the discovered patterns are used as features to represent the classified entities. We demonstrate our approach in the domain of Intensive Care Unit. We start by surveying the background. In the methods section we describe the descretization methods we used for the temporal state abstraction and KarmaLego – the time intervals mining method. Finally, we describe the entire procedure through the rigorous evaluation and results on the ICU dataset.

2 Background

In our approach we first abstract the time series to *states*, then mine them to discover frequent patterns, which are later used as features in the classification task. Thus, in this section we refer to the domain of Temporal Abstraction [18], we survey the recent development in time intervals mining and refer to the concept of classification based on temporal patterns.

2.1 Temporal Abstraction

Temporal abstraction is the aggregation of time series to a summarized and a more comprehensive representation for a human or productive for further data mining tasks. The task of temporal abstraction corresponds to the task of segmenting the time series, having a meaningful symbol for each segment. Segmenting time series [7] is a representation of time series in a piecewise linear representation, which is the approximation of a time series length n with k straight lines, usually $k<<n$.

Knowledge-based temporal abstraction (*KBTA*), as presented by Shahar [18], infers domain-specific interval-based abstractions from point-based raw data, based on a formal domain-specific abstraction and interpolation. However, although the KBTA method applies the temporal-abstraction knowledge to create abstractions that are meaningful to the domain expert, such knowledge is not always available. Moreover, the knowledge provided by the domain expert is not always pertinent to the data-

mining task, such as classification, but rather to the clinical expert's routine activities, such as diagnosis or therapy [20]. Thus, several automated data-driven methods, which provide abstractions are less semantically meaningful, but can be potentially useful for data mining and finally classification, which we will present here. This can be achieved for state abstraction, by simply applying unsupervised discretization methods, such as *equal width discretization* and *equal frequency discretization*. Other methods are *k-means clustering* [8], in which the time series values are grouped into *k* clusters from which the states can be deduced. More sophisticated methods refer to the temporal order of the time points in the time series, include: Symbolic Aggregate approXimation (SAX) [7] is a method for symbolic representation of time series and Persist [8].

2.2 Mining Time Intervals

Mining time intervals is a relatively young research field. One of the earliest studies in the area is that of Villafane et al. [20], which searches for *containments* of intervals in a multivariate symbolic interval series. Kam and Fu [6] were the first to use all of Allen's temporal relations [1]. Höppner [2] was the first to define a non-ambiguous representation of Allen's-based time intervals patterns by a k^2 matrix to represent all of the pairwise relations within a *k*-intervals pattern. Unlike Höppner's naïve mining method, Papapetrou et al. [15] presented a mining method, which results in an *enumeration tree* that enumerates all the symbols and their possible relations combinations, presenting a BFS, DFS and Hybrid approaches, using only five temporal relations. . Additionally, they relaxed the temporal relations with the notion of an epsilon, which we extended to all Allen's relations (fig 1).

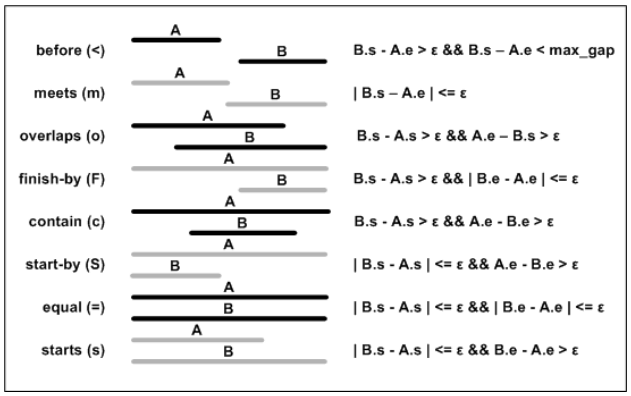


Figure 1 - A flexible extension of Allen's seven relations using the same Epsilon value for all relations. The [eighth] *starts* relation is required when epsilon > 0.

ARMADA presented recently by Winarko and Roddick [23], uses a candidate generation and mining iterations approach. Mörchen [11] proposed an alternative to Allen's relations-based methods, in which time intervals are mined to discover coinciding symbolic time intervals, called *Chords*, while repeating partially ordered chords are called *Phrases*. Sacchi et al [18] use abstracted time series to find temporal association rules by generalizing Allen's rules into a relation called PRECEDES. In [14] we present KarmaLego in more detail and evaluate its performance in comparison to Papapetrou's [15] and ARMADA's [23] methods.

2.3 Classification Using Temporal Patterns

Using unsupervised methods for knowledge discovery as a preprocessing step for classification was suggested first by Liu et al [8]. The authors used association rules mining to discover frequent association rules of each class, called Class Association Rules which later used as features in the classification task. Since then several approaches were proposed to employ association rules for classification, and later sequential mining was employed as well. Further to the recent growth of interest in the use of temporal abstraction as a preprocessing stage and time intervals mining for multivariate temporal knowledge discovery [12] , the idea of employing time intervals related patterns for multivariate temporal data classification was suggested very recently in [2]. The authors used domain expert for state abstraction and gradient abstraction [19] (e.g., increase, decrease) and further mined the intervals for complex patterns consisting on the before and overlap Allen's relations to overcome the lack of robustness in Allen's relations. Finally a Boolean representation of the temporal patterns is used, according to their existence in the classified entity.

3 Methods

3.1 State Abstraction

To abstract the temporal data into states we used a Knowledge Based approach, in which the data was abstracted according to the domain expert definitions, coming from his domain of diagnosis and treatment as described in [20]. As alternative to the knowledge based approach we used four discretization methods which are the focus of this section and this study.

Equal Width Discretization (EQW) determines the cut-points by dividing the value range into equal width bins. The number of values in each bin is based on the distribution of the values.

Equal Frequency Discretization (EQF) divides the value range into bins, so that the frequency of values in each bin is equal, thus, the number of values in each bin is equal.

Symbolic Aggregate Approximation (SAX) [7] consists of two steps. The first step is the Piecewise Aggregate Approximation, in which the granularity of the time series is reduced by averaging several time points into a single valued time point. The second and main step of the SAX method is the discretization of the PAA output. The discretization is based on the assumption that normalized time series have a Gaussian distribution, and on the desire to produce equal probability states. The time series is therefore normalized and discretized into a fixed number of states according to predetermined cut-points which produce equal-sized areas under a Gaussian curve.

Persist [10] is a univariate discretization method designed to maximize the mean duration of the resultant time intervals. The method measures the marginal probability P(Sj) and the self-transition probability A(j,j) = $P(s_i=s_j|s_{i-1}=s_j)$, where S = $\{s_1,\ldots,s_k\}$ is the set of resultant states (symbols), assuming that k states are desired. The algorithm searches for the bins divisions which maximizes the self transition probabilities in comparison to the marginal probabilities, which is expected to maximize the mean duration of the resultant intervals, for the given number of states.

3.2 KarmaLego – Fast Time Intervals Mining

The discovery of temporal interval patterns is computationally highly demanding, since it requires generating all of Allen's basic seven relations. Thus, we developed KarmaLego, a fast algorithm which enumerates all of the patterns whose frequency is above a given support threshold. Due to the lack of space in this paper, whose focus is on the use of KarmaLego for classification of multivariate time series, we will present the algorithm only briefly; a detailed report, including a rigorous comparison to previous methods is available elsewhere [14]. We start with essential definitions and proceed with a general description of the algorithm. KarmaLego consists on a flexible version of Allen's seven relations to increase the robustness of the temporal relations, by adding an epsilon value, as shown in figure 1, where all the seven relations are used and an additional relation when epsilon>0. We enable also to limit the before relation by a maximal allowed gap, which was proposed by [23].

Definition 1. A *symbolic time interval series*, $IS = \{I^1, I^2, .. , I^n\}$, where each symbolic time interval $I = <s, e, sym>$, is an ordered pair of time stamps, start-time (s) and end-time (e), and a symbol (*sym*), which typically includes an abstraction.

Definition 2. A *lexicographic symbolic time interval series* is a time interval series, sorted in the order of the start-time, end-time, and a lexicographic order of the symbols, $IS = \{I^1, I^2, .. , I^n\}$, where

$$\forall I^i, I^j \in IS \ (i < j) \wedge ((I^i_s < I^j_s) \vee (I^i_s = I^j_s \wedge I^i_e < I^j_e) \vee (I^i_s = I^j_s \wedge I^i_e = I^j_e \wedge I^i_{sym} < pI^j_{sym}))$$

Definition 3. A non ambiguous *Lexicographic Time Intervals Relations Pattern* (TIRP) *P* is defined as $P = \{\breve{I}, \breve{R}\}$, where $\breve{I} = \{I^1, I^2, .., I^k\}$ is a set of *k* symbolic time intervals ordered lexicographically and

$$\breve{R} = \bigcap_{i=1}^{k} \bigcap_{j=i+1}^{k} r(I_i, I_j) = \{r_{1,2}(I_1,I_2), r_{1,3}(I_1, I_3), .., r_{1,k}(I_1, I_k), r_{2,3}(I_2, I_3), r_{2,4}(I_2, I_4), .., r_{2,k}(I_2, I_k), \ldots, r_{k-1,k}(I_{k-1}, I_k)\},$$

defining all the relations among each of the $(k^2-k)/2$ pairs of symbolic time intervals in $\breve{I}$.

Definition 4. Given a database of |*E*| entities, The *vertical support* of a TIRP *P* is denoted by the cardinality $|E^P|$ of the set E^P of distinct entities (e.g., different patients), having P at least once, divided by the total number of the entities (e.g., patients) |*E*|: $ver\_sup(P) = |E^P| / |E|$

Definition 5. The *horizontal support* of a TIRP *P* for an entity e_i (e.g., a single patient's record) is the number of instances of the TIRP *P* found in e_i - $hor\_sup(P, e_i)$.

Definition 6. The *mean global horizontal support* of a TIRP *P* is the average of the horizontal support of P for all the entities |E|: $mean\_global\_hor\_sup = \frac{\sum_{i=1}^{|E|} hor\_sup(P,e_i)}{|E|}$.

When the set *E* is restricted to the set E^P, we refer to the *mean local horizontal support* of *P*, *mlh_sup*(*P*), since the horizontal support is computed only for entities in which P was found at least once, which we will use in this study.

Problem Definition

Given a set of entities *E*, described by a symbolic time intervals series *IS*, and a *minimum vertical support* threshold *min_ver_sup*, the goal is to find all the TIRPs whose frequency is above *min_ver_sup*.

KarmaLego consists of two main procedures, first the creation of the first and second levels of the enumeration tree (see figure 2), including all the symbols and the 2-sized TIRPs having above minimal vertical support in a BFS search, as presented in Algorithm 1. After enumerating all the 2-sized TIRPs, the next procedure is a recursive extension of the k-sized TIRPs above the minimal vertical support, as shown in algorithm 2. An illustration of the resulted entire enumeration tree is presented in figure 2, in which each TIRP is presented by a half-matrix which represents the relations of a TIRP.

Algorithm 1 – KarmaLego
Input: *db*; *min_ver_sup*; *epsilon*.
Output: T – an enumerated tree of TIRPs
T^2 – *the tree at second level*, InsVec – a vector containing all the TIRPs instances found in the data, P_vec – the parameters vector in a searched TIRP, Rel_vec – the relations vector of a TIRP.

1. $T^1 \leftarrow S \leftarrow$ *min_ver_sup(db)* //all frequent symbols in *db*
2. $T^2 \leftarrow T^1$ and enumerating all $s \in S$, above *min_ver_sup*
3. foreach $t \in T^2$
4. Search_Extended_TIRPs(T^2, *t*, *3*, *min_ver_sup*, *S*)
5. end

Algorithm 2 – Search_Extended_TIRPs
Input: T^2, *t*, *level*, *min_ver_sup*, *S*
Output: void

1. foreach $s \in S$
2. foreach $r \in R$
3. Generate extended TIRPs t'_s from *t*, *r* and *s*
4. Search for supporting instances of *t* T^2
5. if(ver_sup(t'_s) > min_ver_sup)
6. Search_Extended_TIRPs(T^2, t'_s, level+1, min_ver_sup, S)
7. end

A major challenge in the candidate generation of time intervals pattern is the exponential growth of the candidates with the number of the relations with the size of the time intervals (k). In karmaLego the candidates are generated by exploiting the transitive property of the temporal relations to eliminate the generation of wrong logically

candidates which often made by naïve generation, as explained in details in [14].

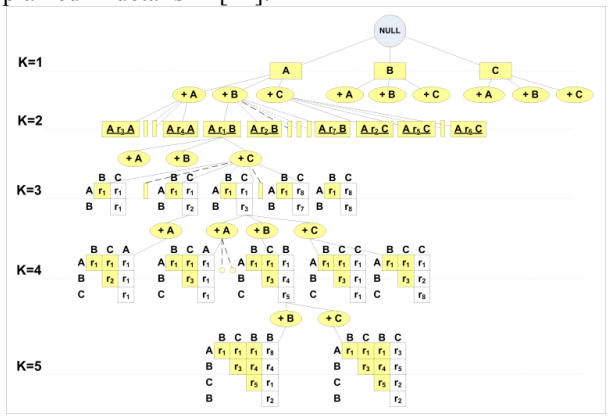


Figure 2 – A KarmaLego enumeration tree, in which the direct expansion of TIRPs is performed. Each node represents a TIRP as a half matrix of the temporal relations.

3.3 Bag of TIRPs

After discovering frequent TIRPs from the entities we can use them as features for the classification task. We adopt the idea of *bag-of-words* from the textual domain [22], in which terms are used to represent documents. In our case it is a *bag-of-tirps*, which are used to represent the patients for the classification task. The discovery process through KarmaLego results with a list of TIRPs, having its *vertical support*, which describes in how many patients it was discovered, and *horizontal support* for each patient, which describes how many times it was found for a specific patient within the period of time in the dataset.

When attempting to classify based on TIRPs a table has to be created, in which the columns (features) are the TIRPs and the last is the label, and the rows are the patients. In this study we suggest two types of representation for a TIRP, related to a patient:

- *binary*, in which a TIRP is represented by 1 it exists and 0 otherwise.
- *horizontal support* (HS), in which the value is the number of instances of the TIRP were discovered in the patient multivariate temporal data.

3.4 Feature Selection

The reduction of the number of TIRPs (features) as often is required in classification procedures is even more essential here since the features (TIRPs) are sometimes extensions of other TIRPs and thus refer to a subset of patients of the extended TIRP, which creates dependency within the features. Thus, we used several filtering feature selection methods, in which each feature is ranked based on some criterion that measures the correlation to the class to estimate the classification potential. Finally we extracted the top ranked features. We used a simple method which is based on the Vertical Support (VS), Gain Ratio and Fisher Score.

- *VS* - Based on the *vertical support* of each TIRP we selected the TIRPs having the highest VS, which were expected to represent more patients.
- *Gain Ratio* - was designed to overcome a bias in the Information Gain measure, and which measures the expected reduction of entropy caused by partitioning the examples according to a chosen feature [8].
- *Fisher Score* - The Fisher score ranking technique calculates the difference, described in terms of mean and standard deviation, between the positive and negative examples relative to a certain feature [4].

4. Evaluation

To evaluate the approach of the classification through temporal patterns on the ICU dataset, we designed an experiment which includes several settings for each parameter.

4.1 ICU Data Set

An ICU dataset was used of patients who underwent cardiac surgery at the Academic Medical Center in Amsterdam, the Netherlands, in the period of April 2002-May 2004. Two types of data were measured: *static data* including details on the patient, such as *age*, *gender*, *surgery type*, whether the patient was mechanically ventilated more than 24 hours during her postoperative ICU stay, and *temporal data*, measured each minute along the first 12 hours of the ICU hospitalization, including: mean arterial blood pressure (ABPm), central venous pressure (CVP), heart rate (HR), body temperature (TMP), and two ventilator variables, namely fraction inspired oxygen (FiO2) and level of positive end-expiratory pressure (PEEP). The data contains 664 patients, among which 196 patients were mechanically ventilated for more than 24hr (29.5%).

4.2 Experimental Plan

We wanted to answer several questions.

1. What is the best Discretization method and how many states are required?
2. What is the best epsilon value for the mining?
3. What is the best TIRP representation?
4. What is the best feature selection method and Top selection?

We used the temporal data of the last *three* hours ninth till twelfth hours. All the variables were abstracted using the four discretization methods [EQW, EQF, SAX, Persist] into 3 and 5 states. Then each abstracted dataset was mined using KarmaLego with *maximal gap* (on the relation *before* – see fig 1) of 100 seconds, minimum vertical support of 20% and the TIRPs were restricted to not more than 5 intervals. The mining was performed with three epsilon values 0, 5 and 10. After that the discovered TIRPs were used to create a matrix of the patients using the two representations: binary and horizontal support. Then the features were selected by the three feature selection methods Vertical Support, Gain Ratio and Fisher Score into three top selection levels 50, 100 and 150. This resulted into 243 evaluation runs, in which we used several classifiers, but we report the results of Random

Forest using weka Classification method, which outperformed the other, with 10 cross validation.

5. Results

In the results analysis we compare the mean accuracy values of all the experiments according to the parameters. The parameters are analyzed according to the procedure described earlier.

5.1 Discretization methods

Figure 3 presents the mean accuracy of all the runs relating to the each discretization method and the states number. In most of the discretization methods abstracting into 5 states outperformed the 3 states, except for Persist.

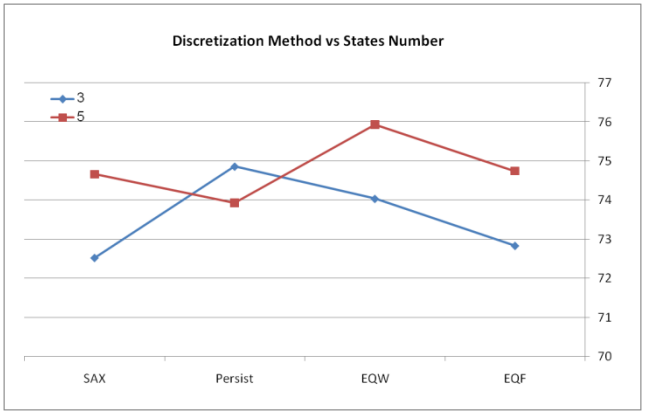


Figure 3 – Abstracting into five states outperformed for most of the methods, out of Persist.

5.2 Epsilon and TIRP representation

Figure 4 presents the mean accuracy for all the runs of the epsilon values and the TIRP representations. The Binary representation outperformed the HS. This indicates that the number of instances of the TIRP is being of less important to the classification task. While the Binary representation accuracy increased for lower epsilon it was slightly improved for the HS when the epsilon value increased. This is actually surprising since it was expected that having a larger epsilon will increase the robustness of the temporal relations and will enable more representative temporal patterns.

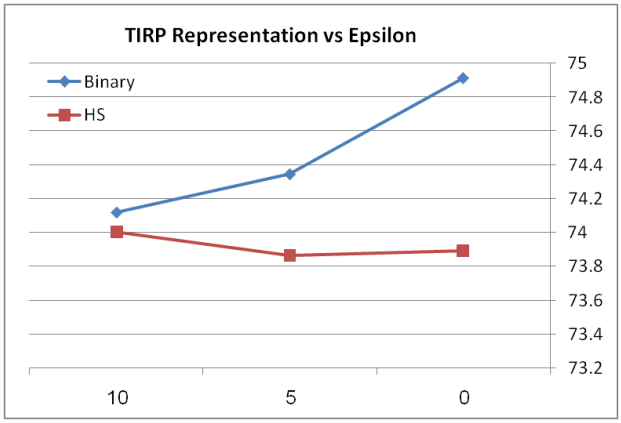


Figure 4 – The binary outperformed the HS representation, especially when the epsilon value was 0.

5.3 Feature Selection

The Fisher Score and the Gain Ratio outperformed the Vertical Support which selects the TIRPs with the highest vertical support. This can be explained by the imbalance of the classes, which might not represented by the TIRPs having the highest VS.

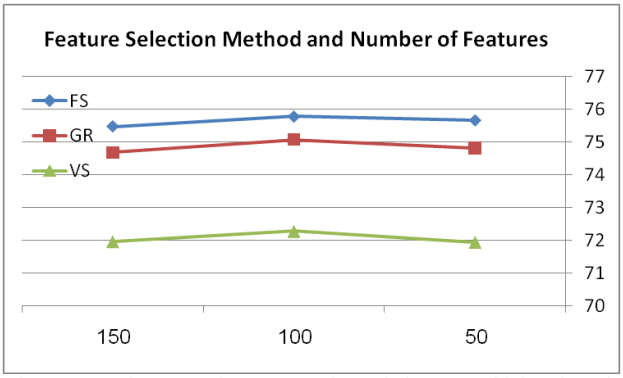


Figure 5 – The FS and GR outperformed the VS, which select the TIRPs with the highest VS.

We present in Table 1 the top six runs which brought the best accuracies. The top runs correspond to the mean accuracy analysis presented in the previous sections, which shows that the best settings for this dataset are abstracting using the EQW into 5 states, which is interesting since it is the simplest discretization method. Mining using low value of epsilon and using the Binary representation for the TIRPs. Finally to reduce the features both Gain Ratio and Fisher Score with 100 to 150 features were the best.

Table 1 – the top evaluation runs with the dominant settings in bold.

discM	|s|	e	TIRP - Rep	|f|	FS	Acc [%]
EQW	**5**	**5**	**Bin**	100	**GR**	79.64
EQW	**5**	10	**Bin**	**150**	FS	79.64
EQW	**5**	0	HS	**150**	**GR**	79.34
EQW	**5**	0	**Bin**	100	**GR**	79.34
EQW	**5**	**5**	**Bin**	**150**	FS	79.03
EQW	**5**	**5**	**Bin**	**150**	**GR**	79.03

6. Discussion and Conclusions

We presented an approach for classification of multivariate temporal data using time intervals patterns, which were discovered after abstracting the multivariate time series into time intervals. In addition to knowledge based state abstraction we used discretization methods. We presented an efficient time intervals mining algorithms, called karmaLego, for the discovery of non-ambiguous patterns consisting on a flexible version of Allen's temporal relations. The discovered patterns (TIRPs) are then used as features for the classification task. We presented two representation approaches, binary and horizontal support which uses the number of instances of the TIRP within the given temporal data. To reduce the number of features we used two commonly used feature selection measures and a measure consisting on the vertical sup-

port. Finally we used the Random Forest classification method for the classification task.

Our analysis shows that abstracting into five states was better for most of the methods out of Persist, which is motivated by creating long intervals (fig 3). The Binary representation though simpler outperformed the Horizontal Support, which can be explained by small values of HS and it might be too noisy. This happened especially for epsilon 0 (fig4). The feature selection measures were much better than the VS measure, which prefers TIRPs with high vertical support. This effect can be because this dataset is imbalanced (fig 5). Finally, we showed in table 1 the best runs with the best settings.

For future work we would like to perform another mining approach, in which each class patients are mined separately to discover its representative TIRPs, which can reduce the problem of imbalanced datasets. Then after discovering the TIRPs of each class to create a matrix based on their unification. Additionally, we plan to develop a discretization method that considers the class of each patient to perform an abstraction which maximizes the difference in the states distribution for each class, which is expected to discover different TIRPs for each class for better classification. Additionally we would like examine the use of smaller set of temporal relations which are more general to increase the number of discovered TIRPs.

References

[1] J. F. Allen. Maintaining knowledge about temporal intervals, Communications of the ACM, 26(11): 832-843, 1983.

[2] I. Batal., L. Sacchi, R. Bellazi, M. Hauskrecht, Multivariate Time Series Classification with Temporal Abstractions, Twenty-Second International FLAIRS Conference, Florida, 2009.

[3] R. N., Bracewell, The Fourier Transform and Its Applications, Boston: McGraw-Hill, 2000.

[4] Golub, T., Slonim, D., Tamaya, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and E. Lander, E., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science, 286:531-537, 1999.

[5] F. Höppner, Learning Temporal Rules from State Sequences, Proceedings of WLTSD-01, 2001.

[6] P. S. Kam and A. W. C. Fu, Discovering temporal patterns for interval based events, In Proceedings DaWaK-00, 2000.

[7] J., Lin, E., Keogh, S., Lonardi, B., Chiu, A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13, 2003.

[8] B. Liu, W. Hsu, Y. Ma, Integrating Classification and Association Rule Mining, In Proceedings of KDD-98, New York, 1998.

[9] Mitchell T. Machine Learning, McGraw-Hill, 1997.

[10] F. Mörchen, and A. Ultsch, Optimizing Time Series Discretization for Knowledge Discovery, In Proceeding of KDD05, 2005.

[11] F. Mörchen, Algorithms for Time Series Knowledge Mining, Proceedings of KDD-06, 2006.

[12] F. Mörchen, Unsupervised Pattern Mining From Symbolic Temporal Data, SIGKDD Explorations, 9 (1), 2007.

[13] R. Moskovitch, and Y. Shahar, Temporal Data Mining Based on Temporal Abstractions, (IEEE) ICDM-05 workshop on Temporal Data Mining, Houston, US, 2005.

[14] R. Moskovitch, Y. Shahar, KarmaLego – Fast Time Intervals Mining, ISE-TECH-REP 23/2009.

[15] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos, Discovering Frequent Arrangements of Temporal Intervals, Proceedings of ICDM-05, 2005.

[16] D. Rajanish, Classification Using Association Rules, IIMA Working Papers 2008-01-05, Indian Institute of Management Ahmedabad, 2008.

[17] L. Sacchi, M. Verduijn. N. Peek, E. de Jonge, B. de Mol, R. Bellazzi. Describing and modeling time series based on qualitative temporal abstraction. Workshop notes of the IDAMAP workshop, 2006.

[18] L. Sacchi, C. Larizza, C. Combi, and R. Bellazi. Data mining with temporal abstractions: learning rules from time series. Data Mining and Knowledge Discovery, (15):217–247, 2007.

[19] Y., Shahar, A framework for knowledge-based temporal abstraction, Artificial Intelligence, 90(1-2):79-133, 1997.

[20] M, Verduijn, L, Sacchi, N, Peek, R, Bellazzi, E, de Jonge, B, Mol, Temporal abstraction for feature extraction: A comparative case study in prediction from intensive care monitoring data. Artificial Intelligence in Medicine, 41(1): 1-12, 2007.

[21] R. Villafane, K. Hua, D Tran, and B. Maulik, Knowledge discovery from time series of interval events, Journal of Intelligent Information Systems, 15(1):71-89, 2000.

[22] G., Salton, A., Wong, and C.S., Yang, A vector space model for automatic indexing. Communications of the ACM, 18:613-620, 1975.

[23] E. Winarko and J. Roddick. Armada - an algorithm for discovering richer relative temporal association rules from interval-based data. Data and Knowledge Engineering, 1(63):76–90, 2007.

Ontology-based semantic similarity in the biomedical domain

Montserrat Batet[1], **David Sanchez**[1], **Aida Valls**[1], **Karina Gibert**[2]
[1] Dept. Computer Science and Mathematics, Universitat Rovira i Virgili (Spain)
{montserrat.batet, david.sanchez, aida.valls}@urv.cat
[2] Dept. Statistics and Operations Research, Universitat Politecnica de Catalunya (Spain)
karina.gibert@upc.edu

Abstract

Computation of semantic similarity between concepts is a very common problem in many language related tasks. In the biomedical field, several approaches have been developed to deal with this issue by exploiting the knowledge available in domain ontologies. In this paper, we study the behaviour of several measures based on the exploitation of the geometrical model of a domain ontology of the biomedical field (SNOMED-CT). Then, we propose a new approach based on the amount of overlapping and non-overlapping taxonomical knowledge between a pair of concepts. Its performance is compared against classical approaches using a standard benchmark composed by manually ranked biomedical terms, showing that our proposal obtains the highest correlation with respect to human experts.

1 Introduction

The computation of the semantic similarity/distance between concepts has been a very active trend in computational linguistics. It gives a clue which quantifies how words extracted from documents or textual descriptions are alike. Similarity measures are usually based on *is-a* relations between concepts which are inherent to the concept's semantics. For example, *bronchitis* and *flu* are similar because both are disorders of the respiratory system.

From a domain independent point of view, the assessment of semantic similarity has many direct applications such as, word-sense disambiguation [Resnik, 1999], document categorization or clustering [Cilibrasi and Vitnyi, 2006], word spelling correction [Budanitsky and Hirst, 2006], automatic language translation [Cilibrasi and Vitnyi, 2006], ontology learning [Sanchez and Moreno, 2008] or information retrieval [Lee *et al.*, 1993].

In the biomedical domain, similarity measures can improve the performance of Information Retrieval tasks, since they are able, for example, to map a user's specific search query (e.g. patient cohort identification) to multiple equivalent formulations [Pedersen *et al.*, 2007]. Other authors have applied semantic similarity measures to discover similar protein sequences [Lord *et al.*, 2003] or to the automatic indexing and retrieval of biomedical documents (e.g. the PubMed digital library) [Wilbu and Yang, 1996].

In general, semantic similarity computation is based on the estimation of semantic evidence observed in some knowledge source. That is, background knowledge is needed in order to measure the degree of similarity between a pair of concepts.

Domain-independent approaches [Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997] typically rely on WordNet [Fellbaum, 1998], which is a freely available lexical database that describes and structures more than 100,000 general English concepts, which are stored as an ontology. An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary [Neches *et al.*, 1991]. However, in specific domains it is more appropriate to use domain ontologies that have been built to describe precisely and completely the information related to a certain domain of knowledge. In biomedicine, there exist a growing number of ontologies that organize medical concepts into hierarchies and semantic networks like the Unified Medical Language System (UMLS) of the National Library of Medicine. SNOMED-CT is one of the largest sources included in the UMLS and contains a number of medical concepts interrelated by different conceptual hierarchies corresponding to different scopes (procedures, substances etc) (see section §3).

In the past, some classical similarity computation measures have been adapted to the biomedical domain [Pedersen *et al.*, 2007] by exploiting medical ontologies (SNOMED-CT). In this paper, we expand this study to other classical similarity measures [Wu and Palmer, 1994; Maedche and Zacharias, 2002] based on the exploitation of the ontology's geometric model. Considering the limited performance obtained by previous attempts applied to the biomedical domain, we present a new method which is able to overpass them when evaluated against a benchmark of medical concepts. It is based on the computation of the amount of common taxonomical knowledge between a pair of concepts.

The rest of the paper is organized as follows. Section 2 presents some similarity computation paradigms and the way in which they have been used in the past to deal with biomedical concepts. Section 3 presents classical similarity measures based on the ontology's geometric model and how can they be adapted to use SNOMED-CT as ontology. Section 4 introduces a new measure aimed to provide a better performance than previous approaches in the biomedi-

cal field. In section 5, all the presented measures are evaluated using a standard benchmark composed by 30 medical terms whose similarity has been assessed by expert physicians of the Mayo Clinic [Pedersen *et al.*, 2007]. The final section will present the conclusions of this study and some lines of future work.

2 Related Work

In the literature, we can distinguish several different approaches to compute semantic similarity between concepts according to the techniques employed and the knowledge exploited to perform the assessment. First, there are unsupervised approaches in which semantics are inferred from the information distribution of terms in a given corpus [Etzioni *et al.*, 2005; Landauer and Dumais, 1997]. Statistical analysis and shallow linguistic parsing are used to measure the degree of co-occurrence between terms which is used as an estimation of similarity [Lemaire and Denhire, 2006]. These measures need a corpus as general as possible in order to estimate social-scale word usage. However, due to their completely unsupervised nature and the lack of semantic analysis over the text, they offer a limited performance, specially when dealing with concrete domain such as biomedicine [Pedersen *et al.*, 2007]. This is motivated by the lack of domain coverage of a general domain corpus and the difficulty of compiling a relevant domain corpus big enough to obtain robust statistics.

Other trends exploit structured representations of knowledge as the base to compute similarities. Typically, subsumption hierarchies, which are a very common way to structure knowledge [Gómez-Pérez *et al.*, 2004], have been used for that purpose. The evolution of those basic semantic models has given the origin to ontologies in which many types of relationships and logical descriptions can be specified to formalize knowledge [Pedersen *et al.*, 2007]. In the biomedical field, many domain ontologies are available, being SNOMED-CT or MeSH some of the most successful examples.

From the similarity point of view, there exist ontology-based measures which combine the knowledge provided by an ontology and the Information Content (IC) of the concepts that are being compared. IC measures the amount of information provided by a given term from its probability of appearance in a corpus. Consequently, infrequent words are considered more informative than common ones. Based on this premise, Resnik [Resnik, 1995] presented a seminal work in which the similarity between two terms is estimated as the amount of information they share in common. In a taxonomy, this information is represented by the Least Common Subsumer (LCS) of both terms. So, the computation of the IC of the LCS results in an estimation of the similarity of the subsumed terms. The more specific the subsumer is (higher IC), the more similar the subsumed terms are, as they share more information. Several variations of this measure have been developed [Lin, 1998; Jiang and Conrath, 1997]. They have been adapted by Pedersen et al. [Pedersen *et al.*, 2007] to the biomedical domain by using SNOMED-CT as ontology and a source of clinical data as corpus. Those measures can be affected by the availability of the background corpus and their coverage with respect to the evaluated terms. Data sparseness (i.e. the fact that not enough data is available for certain concepts to reflect an appropriate semantic evidence) is the main problem [Jiang and Conrath, 1997].

Without relying on a domain corpus, other approaches consider taxonomies and, more generally, ontologies, as a graph model in which semantic interrelations are modelled as links between concepts. Several measures have been developed to exploit this geometrical model, computing concept similarity as inter-link distance (also called Path Length) [Wu and Palmer, 1994; Rada *et al.*, 1989; Leacock and Chodorow, 1998]. In the past, this idea has been applied to the MeSH (*Medical Subject Headings*) semantic network [Rada *et al.*, 1989] in order to improve the information retrieval by ranking document from MEDLINE, a corpus made up of abstracts of biomedical journal articles. Taking a similar approach, several authors [Caviedes and J.Cimino, 2004; Nguyen and Al-Mubaid, 2006] developed measures for finding path lengths in the UMLS hierarchy. The advantage of this kind of measures is that they only use a domain ontology as the background knowledge, so, no corpus with domain data is needed. In this paper we centre the study on this kind of measures when applying them to the biomedical field by using SNOMED-CT as ontology.

3 Semantic similarity measures based on the taxonomical structure

In an is-a hierarchy, the simplest way to estimate the distance between two concepts c_1 and c_2 is calculating the shortest *Path Length* connecting these concepts (i.e. the minimum number of links) [Rada *et al.*, 1989].

$$sim_{pL}(c_1,c_2)=min\ \#\ of\ is{-}a\ edges\ connecting\ c_1\ and\ c_2 \quad (1)$$

Several variations of this measure have been developed such as the one proposed by Wu and Palmer [Wu and Palmer, 1994]. They propose a Path Length measure that also takes into account the depth of the concepts in the hierarchy (2).

$$sim_{w\&p}(c_1, c_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \quad (2)$$

, where N_1 and N_2 is the number of is-a links from c_1 and c_2 respectively to the LCS c, and N_3 is the number of is-a links from c to the root ρ of the ontology. It scores between 1 (for similar concepts) to 0.

Leacock and Chodorow [Leacock and Chodorow, 1998] proposed a measure that considers both the shortest path between two concepts (in fact, the number of nodes N_p from c_1 to c_2) and the depth D of the taxonomy in which they occur (3).

$$sim_{l\&c}(c_1, c_2) = -\log N_p/2D \quad (3)$$

Notice, that if the pair of concepts inherits from many is-a hierarchies, all the possible paths between two concepts are calculated but only the shortest one is considered.

Related to the taxonomical aspect of ontologies, another interpretation of these measures is possible, considering

that the similarity is assessed from the minimum number of shared superclasses of the pair of compared concepts.

Having into account these two issues, Maedche and Zacharias [Maedche and Zacharias, 2002] defined the Concept Match (CM) measure (4) based on the definition of the Upward Cotopy(UC)[Maedche *et al.*, 2001].

Definition: The Upward Cotopy (UC) of a concept c_i is restricted to the set of superconcepts (upper concepts of a concept in an is-a hierarchy) of c_i, and the reflexive relationship of c_i to itself. More formally, the $UC(c_i, H^C)$ of a set of concepts C with the associated partial order H^C (all the is-a hierarchies of concepts, that is a directed, transitive relation $H^C \subseteq C \times C$) is defined as:

$$UC(c_i, H^C) = \{c_j \in C | H^C(c_i, c_j) \vee c_i = c_j\}$$

where $H^C(c_i, c_j)$ means that c_i is a sub-concept of c_j.

Concept Match considers a proportion between the number of common UC from the total of UC of both concepts.

$$sim_{CM}(c_i, c_j) = \frac{|UC(c_i, H^C) \cap UC(c_j, H^C)|}{|UC(c_i, H^C) \cup UC(c_j, H^C)|} \quad (4)$$

Classical approaches use those measures relying on WordNet [Fellbaum, 1998] as the ontology to obtain the similarities between terms. However, due to the limited WordNet's coverage of biomedical terms [Burgun and Bodenreider, 2001], the performance obtained in this specific domain concepts is poor [Pedersen *et al.*, 2007]. So, as stated in the previous section, they have been adapted to the biomedical domain by exploiting SNOMED-CT instead of WordNet.

SNOMED-CT (*Systematized Nomenclature of Medicine, Clinical Terms*) is an ontological/terminological resource distributed as part of the UMLS and it is used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerized physician order entry, disease surveillance, image indexing and consumer health information services. It contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into 13 overlapping hierarchies: clinical findings, procedures, observable entities, body structures, organisms, substances, physical objects, physical forces, events, geographical environments, social contexts, context-depedent categories, and staging and scales. Each concept may belong to one or more of those hierarchies by multiple inheritance. Concepts are linked with approximately 1.36 million relationships. In such a complete domain description, *is-a* relationships can be exploited to calculate the similarity between a pair of terms.

4 Superconcept-based distance

Path length-based measures only consider the minimum path between a pair of concepts, omitting the rest of the taxonomical knowledge available in the ontology. For complex taxonomies, such as SNOMED-CT, with thousands of interrelated concepts with multiple hierarchies that classify the concepts, this kind of measures wastes a great amount of knowledge. For this reason, it seems reasonable that a measure that takes into account the whole taxonomical hierarchy involving the evaluated concepts could provide more accurate similarity assessments.

Taking this into account, a measure was defined in [Batet *et al.*, 2008b] based on the amount of non-shared taxonomical information of a pair of concepts considering the complete is-a hierarchy and exploiting ontologies with multiple is-a hierarchies. In particular, this measure is based on the number of non-common superconcepts of the pair of compared concepts. The measure has already shown successful behaviour in the context of finding clusters when ontologies provide additional semantical information for some of the variables used in the objects description [Batet *et al.*, 2008b; 2008a]. In this work the original measure defined in [Batet *et al.*, 2008b] is normalized to take also into account the proportion between common and non-common superconcepts and better behaviour is observed.

The set of superconcepts of a concept c_i is represented by a binary vector $x_i = (x_{i1} \dots x_{in})$, being n the number of concepts of the ontology. Each element x_{ik} represents the existence of an is-a relation between c_i and $c_k, k = 1 : n$, such as:

$$x_{ik} = \begin{cases} 0, \text{if } c_k \notin UC(c_i, H^C) \\ 1, \text{if } c_k \in UC(c_i, H^C) \end{cases}$$

Having a vectorial representation of the concepts, the distance between two concepts c_i, c_j can be defined as the Euclidean distance between the associated vectors x_i, x_j:

$$d(c_i, c_j) = d(x_i, x_j) = \sqrt{\sum_{i=k}^{n} (x_{ik} - x_{jk})^2}$$

In this case, this measure has a very clear interpretation. As the values in the vectors can only be 0 or 1, the difference $(x_{ik} - x_{jk})$ can only be equal to 1 if and only if c_k is a superconcept of c_i and it is not a superconcept of c_j (or viceversa). Therefore, $\sum_{k=1:n}(x_{ik} - x_{jk})^2$ is, in fact, equal to the number of non-shared superconcepts between c_i and c_j.

Based on this interpretation, the distance can be rewritten in terms of the set of superconcepts of c_i (UC) providing a more compact expression, which is more efficient for evaluation in the scope of the treated ontologies with thousands of concepts, and which do not require the explicit construction of the binary matrix associated to the ontology, too big and hardly to manage in big ontologies:

$$\begin{aligned} &d_e(c_i, c_j) = \\ &= \sqrt{|UC(c_i, H^C) \cup UC(c_j, H^C)| - |UC(c_i, H^C) \cap UC(c_j, H^C)|} \end{aligned} \quad (5)$$

It is worth to note that the distance d_e only considers the non-common information of two concepts but does not evaluate the amount of common information. So, it is not capable to distinguish between cases in which the number of common superconcepts between a pair of concepts is small from those cases in which the number of common superconcepts is high. For example, in figure 1 the distance between concepts c_1 and c_2 is equal to the distance between concepts c_3 and c_4. However, it makes sense to modify the definition of d_e in such a way that $d_e(c_1, c_2) < d_e(c_3, c_4)$

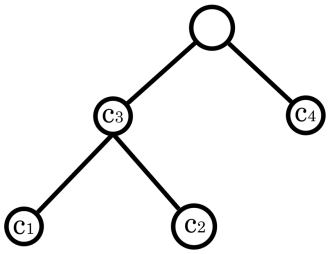


Figure 1: Taxonomy example

owing to the higher number of common superconcepts of the pair (c_1, c_2). This means that c_1 and c_2 are more specific terms that share more is-a relations in the taxonomy.

$$d_e(c_1,c_2)= \\ =\sqrt{|UC(c_1,H^C)\cup UC(c_2,H^C)|-|UC(c_1,H^C)\cap UC(c_2,H^C)|}= \\ =\sqrt{4-2}=\sqrt{2}$$

$$d_e(c_3,c_4)= \\ =\sqrt{|UC(c_3,H^C)\cup UC(c_4,H^C)|-|UC(c_3,H^C)\cap UC(c_4,H^C)|}= \\ =\sqrt{3-1}=\sqrt{2}$$

In order to take into account the number of common superconcepts, d_e is normalised by the total number of superconcepts of c_i and c_j. The sum of common and non-common superconcepts is $|UC(c_i, H^C) \cup UC(c_j, H^C)|$. This permits to include the information about the number of common superconcepts and the Superconcept-based Distance is defined as:

Definition: Superconcept-based Distance (SCD)

$$d_{SCD}(c_i, c_j) = \\ = \sqrt{\frac{|UC(c_i,H^C)\cup UC(c_j,H^C)|-|UC(c_i,H^C)\cap UC(c_j,H^C)|}{|UC(c_i,H^C)\cup UC(c_j,H^C)|}} \quad (6)$$

This definition introduce a desired penalization to those cases in which the number of shared superconcepts is small too. So, we are able to compare a pair of concepts on the basis of the ratio between the non-overlapping and the overlapping taxonomical knowledge between them.

Using the previous example, now the distance between concepts has changed to a better approximation of the real situation. The result is smaller as bigger is the common information, and vice versa.

$$d_{SCD}(c_1, c_2) = \sqrt{\frac{4-2}{4}} = \sqrt{0.5}$$

$$d_{SCD}(c_3, c_4) = \sqrt{\frac{3-1}{3}} = \sqrt{0.66}$$

In the next section, the results obtained with the proposed SCD measure and those presented in section 3 are compared, showing that considering both the amount of common and non-common information between a pair of concepts results in more accurate estimation of semantic similarity for concepts in the biomedical domain.

5 Evaluation

The most common way of evaluating similarity measures is by using a set of word pairs whose similarity has been assessed by a group of human experts and computing their correlation with the results of the computerized measures. In a general setting, the most commonly used benchmark is the Miller and Charles set [Miller and Charles, 1991] of 30 domain-independent word pairs.

For the biomedical domain, Pedersen et al. [Pedersen *et al.*, 2007], in collaboration with Mayo Clinic experts, created a set of 30 word pairs referring to medical disorders. Their similarity was assessed in a scale from 1 to 4 by a set of 9 medical coders who were aware about the notion of semantic similarity and a group of 3 physicians who were experts in the area of rheumathology. For each pair of terms, the averaged scores for each group of experts is presented in Table 1. The correlation between physician judgements was 0.68 and between the medical coders was 0.78.

Table 1: Set of 30 medical term pairs with associated averaged expert's similarity scores (extracted from Pedersen et al.)

Term 1	Term 2	Phys.	Coder
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3
Chronic obstructive pulmonary disease	Lung infiltrates	2.3	1.9
Rheumatoid arthritis	Lupus	2.0	1.1
Brain tumor	Intracranial hemorrhage	2.0	1.3
Carpal tunnel syndrome	Osteoarthritis	2.0	1.1
Diabetes mellitus	Hypertension	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Cortisone	Total knee replacement	1.7	1.0
Pulmonary embolus	Myocardial infarction	1.7	1.2
Pulmonary fibrosis	Lung cancer	1.7	1.4
Cholangiocarcinoma	Colonoscopy	1.3	1.0
Lymphoid hyperplasia	Laryngeal cancer	1.3	1.0
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Rectal polyp	Aorta	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Depression	Cellulitis	1.0	1.0
Varicose vein	Entire knee meniscus	1.0	1.0
Hyperlipidemia	Metastasis	1.0	1.0

We used the same benchmark to evaluate the measures presented in this paper, using SNOMED-CT as the domain ontology. Note that the term pair *"chronic obstructive pulmonary disease" - "lung infiltrates"* was excluded from the test bed as the later term was not found in the SNOMED-

CT terminology.

Table 2: Correlations obtained for each measure against Physicians, Coders and both

Measure	Physician	Coder	Both
Path Length	0.33	0.395	0.386
Wu and Palmer	0.293	0.364	0.353
Leacock and Chodorow	0.453	0.585	0.548
CM	0.56	0.685	0.656
SCD	**0.589**	**0.744**	**0.7**

As some of the measures involved in the test compute similarity (Wu and Palmer, Leacock and Chodorow and Concept-match) and others evaluate dissimmilarity (Path Length and Superconcept-based distance), for a consistent comparison, all the results have been converted into similarity values. So, $sim(c_i) = max_d - d(c_i)$, where max_d is the maximal value that can be obtained by the distance function d [Blanchard *et al.*, 2008]. In this case, max_d corresponds to 2*maximum depth of any taxonomical branch in SNOMED-CT. Note that this conversion does not affect the result of the evaluation, since a linear transformation of the values will not change the magnitude of the resulting correlation coefficient.

The correlation values between the results of the different similarity measures with respect to the human expert scores (including physicians, coders and the averaged scores of both) are presented in Table 2.

Considering the correlation values between human experts (0.68 for physicians and 0.78 for coders) which represent an upper bound for a computerized approach, it can be seen that Path Length-based measures offer a limited performance with correlations smaller than 0.45 and 0.59 respectively. These results show that poor results are obtained when estimating semantic similarity from the minimum inter-concept path in complex domain ontologies, such as SNOMED-CT, where multiple paths between concepts from several overlapping taxonomies are available.

On the other hand, similarities computed using the measures considering much more ontological knowledge (the whole subsumer's hierarchy like CM and SCD) correlate significantly better than Path Length-based ones. Furthermore, the SCD measure has the best performance compared against the others and it is quite close to the correlation between human manual evaluation: 0.589 vs 0.68 in the case of physicians and 0.744 vs 0.78 with respect to medical coders.

In conclusion, the SCD measure is able to extract a robust semantic evidence from highly complex ontologies in biomedicine. The consideration of non common information between concepts and its relative importance with the common information, in addition with smoothing this relation with the root (given by the Euclidean measure), provides a more accurate estimation of the semantic distance.

6 Conclusions

In this paper, we studied the behaviour of several ontology-based semantic similarity measures exploiting the geometrical model of ontologies when applied to the biomedical domain. The main advantage of those measures is that they do not rely on a domain corpus in order to extract semantic evidences. This is especially interesting in domains such as biomedicine in which the access to the required amounts of data is typically difficult due to the sensitivity of medical information. The main drawback is that their performance completely depends on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology. A priori, massive ontologies such as SNOMED-CT with thousands of interrelated concepts with a high degree of taxonomic specialisation are a good knowledge source to apply those measures [Resnik, 1999]. For other more specific domain ontologies with a limited scope, the graph model may be partial; in this case, these measures will be affected by the bias introduced by the partial knowledge modelling [Cimiano, 2006].

Even using a wide ontology like SNOMED-CT, classical approaches based on Path Length have shown a poor performance. Due to the inherent complexity of taxonomical links modelled in that ontology, with relationships of multiple inheritance between concepts, the computation of the minimum path between a pair of concepts only represents a partial view of the modelled knowledge.

In this paper, another measure that takes into account the ratio between the shared and non-shared taxonomicaly related concepts of the compared pair of concepts is analysed. Provided that the semmantics of the proposed expression is very intuitive in terms of the set of superconcepts of the compaired concepts, an equivalent expression of the original Euclidean expression over a set of binary vectors was found. The proposed distance can be rewritten in terms of the set of superconcepts by using UC expressions. This permits, on the one hand, to skip the explicit construction of the binary matrix (quite interesting in big ontologies), and, on the other hand, a very efficient computation since the UC is directly retrievable from the ontology itself.

As shown in the evaluation, with this strategy we are able to clearly outperform previous attempts, exploiting the taxonomic network complexity of SNOMED-CT. At the end, the correlation obtained by our approach with respect to human expert judgements is quite near to the maximum upper-bound (the inter-expert agreement, both for medical coders and physicians), showing the reliability of the obtained results.

After this initial study, we plan to evaluate the SCD measure with other medical ontologies such as UMLS or MeSH. In addition, other non-taxonomic relationships available in those ontologies can be also considered in the future as an statement of concept relatedness. Tests with more reduced domain ontologies will be also interesting in order to evaluate the dependency of the similarity values in relation to the ontology coverage.

Acknowledgments

This work has been partially supported by the K4Care European research project (IST-2004-026968) and the HYGIA project (TIN2006-15453-C04-01). Montserrat Batet is also supported by a research grant provided by the Univesity Rovira i Virgili.

References

[Batet *et al.*, 2008a] M. Batet, A. Valls, and K . Gibert. Improving classical clustering with ontologies. In *Proceed-ings of the*

4th World Conference of the International Association for Statistical Computing, Yoko-hama, Japan, 2008.

[Batet *et al.*, 2008b] M. Batet, A. Valls, and K. Gibert. Measuring similarity in ontologies by means of boolean matrices. Technical Report 2008/3, Universitat Rovira i Virgili, 2008.

[Blanchard *et al.*, 2008] E. Blanchard, M. Harzallah, and P. Kuntz. A generic framework for comparing semantic similarities on a subsumption hierarchy. In *Proceedings of 18th European Conference on Artificial Intelligence (ECAI)*, volume 178, pages 20–24, Patras, Greece, July 21-25 2008. IOS Press.

[Budanitsky and Hirst, 2006] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47, 2006.

[Burgun and Bodenreider, 2001] A. Burgun and O. Bodenreider. Comparing terms, concepts and semantic classes in wordnet and the unified medical language system. In *Proc. of the NAACL 2001 Workshop: WordNet and other lexical resources: Applications, extensions and customizations*, pages 77–82, Pittsburgh, PA, 2001.

[Caviedes and J.Cimino, 2004] J. Caviedes and J.Cimino. Towards the development of a conceptual distance metric for the umls. *Journal of Biomedical Informatics*, 37:77–85, 2004.

[Cilibrasi and Vitnyi, 2006] Rudi L. Cilibrasi and Paul M.B Vitnyi. The google similarity distance. *IEEE Transaction on Know-ledge and Data Engineering*, 19(3):370–383, 2006.

[Cimiano, 2006] P. Cimiano. *Ontology Learning and Population from Text. Algorithms, Evaluation and Appli-cations*. Springer-Verlag, 2006.

[Etzioni *et al.*, 2005] O. Etzioni, M. Cafarella, D. Downey, AM. Popescu, T. Shaked, S. Soderland, DS. Weld, and A. Yates. Unsupervised named-entity extraction form the web: An experimental study. *Artificial Intelligence*, 165:91–134, 2005.

[Fellbaum, 1998] C Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press. More information: http://www.cogsci.princeton.edu/ wn/, Cambridge, Massachusetts, 1998.

[Gómez-Pérez *et al.*, 2004] Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering*. 2nd printing. Springer-Verlag. ISBN: 1-85233-551-3, 2004.

[Jiang and Conrath, 1997] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19–33, Sep 1997.

[Landauer and Dumais, 1997] TK. Landauer and ST. Dumais. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[Leacock and Chodorow, 1998] Claudia Leacock and Martin Chodorow. *Combining local context and WordNet similarity for word sense identification*, chapter WordNet: An electronic lexical database, pages 265–283. MIT Press, 1998.

[Lee *et al.*, 1993] J.H. Lee, M.H. Kim, and Y.J. Lee. Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2):188–207, 1993.

[Lemaire and Denhire, 2006] B Lemaire and G Denhire. Effects of high-order co-occurrences on word semantic similarities. *Current Psychology Letters - Behaviour, Brain and Cognition*, 18(1):http://cpl.revues.org/document471.html., 2006.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In Jude W. Shavlik, editor, *Proceedings of the 15th International Conference on Machine Learning (ICML98)*, pages 296–304, Madison, Wisconson, USA, 1998. Morgan Kaufmann.

[Lord *et al.*, 2003] P. Lord, R. Stevens, A. Brass, and C. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[Maedche and Zacharias, 2002] Alexander Maedche and Valentin Zacharias. Clustering ontology-based metadata in the semantic web. In *Proc. of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, volume 2431 of *Lecture Notes in Computer Science*, pages 348–360. Springer-Verlag, 2002.

[Maedche *et al.*, 2001] Alexander Maedche, Steffen Staab, Nenad Stojanovic, Rudi Studer, and York Sure. Seal - a framework for developing semantic web portals. In Brian J. Read, editor, *BNCOD*, volume 2097 of *Lecture Notes in Computer Science*, pages 1–22. Springer, 2001.

[Miller and Charles, 1991] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[Neches *et al.*, 1991] R. Neches, R. Fikes, T. Finin, T. Gruber, T. Senator, and W. Swartout. Enabling technology for knowledge sharing. *AI Magazine*, 12(3):36–56, 1991.

[Nguyen and Al-Mubaid, 2006] H.A. Nguyen and H. Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *IEEE conference on Granular Computing*, pages 623–628, 2006.

[Pedersen *et al.*, 2007] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40:288–299, 2007.

[Rada *et al.*, 1989] R. Rada, H. Mili, E. Bichnell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):17–30, 1989.

[Resnik, 1995] P. Resnik. Using information content to evalutate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, pages 448–453, Montreal, Canada, 1995.

[Resnik, 1999] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[Sanchez and Moreno, 2008] David Sanchez and Antonio Moreno. Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowledge Engineering*, 63(3):600–623, 2008.

[Wilbu and Yang, 1996] W. Wilbu and Y. Yang. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine*, 26:209–222, 1996.

[Wu and Palmer, 1994] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico, USA, 1994. Association for Computational Linguistics.

Schizophrenia Classification Using Regions of Interest in Brain MRI

D. S. Cheng[1], **M. Bicego**[1], **U. Castellani**[1*], **S. Cerruti**[2,3], **M. Bellani**[2,3]
G. Rambaldelli[2,3], **M. Atzori**[2,3], **P. Brambilla**[†], **V. Murino**[1‡]
[1]Dipartimento di Informatica, University of Verona, Italy
[2]Department of Medicine and Public Health, University of Verona, Italy
[3]Inter-University Centre of Behavioural Neurosciences (ICBN), University of Verona, Italy

Abstract

Is it possible to identify human schizophrenic patients just by analyzing their brain images? This is the fundamental question of magnetic resonance imaging (MRI) based studies of human brains for people affected by schizophrenia and other mental illnesses traditionally diagnosed by self-reports and behavioral observations. The appeal of this approach is at least two-fold: to provide a non-invasive diagnostic tool for mass analyses and early diagnoses, and to characterize mental illnesses with specific and detectable brain abnormalities. Using a dataset of 124 subjects and 7 expert-traced regions of interest (ROI), corresponding to well-known functional parts in the brain, we provide some supportive evidence that this question can be answered positively. By applying several techniques of pattern recognition to the task of discriminating between the 64 patients and the 60 controls, we obtain results and indicative conclusions that find some encouraging agreements with previous medical studies in schizophrenia research.

1 Introduction

Computational neuroanatomy using magnetic resonance imaging (MRI) is a growing research field, which employs image analysis methods to quantify morphological characteristics of different brains [Giuliania *et al.*, 2005]. The ultimate goal is to identify structural brain abnormalities by comparing normal subjects with patients affected by a certain disease.

Roughly speaking, there are two main categories of methods: (i) methods based on the analysis of regions of interest (ROI), and (ii) methods based on Voxel-based-Morphometry (VBM) (see [Ashburner and Friston, 2000]). ROI-based methods focus on a limited set of brain subparts which are manually traced by experts. Methods based on VBM use the whole brain after a normalization procedure that maps the current brain onto a standard reference, called the *stereotaxic* space. In this way, a voxel-by-voxel correspondence is available among the analyzed subjects.

In this work, we apply pattern recognition techniques to the problem of discriminating subjects affected by schizophrenia. We build our framework on top of several previous investigations that confirmed the presence of abnormalities in these subjects [Agarwal *et al.*, 2008; Baiano *et al.*, 2008; Bernasconi *et al.*, 1999; Brambilla *et al.*, 2003; Emmorey *et al.*, 2003; Potkin *et al.*, 2009; Prasad *et al.*, 2005] and extend it to classify healthy (i.e., controls) and unhealthy (i.e., patients) subjects.

Several works have been proposed recently for human brain classification in the context of schizophrenia research [Fan *et al.*, 2007; Gering *et al.*, 2001; Yoon *et al.*, 2007]. Beside standard volumetric methods [Ashburner and Friston, 2000; Baiano *et al.*, 2008], the most promising approaches focus on: (i) shape characterization [Gering *et al.*, 2001], (ii) surface computation [Yoon *et al.*, 2007], and (iii) high dimension pattern classification [Fan *et al.*, 2007]. In [Gering *et al.*, 2001], a ROI-based morphometric analysis is introduced by defining spherical harmonics and a 3D skeleton as shape descriptors. Improvement of such shape-descriptor-based approach with respect to classical volumetric techniques is experimentally shown. In [Yoon *et al.*, 2007], a support vector machine (SVM) has been proposed to classify cortical thickness that has been measured by calculating the Euclidean distance between linked vertices on the inner and outer cortical surfaces. In [Fan *et al.*, 2007], a new morphological signature has been defined by combining deformation-based morphometry with SVM. In this fashion, multivariate relationships among various anatomical regions have been captured to characterize more effectively the group differences.

In this work, we go beyond volumetric measurements, by classifying intensity histograms of the given ROIs. In order to be able to compare intensity values effectively, we perform a preliminary scale normalization based on landmark matching between histograms [Nyúl *et al.*, 2000].

2 Methods

Quantitative data collection and processing in MRI based research implies facing several methodological issues to minimize biases and distortions. The standard approach to dealing with these issues is following well established

*Corresponding author (umberto.castellani@univr.it)
†ICBN, University of Udine, Italy. Department of Pathology and Experimental & Clinical Medicine, University of Udine, Italy. Scientific Institute 'E. Medea', Udine (Italy). CERT-BD, Department of Psychiatry, University of North Carolina (USA).
‡Istituto Italiano di Tecnologia (IIT), Genova, Italy.

	Group mean (and SD)*		Statistics		
	Control	Schizophrenia			
Characteristic	(*n* = 60)	(*n* = 64)	Test	*df*	*p*
Age, yr	39.95 (11.25) [range 23-60]	38.84 (11.96) [range 18-62]	$t = 0.53$	122	0.60
Male/female	32/28	43/21	$\chi^2 = 2.49$	1	0.11
Age at onset, yr		26.28 (9.17)			
Duration of illness, yr		13.37 (10.30)			

SD = standard deviation; *df* = degrees of freedom; *p* = value of significance.
* Unless otherwise indicated.

Table 1: Demographic and clinical characteristics of the study groups. The Student's *t*-test of the age means rejects (at a two-tailed significance level of $p < 0.05$) the hypothesis that the groups are significantly different in age, and Pearson χ^2 confirms the same for the gender differences.

guidelines, dictated by international organizations, such as the World Health Organization (WHO), or codified by respected institutions, such as leading universities. For a detailed description of the procedures followed in collecting and processing the data, please see [Cheng *et al.*, 2009].

2.1 The Dataset

The dataset used in this work originates from a database of MRI scans of hundreds of human brains, containing patients affected by schizophrenia or bipolar disorder (not considered in this particular study), and healthy control subjects. This database has been investigated several times, for example to produce large sample studies aimed at confirming previous reports of physiological abnormalities associated with the given mental illnesses [Agarwal *et al.*, 2008; Baiano *et al.*, 2008; Potkin *et al.*, 2009]. Each of these studies focuses on a particular sub-part of the brain, a so-called *region of interest* (ROI), whose abnormal activity is noted to affect cognitive processes.

In this dataset, we combine data processed from several different studies by creating a single data ensemble where each subject is described by multiple ROIs. In particular, this dataset involves 64 patients with schizophrenia and 60 healthy control subjects (Table 1). Images were acquired and processed on PC workstations for ROI *tracing*, *i.e.* manual annotation of the images, performed by drawing contours enclosing the intended region. It is carried out by a trained expert following a specific protocol for each ROI.

The ROIs contained in this dataset (each presenting two disconnected portions, a left and a right hemisphere part) are the following:

- Amygdala (*amyg* in short);
- Dorso-lateral PreFrontal Cortex (*dlpfc*);
- Entorhinal Cortex (*ec*);
- Heschl's Gyrus (*hg*);
- Hippocampus (*hippo*);
- Superior Temporal Gyrus (*stg*);
- Thalamus (*thal*).

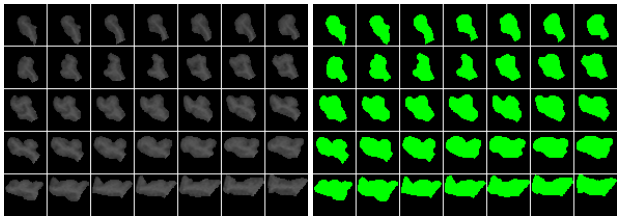


Figure 1: Montage of the slices in the ROI volume ($41 \times 40 \times 35$) of *r_stg* for subject 11. On the left, the MRI values; on the right, the corresponding binary masks.

For the sake of clarity, in the following we separately identify the left and right portion of a ROI, effectively obtaining 14 data records for each of the 124 subjects. We thus number and identify these records in the following way:

1. Left Amygdala (*l_amyg*);
2. Right Amygdala (*r_amyg*);
3. Left Dorso-lateral PreFrontal Cortex (*l_dlpfc*);
4. Right Dorso-lateral PreFrontal Cortex (*r_dlpfc*);
5. Left Entorhinal Cortex (*l_ec*);
6. Right Entorhinal Cortex (*r_ec*);
7. Left Heschl's Gyrus (*l_hg*);
8. Right Heschl's Gyrus (*r_hg*);
9. Left Hippocampus (*l_hippo*);
10. Right Hippocampus (*r_hippo*);
11. Left Superior Temporal Gyrus (*l_stg*);
12. Right Superior Temporal Gyrus (*r_stg*);
13. Left Thalamus (*l_thal*);
14. Right Thalamus (*r_thal*).

In Figure 1, we show a sample from the dataset, specifically the ROI volume of *r_stg* for subject 11. This volume is made up of 35 slices of size 41×40 and can be arranged as a montage of images (ordered from left to right, top to bottom). Within this bounding box, the ROI itself is irregularly shaped, as can be clearly seen from the corresponding binary masks on the right, artificially colored to highlight the ROI shape. Additionally, another important ROI that is traced is the *intracranial volume* (ICV), *i.e.* the volume occupied by the brain in the cranial cavity leaving out the brainstem and the cerebellum. This information is extremely useful for normalizing volume values against differing overall brain sizes.

We included in this study only subjects that featured the complete set of all 14 ROIs, and hence we have a slightly reduced pool compared to previous studies conducted on the same database (*e.g.*, 17 less controls and 6 less patients than [Baiano *et al.*, 2008]).

2.2 Statistical Analysis

An analysis of covariance (ANCOVA) using age, gender and ICV as covariates is usually performed to compare the volumes of ROIs between patients with schizophrenia and healthy normal controls. The purpose of ANCOVA is to find out whether data from several groups have a common

ROI volumes (cm^3)	Group mean (and SD) Control ($n = 60$)	Schizophrenia ($n = 64$)	Statistics F	p
l_amyg	1.46 (0.27)	1.37 (0.28)	3.07	0.08
r_amyg	1.53 (0.27)	1.43 (0.30)	3.74	0.06
l_dlpfc	15.08 (7.07)	14.35 (7.25)	0.42	0.52
r_dlpfc	15.93 (6.83)	13.32 (6.75)	5.73	0.02
l_ec	1.05 (0.22)	1.02 (0.22)	0.67	0.41
r_ec	1.16 (0.23)	1.08 (0.24)	7.27	0.008
l_hg	2.22 (0.70)	2.37 (0.67)	2.33	0.13
r_hg	2.04 (0.58)	2.16 (0.70)	1.33	0.25
l_hippo	1.73 (0.29)	1.75 (0.41)	0.05	0.82
r_hippo	1.77 (0.32)	1.76 (0.33)	0.09	0.76
l_stg	13.75 (1.95)	13.78 (2.17)	<0.01	0.99
r_stg	14.55 (1.98)	14.56 (2.36)	0.01	0.94
l_thal	4.77 (0.52)	4.66 (0.59)	3.10	0.08
r_thal	5.00 (0.63)	5.02 (0.67)	0.17	0.68

SD = standard deviation.

Table 2: Analysis of covariance for ROI volumes in normal controls and patients with schizophrenia show significant differences in the group means only for *r_ec* and *r_dlpfc*.

mean, after adjusting for sources of bias in observational studies. That is, to determine whether the groups are actually different in the measured characteristic when some known sources are factored in. When studying volumetric properties of the brain or parts of it, it is well known that the overall size shrinks with age, it is smaller on average for females, and there is considerable variation between subjects.

ANCOVA retrieves adjusted means for the population groups, in this case patients and controls, that can be tested for significance against the simpler hypothesis that there is no difference between them (the so-called *null* hypothesis). This test is a straightforward application of the F test, with an accompanying p value of significance. From the results reported in Table 2, we can draw the following evidence:

- *r_ec* shows strongly significant differences ($p = 0.008$) between the volume averages for the two groups;
- *r_dlpfc* shows significance at the $p = 0.05$ level;
- some other ROIs are close to being significant, but, overall, the normal distributions of volumes are too close or too overlapping in spread to be usable for discrimination.

These results are well in accordance with previous studies [Agarwal *et al.*, 2008; Baiano *et al.*, 2008].

2.3 MRI Intensity Scale Normalization

A major disadvantage of MRI compared to other imaging techniques is the fact that its intensities are not standardized. Even MR images taken for the same patient on the same scanner with the same protocol at different times may differ in content due to a variety of machine-dependent reasons, therefore, image intensities do not have a fixed meaning [Nyúl *et al.*, 2000]. This implies a significant effect on the accuracy and precision of the following image processing, analysis, segmentation and registration methods relying on intensity similarity.

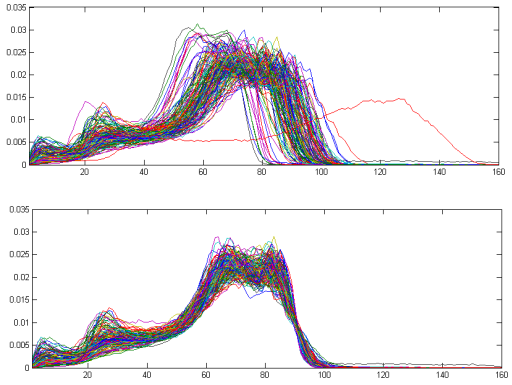


Figure 2: ICV intensity histograms (treated like probability density functions), before (top) and after (bottom) the normalization process.

A successful technique used to calibrate MR signal characteristics at the time of acquisition employs *phantoms* [Edelstein *et al.*, 1984], by placing physical objects with known attributes within the scanning frame. Unfortunately, this technique is not always exploited, which is our present case. Alternatively, it is possible to obtain good results by retrieving deformation mappings for the image intensities, that is, by developing histogram mappings [Jager and Hornegger, 2009; Nyúl *et al.*, 2000].

In this work, we have decided to retrieve the rescaling parameters from the ICV histograms (see Figure 2). In this way, we focus on the interesting content of the images, which usually contain "noise" in the form of bone and muscle tissue surrounding the brain matter proper. It is also easier to identify landmarks on the histograms that match the canonical subdivision of intracranial tissue into white matter, gray matter and cerebrospinal fluid. We have opted to select a simple rescaling mapping that conserves most of the signal in the gray matter - white matter area, corresponding to the two highest bumps in the range 60-90, since ROIs primarily contain those kinds of tissue.

3 Classification Experiments

We performed several classification experiments under varying conditions of histogram pre-processing, feature selection and classifiers to seek the most promising settings for further investigations. Experiments were carried out in Matlab using PRTools [Duin *et al.*, 2007] and accuracies figures for each test run where obtained through leave-one-out (LOO) cross-validation.

In this preliminary work, each ROI was treated independently of the others, much like in common medical analyses, to assess the individual discriminatory capabilities and to be able to effectively compare results with previous medical studies. In the following, all references to "histograms" are intended to be the histograms scale-normalized as in Subsection 2.3.

We tested the following classifiers [Duda *et al.*, 2001]:

- Gaussian radial basis support vector classifier (*svm*), where the standard deviation is estimated by cross validation;

ROI	Best accuracy (classifier,features)	Average accuracy over classifiers and features*
l_amyg	71.0% (svm,13/15)	60.0% (63.7%)
r_amyg	68.5% (knn,13)	57.2% (59.9%)
l_dlpfc	71.0% (svm,1/5)	59.3% (60.9%)
r_dlpfc	70.2% (svm,16)	55.4% (55.0%)
l_ec	66.1% (svm,15)	58.1% (59.0%)
r_ec	66.9% (parzen,13)	58.6% (61.2%)
l_hg	66.1% (svm,12)	56.4% (58.5%)
r_hg	64.5% (knn,16)	53.0% (53.4%)
l_hippo	73.4% (svm,8)	58.8% (61.0%)
r_hippo	65.3% (knn,1/5/12)	53.7% (55.5%)
l_stg	64.5% (knn,13)	56.0% (58.1%)
r_stg	65.3% (knn,10)	55.1% (57.5%)
l_thal	68.5% (knn,10)	55.3% (56.9%)
r_thal	66.9% (knn,4/8/12)	57.0% (58.8%)

Table 3: Best and average accuracies of classification for each ROI. In brackets, the combination of classifier and features (more than one number indicates that multiple settings attained the same accuracy) that achieved the given performance (see text for the numbering of feature methods). * In brackets, the average of only the three best classifiers (*svm*, *knn*, *parzen*).

- Parzen classifier (*parzen*), where the optimum smoothing parameter is retrieved by LOO using the [Lissack and Fu, 1976] estimate for the classification error;
- Fisher's least square linear classifier (*fisher*);
- K-nearest neighbor classifier (*knn*), with K automatically optimized with respect to the LOO error;
- Nearest neighbor (*nn)*;
- Linear Bayes normal classifier (*ldc*);
- Logistic linear classifier (*loglc*).

In addition to using raw histograms, we tried a variety of methods to enhance the discriminatory signal by employing feature selection or processing techniques and dimensionality reduction procedures. The following list shows the full array of methods we devised:

1. raw histograms, *i.e.*, with frequencies of values in natural bins (no quantization);
2. histograms normalized as probability density functions (pdf), to emphasize the relative distribution of frequencies;
3. histograms scaled bin by bin to have zero mean and unit variance (bin scaling), to emphasize deviations from average frequencies;
4. pdfs with bin scaling, to emphasize the relative distribution of deviations;
5. histograms with zero variance bins eliminated, to eliminate constant bins;
6. pdfs with zero variance bins eliminated;
7. as (5) with bin scaling;
8. as (6) with bin scaling;

Classifier	Best accuracy (ROI,features)	Average accuracy over ROIs and features
svm	73.4% (*l_hippo*,8)	59.6%
knn	69.4% (*l_dlpfc*,12)	58.1%
parzen	69.4% (*l_amyg*,16)	57.9%
fisher	67.7% (*l_amyg*,16)	55.8%
loglc	66.9% (*l_hippo*,2/4)	55.6%
nn	69.4% (*l_dlpfc*,12)	55.0%
ldc	67.7% (*l_amyg*,16)	55.1%

Table 4: Best and average accuracies for each classifier. In brackets, the combination of ROI and features that achieved the given performance.

Feature	Best accuracy (classifier,ROI)	Average accuracy over classifiers and ROIs
1	71.0% (svm,*l_dlpfc*)	56.1%
2	66.9% (loglc,*l_hippo*)	54.7%
3	72.6% (svm,*l_hippo*)	55.9%
4	71.8% (svm,*l_hippo*)	55.4%
5	71.0% (svm,*l_dlpfc*)	56.2%
6	66.1% (svm,*l_dlpfc*)	54.8%
7	71.0% (svm,*l_hippo*)	55.7%
8	73.4% (svm,*l_hippo*)	55.3%
9	69.4% (svm,*l_amyg*/*l_dlpfc*)	58.7%
10	68.5% (knn,*l_thal*)	58.8%
11	68.5% (svm,*l_dlpfc*)	58.8%
12	69.4% (knn/nn,*l_dlpfc*)	58.3%
13	71.0% (svm,*l_amyg*)	57.5%
14	66.1% (svm,*l_amyg*)	57.1%
15	71.0% (svm,*l_amyg*)	57.1%
16	72.6% (svm,*l_hippo*)	56.8%

Table 5: Best and average accuracies for each feature method. In brackets, the combination of classifiers and ROIs that achieved the given performance.

9. histograms quantized into 4 bins;
10. pdfs quantized into 4 bins;
11. as (9) with bin scaling;
12. as (10) with bin scaling;
13. dimensionality reduction with principal component analysis (PCA) at 99% of variance of (5);
14. PCA of (6);
15. PCA of (7);
16. PCA of (8).

In Tables 3, 4 and 5, we report a summary of the results we obtained. Table 3 shows the best and average classification accuracies for each ROI over all tested classifiers and feature methods. It gives a rough indication of which ROIs are more discriminatory. Table 4 shows the best and average classification accuracies for each classifier over all feature methods and ROIs, giving a ranking of the most performing classifiers. Finally, Table 5 shows the best and average accuracies for each feature method over all classifiers and ROIs, highlighting the most promising data processing techniques: quantization and PCA. Note that the average accuracies can be used to sort the *relative* discriminatory

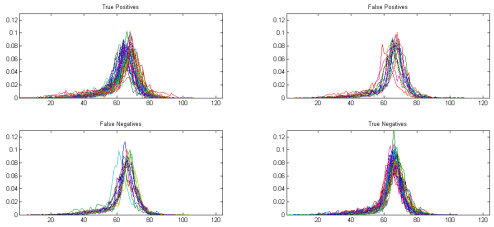


Figure 3: Plots of the subjects' histograms divided according to the classification experiment that achieved the best accuracy (*svm* with feature method 8 on *l_hippo*). From top to bottom, left to right, the four plots identify true positives, false positives, false negatives and true negatives (precision = 0.74, recall = 0.75).

power between different choices (*e.g.*, *svm* is more powerful than *nn* among classifiers), but they have no meaning in absolute terms (59.6% for *svm* is a rather dreadful classification rate, barely above chance).

Overall, results are suggestive and encouraging, in fact they seem to support the main scientific claim that it is possible to identify schizophrenic patients from healthy people. However, best and average accuracies are not conclusive, showing a significant rate of false positives and false negatives (see Figure 3). This confirms how tough the problem is.

Nevertheless, we can draw some limited conclusions:

- *svm* was the best performing classifier, both on average and with the best overall accuracy, obtained on *l_hippo* with 73.4%;
- the three best performing ROIs are *l_hippo*, *l_amyg* and *l_dlpfc* with respectively 73.4%, 71.0% and 71.0%;
- left portions of the ROIs are almost always more discriminative, which is interesting since medical analyses found an overall difference between healthy subjects and schizophrenic patients with higher statistical difference on right ROIs of *amyg*, *hippo* and *dlpfc*;
- while not achieving the best peak performances, quantization seems to enhance discrimination power over all classifiers and ROIs;
- average accuracies are very low, in fact just above chance;
- *nn*, *ldc* and *loglc* always perform worse or equal to the other classifiers, bringing the averages down.

The abnormalities in the amygdala, hippocampus and dorsolateral prefrontal cortex, in particular in the left side, are among the most consistent findings in MRI studies on schizophrenia [Meisenzahl *et al.*, 2008; Shenton *et al.*, 2001]. Suggesting that these structures play a major role for the pathophysiology of the disease [Lopez-Garcia *et al.*, 2006]. In particular, the dorsolateral prefrontal cortex, along with the thalamus and the hippocampus, is a critic component of the brain circuitry underlying higher cognitive functions, such as attention, executive function and context processing [Procyk and Goldman-Rakic, 2006]. The amygdala plays a critical role in the neural system that is involved in emotional and in fear-related responses [Swanson and Petrovich, 1998]; the hippocampus is involved in long term memory and in regulating stress response [Sala *et al.*, 2004; Tulving and Markowitsch, 1998].

4 Conclusions

In this report, we have provided some supportive evidence that it is possible to discriminate between schizophrenic patients and healthy people based on analyses of brain MR images. We have built upon previous medical studies that have focused on volumetric measurements of selected portions of the brain, namely ROIs, corresponding to well-known functional units that affect human cognitive behavior in schizophrenia.

Classification results achieved with several off-the-shelf pattern recognition techniques under different data processing methods seem to suggest that the content of these ROIs, in the form of histograms, can be used to characterize the population of schizophrenic patients. We find results that are consistent with MRI studies in schizophrenia, especially concerning the amygdala, hippocampus and dorsolateral prefrontal cortex, in particular in the left side.

Moreover, we find that *svm* performs best both on peak and average accuracy, but not by a large margin with respect to a simpler classifier such as *knn*. At the same time, the feature extraction and processing methods we employed have increased accuracies, but not dramatically and decisively.

We envisage further work in two main directions: extracting more significant features from the data and building more complex models, starting from smart combinations of all ROI data.

Acknowledgments

We acknowledge financial support from the FET programme within the EU-FP7, under the SIMBAD project (contract 213250).

The dataset used in this work is part of a larger database cared by the Research Unit on Brain Imaging and Neuropsychology (RUBIN) at the Department of Medicine and Public Health-Section of Psychiatry and Clinical Psychology of the University of Verona.

References

[Agarwal *et al.*, 2008] N. Agarwal, G. Rambaldelli, C. Perlini, N. Dusi, O. Kitis, M. Bellani, R. Cerini, M. Isola, A. Versace, M. Balestrieri, A. Gasparini, R. Mucelli, M. Tansella, and P. Brambilla. Microstructural thalamic changes in schizophrenia: a combined anatomic and diffusion weighted magnetic resonance imaging study. *Journal of Psychiatry & Neuroscience*, 33(5):440–448, 2008.

[Ashburner and Friston, 2000] J. Ashburner and K. J. Friston. Voxel-based morphometry - the methods. *Neuroimage*, 11:805–821, 2000.

[Baiano *et al.*, 2008] M. Baiano, C. Perlini, G. Rambaldelli, R. Cerini, N. Dusi, M. Bellani, G. Spezzapria, A. Versace, M. Balestrieri, R. Mucelli, M. Tansella, and

P. Brambilla. Decreased entorhinal cortex volumes in schizophrenia. *Schizophr. Res.*, 102:171–180, 2008.

[Bernasconi *et al.*, 1999] N. Bernasconi, A. Bernasconi, F. Andermann, F. Dubeau, W. Feindel, and D. C. Reutens. Entorhinal cortex in temporal lobe epilepsy: a quantitative MRI study. *Neurology*, 52(9):1870–1876, June 1999.

[Brambilla *et al.*, 2003] P. Brambilla, K. Harenski, M. Nicoletti, R. B. Sassi, A. G. Mallinger, E. Frank, D. J. Kupfer, M. S. Keshavan, and J. C. Soares. MRI investigation of temporal lobe structures in bipolar patients. *Journal of Psych. Res.*, 37:287–295, 2003.

[Cheng *et al.*, 2009] D. S. Cheng, M. Bicego, U. Castellani, S. Cerruti, M. Bellani, G. Rambaldelli, M. Atzori, P. Brambilla, and V. Murino. Schizophrenia classification using regions of interest in brain MRI. Technical report, Dipartimento di Informatica, University of Verona, Italy, 2009.

[Duda *et al.*, 2001] R. Duda, P. Hart, and D. P. Stork. *Pattern Classification*. John Wiley & Sons, 2nd edition, 2001.

[Duin *et al.*, 2007] R. P. W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D. M. J. Tax, and S. Verzakov. *PRTools4.1, A Matlab Toolbox for Pattern Recognition*. Delft University of Technology, 2007.

[Edelstein *et al.*, 1984] W. A. Edelstein, P. A. Bottomley, and L. M. Pfeifer. A signal-to-noise calibration procedure for NMR imaging systems. *Medical Physics*, 11:180–185, 1984.

[Emmorey *et al.*, 2003] K. Emmorey, J. S. Allen, J. Bruss, N. Schenker, and H. Damasio. A morphometric analysis of auditory brain regions in congenitally deaf adults. *Proceedings of the National Academy of Sciences of the United States of America*, 100:10049–10054, 2003.

[Fan *et al.*, 2007] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos. COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Trans. on Medical Imaging*, 26(1):93–105, 2007.

[Gering *et al.*, 2001] G. Gering, M. Styner, and J. Lieberman. Shape versus size: Improved understanding of the morphology of brain structures. In *Medical Image Comp. Computer-Assisted Intervention (MICCAI)*, 2001.

[Giuliania *et al.*, 2005] N. R. Giuliania, V. D. Calhon, V. D. Pearlson, A. Francisd, and R. W. Buchanan. Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia. *Schizophr. Res.*, 74:135–147, 2005.

[Jager and Hornegger, 2009] F. Jager and J. Hornegger. Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 28(1):137–150, January 2009.

[Lissack and Fu, 1976] T. Lissack and K. S. Fu. Error estimation in pattern recognition via l-distance between posterior density functions. *IEEE Trans. Inform. Theory*, 22:34–45, 1976.

[Lopez-Garcia *et al.*, 2006] P. Lopez-Garcia, H. J. Aizenstein, B. E. Snitz, R. P. Walter, and C. S. Carter. Automated roi-based brain parcellation analysis of frontal and temporal brainvolumes in schizophrenia. *Psychiatry Res*, 147(2-3):153–161, Oct 2006.

[Meisenzahl *et al.*, 2008] E. M. Meisenzahl, N. Koutsouleris, R. Bottlender, J. Scheuerecker, M. Jäger, S. J. Teipel, S. Holzinger, T. Frodl, U. Preuss, G. Schmitt, B. Burgermeister, M. Reiser, C. Born, and H. J. Möller. Structural brain alterations at different stages of schizophrenia: A voxel-based morphometric study. *Schizophrenia Res.*, 104(1-3):44–60, Sep 2008.

[Nyúl *et al.*, 2000] L. G. Nyúl, J. K. Udupa, and X. Zhang. New variants of a method of mri scale standardization. *IEEE Trans. Med. Imaging*, 19(2):143–150, 2000.

[Potkin *et al.*, 2009] S. G. Potkin, J. A. Turner, G. G. Brown, G. McCarthy, D. N. Greve, G. H. Glover, D. S. Manoach, A. Belger, M. Diaz, C. G. Wible, J. M. Ford, D. H. Mathalon, R. Gollub, J. Lauriello, D. O'Leary, T. G. Van Erp, A. W. Toga, A. Preda, and K. O. Lim. Working memory and DLPFC inefficiency in schizophrenia: The FBIRN study. *Schizophr. Bulletin*, 35(1):19–31, 2009.

[Prasad *et al.*, 2005] K. M. Prasad, S. D. Sahni, B. R. Rohm, and M. S. Keshavan. Dorsolateral prefrontal cortex morphology and short-term outcome in first-episode schizophrenia. *Psychiatry Research*, 140(2):147–155, 2005.

[Procyk and Goldman-Rakic, 2006] E. Procyk and P. S. Goldman-Rakic. Modulation of dorsolateral prefrontal delay activity during self-organized behavior. *J Neurosci*, 26(44):11313–23, Nov 2006.

[Sala *et al.*, 2004] M. Sala, J. Perez, P. Soloff, S. Ucelli di Nemi, E. Caverzasi, J. C. Soares, and P. Brambilla. Stress and hippocampal abnormalities in psychiatric disorders. *Eur. Neuropsychopharmacology*, 14(5):393–405, Oct 2004.

[Shenton *et al.*, 2001] M. E. Shenton, C. C. Dickey, M. Frumin, and R. W. McCarley. A review of mri findings in schizophrenia. *Schizophr. Res.*, 49(1-2):1–52, Apr 2001.

[Swanson and Petrovich, 1998] L. W. Swanson and G. D. Petrovich. What is the amygdala? *Trends in Neurosciences*, 21(8):323–331, 1998.

[Tulving and Markowitsch, 1998] E. Tulving and H. J. Markowitsch. Episodic and declarative memory: role of the hippocampus. *Hippocampus*, 8(3):198–204, 1998.

[Yoon *et al.*, 2007] U. Yoon, J. Lee, K. Im, W. Shin, B. H. Cho, I. Y. Kim, J. S. Kwon, and S. I. Kim. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. *Neuroimage*, 34:1405–1415, 2007.

Decision Nomograms

Janez Demšar[1], **Aleksander Sadikov**[1], **Tanja Čufer**[2]
[1] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[2] University Clinic Golnik, Medical Faculty Ljubljana, University of Ljubljana, Slovenia

Abstract

We present a new technique for induction of models which can help in selecting the optimal therapy for a particular patient. The model is based on Bayesian reasoning and visualized with a nomogram. To evaluate the method, we present a case study on choosing the chemotherapy for treatment of breast cancer.

Machine learning methods are often used in medical diagnostic and prognostic, while their use in medical decision making is most often only indirect. We can either train models to mimic the gold standard decisions or construct the prognostic models which use the chosen therapy as one of the factors and thus implicitly suggest the therapy leading to the better outcome.

We propose a simple approach inspired by naive Bayesian classifier and, in particular, the work of Možina *et al.* [2004]. We compute how much each possible value of each feature indicates a particular therapy, and visualize the model as a nomogram, so it can be easily printed out and used without a computer.

1 Method

Log odds ratio for conditional probabilities is defined as

$$\log \frac{p(Y=1|T=1)/p(Y=0|T=1)}{p(Y=1|T=0)/p(Y=0|T=0)} \tag{1}$$

In our case, Y will be the outcome, where Y=1 is the *desired outcome* (*e.g.* patient recovers), and T will be the chosen therapy. Both Y and T are binary.

The numerator and denominator can be interpreted as conditional relative "risks": they tell how more likely is the desired class than the undesired. The entire formula is the ratio of ratios. Its value is positive if Therapy 1 is more successful than Therapy 0, and vice versa.

Equation (1) can be simplified to

$$\log \frac{p(Y=1,T=1)\ p(Y=0,T=0)}{p(Y=1,T=0)\ p(Y=0,T=1)} \tag{2}$$

Now consider a specific patient described with feature vector $\mathbf{A} = [A_1, A_2, \ldots, A_n]$. We can rewrite (2) into conditional log odds ratio, with patient's data as condition

$$\log \frac{p(Y=1,T=1|\mathbf{A})\ p(Y=0,T=0|\mathbf{A})}{p(Y=1,T=0|\mathbf{A})\ p(Y=0,T=1|\mathbf{A})} \tag{3}$$

Applying the Bayesian rule gives

$$\log \left(R \times \frac{p(\mathbf{A}|Y=1,T=1)\ p(\mathbf{A}|Y=0,T=0)}{p(\mathbf{A}|Y=1,T=0)\ p(\mathbf{A}|Y=0,T=1)} \right), \tag{4}$$

where

$$R = \frac{p(Y=1,T=1)\ p(Y=0,T=0)}{p(Y=1,T=0)\ p(Y=0,T=1)} \tag{5}$$

Assuming, like in the naive Bayesian classifier, the independence of attributes with regard to the class, we get

$$\log \left(R \times \prod_i \frac{p(A_i|Y=1,T=1)\ p(A_i|Y=0,T=0)}{p(A_i|Y=1,T=0)\ p(A_i|Y=0,T=1)} \right) \tag{6}$$

which equals

$$\log R + \sum_i \log \frac{p(A_i|Y=1,T=1)\ p(A_i|Y=0,T=0)}{p(A_i|Y=1,T=0)\ p(A_i|Y=0,T=1)} \tag{7}$$

The terms being summed represent the contributions of individual attribute values. We shall denote these contributions by CLORC(A_i) (conditional log odds ratio contribution), so (7) becomes

$$\log R + \sum_i \text{CLORC}(A_i) \tag{8}$$

Probabilities needed to compute CLORC(A_i) can be estimated from the training data.

2 Visualization

The model can be visualized in a nomogram (see Figure 1). Each axis in the main part corresponds to a feature and positions of values correspond to their respective CLORCs. Choosing the therapy requires reading the contributions from the nomogram, summing them and adding $\log R$. If the result is positive we choose Therapy 1 and vice versa.

To compute by how much one therapy is better than another (more accurately, what is the ratio of ratios of positive to negative outcomes for the two therapies), we use the bottom part of the nomogram, which represents the function e^{logR+x}: we find the sum of CLORCs on the upper axis and read the ratio on the bottom one.

Similar to naive Bayesian nomograms [Možina *et al.*, 2004], this nomogram:

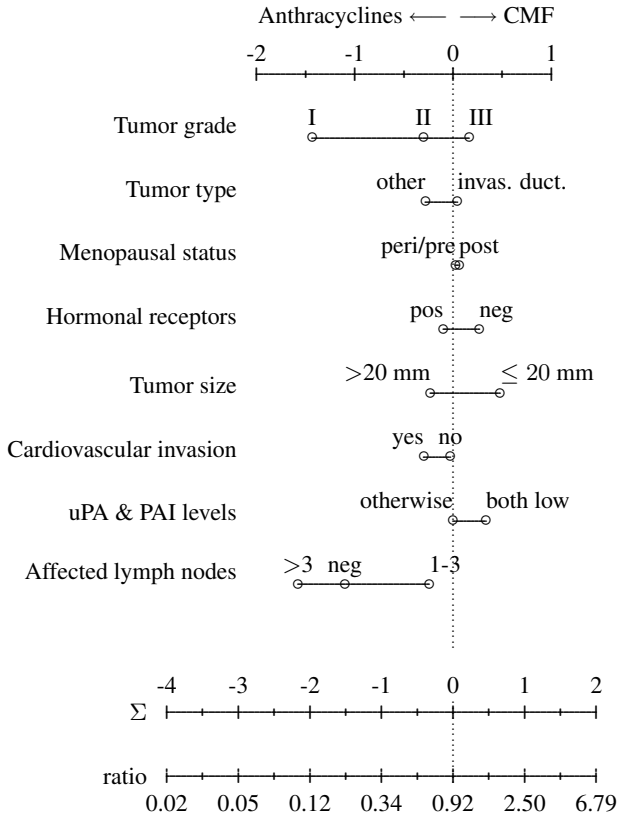


Figure 1: Nomogram for choosing the optimal therapy for treating breast cancer

- tells which features indicate which therapy,
- can be used to choose the optimal treatment for a particular patient,
- gives the arguments for and against each therapy for a particular patient.

3 Case Study

Between 1997 and 2001, 696 patients with early breast cancer were treated at the Institute of Oncology Ljubljana and had their uPA and PAI-1 levels determined in the tumor extract. All of the patients had a histologically confirmed invasive breast cancer. The established clinical and histomorphological factors such as menopausal status, tumor size, tumor type, nodal status, pathological grade, hormonal receptors and tumor vessel invasion were also determined. Treatment decisions regarding primary surgery and adjuvant systemic therapy were based primarily on guidelines valid at the time and did not take into account neither uPA nor PAI-1 levels; 368 (53%) received CMF (cyclophosphamide, methotrexate, and fluorouracil) and 328 (47%) received anthracycline-containing regiments. All patients underwent radical local treatment. In this paper, we consider the treatment as successful if the disease did not recur for at least three years. Figure 1 shows the model built from this data.

First, the nomogram offers some general insight into the problem. It would seem that the tumor grade and the number of nodes are the most important features to be considered. Both can however strongly suggest only the anthracyclines therapy (namely, having a tumor of grade I, or more than 3 nodes, are strong indicators for the anthracyclines chemotherapy). Low uPa and PAI levels indicate the CMF therapy. Small tumor sizes are another indicator for CMF, though their importance is far behind the number of nodes and the tumor grade. Menopausal status is irrelevant for the decision.

Second, given a patient, we can decide for the therapy. Assume we have a patient with grade II tumor (–0.3), negative hormone receptors (+0.3), small tumor size (+0.5) and low uPA and PAI (+0.4), while other data is unknown. The total is +0.9 and with the lower part of the nomogram we discover that the success ratio for the CMF therapy is around 2.3 times greater than that of anthracycline.

Finally, we see that for this particular patient, the only argument against using CMF is the tumor grade.

The model fairly agrees with the standard practice. Anthracyclines are usually used for more advanced cancers, that is, larger tumors ($\leq$ 20 mm), cardiovascular invasion, higher uPA and PAI levels and larger number of affected lymph nodes ($>$ 3). The most obvious discordance is the tumor grade, where grade III should indicate anthracycline therapy and grade I is typically treated with CMF. Another problem is that the number of affected lymph nodes is not ordered correctly, with zero infected nodes placed between 3 and 1-3. We suspect that this is related to the fact that only 25 out of 221 patients with no affected lymph nodes were treated with CMF, so the probability estimates are unreliable. Menopausal status is indeed believed to be unrelated to the success of these specific therapies.

4 Conclusion

The proposed model has a simple mathematical foundation and is easy to understand and use. Its biggest potential problem are unreliable estimates of probabilities needed for its construction: physicians typically choose the better therapy, so the prescribed therapy in such data is not independent from the attributes.

Future work on the method will include its generalization to multiple therapies and outcomes, computation of confidence intervals for contributions and the final ratio, and, most importantly, its evaluation on a larger set of problems.

Acknowledgment

This work was supported by grants from the Slovenian Ministry of Science (J3-4291-0302-03) and Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

References

[Možina *et al.*, 2004] Martin Možina, Janez Demšar, Michael Kattan, and Blaž Zupan. Nomograms for visualization of naive bayesian classifier. In *PKDD '04: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 337–348, New York, NY, USA, 2004. Springer-Verlag New York, Inc.

Clustering of Electronic Medical Records of MRSA Patients

Anna L. Buczak, PhD, Brian Feighner, MD, MPH, Linda J. Moniz, PhD, Joseph Lombardo, MS
The Johns Hopkins University Applied Physics Laboratory (JHU/APL), Laurel, MD, USA
anna.buczak@jhuapl.edu

Abstract

In this paper we describe a data mining methodology developed for grouping patients with *Staphylococcus aureus* (SA) and Methicillin-Resistant *Staphylococcus aureus* (MRSA) infections into clusters according to the pattern of care, the subtype of MRSA, and the outcome of the illness. A new cluster validation methodology is presented. The results of three final clusterings and their medical interpretations are described.

1 Introduction

In recent years, there has been an increased use of information technology to support the public health mission of monitoring for disease and reportable conditions. Until recently, limited health indicator data has been available to public health for testing surveillance algorithms that operate on the entire Electronic Medical Record (EMR).

The aim of this project is to produce fully synthetic background data based on the full EMR from the Biosense program. Before synthetic records can be generated, it is necessary to determine for each of the diseases/sets of symptoms present in the data what sequence of health care events patients are experiencing. Due to a large variability in the care events that patients experience, clustering techniques will be employed to group patients into clusters of similar care procedures. In this paper we concentrate on *Staphylococcus aureus* (SA) and Methicillin-Resistant *Staphylococcus aureus* (MRSA) infections.

2 The Data Mining Methodology

The goal of the data mining methodology developed is to derive (from the available medical records) a care model of how patients are treated for a given illness or a set of symptoms. This method has the following main steps (Fig. 1):

1) Identify a subset of patients of interest (patients who have a set of symptoms of interest).
2) Build Patient Care Instances - sequences of patient care events for each patient identified in step 1.
3) Build Patient Care Descriptors summarizing each Patient Care Instance.
4) Perform clustering on Patient Care Descriptors to detect clusters with similar care instances.
5) Use derived clusters to characterize patient care model for a given illness/set of symptoms.

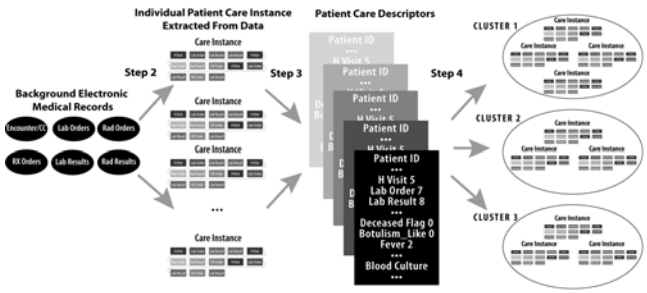


Figure 1 Data mining steps to characterize Patient Care Models.

3 Patient Care Instances

A Patient Care Instance (PCI) is a sequence of the health care encounters present in the data set for a given patient [Buczak *et al*., 2008]. PCI consists of up to 7 types of events: 1) hospital visit (HVisit); 2) lab order; 3) lab result; 4) radiology order; 5) radiology result; 6) Rx (drug) order); and 7) death event. The events in the PCI are sequentially ordered based on the dates and times they occurred. The information in each of the events is extremely detailed. Let us consider a lab order event: it contains information such as test name (e.g., Blood Culture-BLC), specimen type (e.g., blood), order number and date/time of test. A lab result event has even more information: test name (e.g., fungal culture/smear), specimen type (e.g., ELB), order number, date/time of test, order LOINC component (e.g., fungus identified), collection date, etc.

Individual PCIs are of different lengths (depending on the number of visits and specific information in lab orders, lab results, etc.) They range from 90 fields to 58,410 fields. People who came only once and did not have any lab or radiology orders, have a PCI with one record. People who came many times and had many lab or radiology orders and results, have hundreds of records in their PCI.

4 Patient Care Descriptors

The next step is to build the patient care descriptors which summarize the PCIs. For each PCI, one patient care descriptor is computed. Each descriptor has attributes speci-

fying the number of hospital visits, overall lab orders, each of the specific lab orders (e.g., blood culture, respiratory culture, urine culture, Aerobic Culture/Smear, Platelet Auto AB), microorganisms identified (e.g., *Enterobacter Cloacae*, *Staphylococcus Aureus*, MRSA), types of radiology orders (e.g., DX Chest, DX Abdomen), syndromes (e.g., Fever, Gastrointestinal), subsyndromes (e.g., malaise and fatigue, myalgia). The values of all attributes mentioned so far depict how many times a given syndrome / subsyndrome / lab test occurred in the PCI (positive integers or zero). Text attributes include the patient's race (e.g., *White*) and ethnic group (e.g., *Hispanic or Latino*). Table 1 shows an example of a patient care descriptor. For a given data set, each patient descriptor has the same set of attributes allowing for easy subsequent computations.

Table 1 Example Patient Care Descriptor

Patient ID	121845
Age Range	50+
Gender	M
Race	White
Ethnic Group	Not Hispanic or Latino
Deceased Flag	
HVisit	2
Lab Order	3
Lab Result	
Rad Order	1
Rad Result	1
Rx Order	
Botulism-Like	
Fever	
GI	
Hemorrhagic_Illness	
Localized_Cutaneous_Lesion	1
…	
Abdominal Pain	
…	
Diabetes mellitus	1
…	
Heart disease, ischemic	1
…	
Aerobic Culture and Smear	1
Blood Culture	2
C Reactive Protein CRP	1
…	
MRSA Culture (CMRSA)	3
Urine Culture	
Staphylococcus Aureus MR	1
…	
DX Chest	
PX Ankle	2
…	

5 Clustering of Patient Care Descriptors

The clustering method used was hierarchical agglomerative clustering using Ward's linkage. The agglomerative clustering starts with each pattern in a singleton cluster, and successively merges clusters, until a stopping criterion is satisfied. Ward's linkage [Ward, 1963] uses an analysis of variance approach to evaluate the distance between clusters. At each stage, the combination of each cluster pair is analyzed, and the two clusters whose union generates minimum increase in *information loss* are joined. Ward defines the information loss in terms of the error sum-of-squares (ESS) criterion. For a set *X* the *ESS* is described by:

$$ESS\ (X) = \sum_{i=1}^{N_x} | x_i - \frac{1}{N_x} \sum_{j=1}^{N_x} x_j |^2 \qquad (1)$$

where |.| is the absolute value of the norm of a vector. The Ward's linkage function (i.e., distance between clusters X and Y) is described by:

$$D(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)] \qquad (2)$$

where XY is the combined cluster resulting from merging of clusters X and Y; ESS(.) is the error sum of squares described in Eq. (1).

We used SAS to perform clustering and employed the Euclidean distance. The choice of attributes for the distance measure is discussed in Section 9.

6 Cluster Validation

When performing clustering, an important issue is the validation of the clustering results. There are three broad approaches for investigating the cluster validity [Theodoridis and Koutroubas, 1999; Halkihi *et al.,* 2001; Kovacs *et al.*, 2005]. The external method entails that the clustering results are evaluated based on a pre-specified structure that is imposed on the data set and reflects an expert's intuition about that set. The internal criteria approach evaluates the results in terms of quantities that involve the patterns themselves. The relative criteria approach compares the results of clustering obtained by one procedure with those obtained by a different procedure or by the same procedure but with different parameters.

With agglomerative clustering, the main goal of validation is to determine how many underlying clusters are in the data set. In the descriptions below, we will use the notation SS to describe the sum of squares:

$$SS = \sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad (3)$$

SS_w refers to within cluster SS; SS_b refers to between cluster SS; SS_t refers to total (for the whole data set) SS.
Several internal validity indices can be used simultaneously to determine the number of clusters:

- Root Mean Square Standard Deviation (*RMSSDT*): measures how homogenous a given clustering is (lower RMSSTD value usually means better clustering).
- Semi-Partial R squared (*SPR*): measures the loss of homogeneity after merging the difference between the pooled SS_w of the new cluster and the sum of the pooled SS_w values of clusters joined to obtain the new cluster divided by the pooled SS_t for the whole data set. The value of this index should be low.
- Pseudo F statistic [Calinski and Harabasz, 1974]: measure of separation between clusters: $\frac{SS_b/(k-1)}{SS_w/(n-k)}$ where *k* is the number of clusters and *n* is the number of observations. The higher the value of the statistic, the greater the separation between clusters. The cluster solution immediately prior to the decease in Pseudo F should be selected.
- Pseudo T square statistic: $\frac{(SS_{wt} - SS_{wr} - SS_{ws})(n_r + n_s - 2)}{SS_{wr} + SS_{ws}}$ - where SS_{wt}, SS_{wrt}, SS_{ws} represent respectively the within cluster sum of squares in cluster t, r, s; n_r and n_s represent respectively the number of elements of cluster *r* and *s*. Clusters r and s

are joined in this step to form cluster t. A sudden increase of the statistic indicates a joining of two distinct clusters (a cluster solution immediately preceding this increase should be selected).

In our approach to cluster validation we use both internal and external criteria, with special emphasis on the internal ones. The methodology consists of computing all of the above cluster validity measures for each subset of features chosen. Then we determine the underlying cluster number for each method and each feature subset. In several cases a measure hints to several possible numbers of clusters. The next step is to compute a cluster validity value which is set to 1/*n* where *n* is the number of equiprobable cluster choices as determined by a given metric for a given feature subset. For example, if the RMSSTD metric for the feature subset 3 determines that there are 5 or 10 clusters, the cluster validity measure (*ClustConf*) for each of those choices will be ½ (since there are two equiprobable choices).

We define the *Cluster FS (Feature Set) Confidence* as:

$$ClusterFSConf(i, j) = \sum_{k=1}^{NumMetrics} ClusterConf(i, k, j) \quad (4)$$

where *i* is the Feature Set (FS) number, *j* is the number of clusters, and *k* is the metric number.

The Cluster Metric Confidence is defined as:

$$ClusterMetricConf(k, j) = \sum_{i=1}^{NumFeatureSets} ClustConf(i, k, j) \quad (5)$$

We define the overall confidence for a given number of clusters as:

$$OverallClustConf(j) = \sum_{i=1}^{NumFeatureSets} ClusterFSConf(i, j) \quad (6)$$

The choice of the underlying number of clusters is performed based on the highest values of the *Overall Clust Conf*. For each number of clusters corresponding to three top *Overall Clust Conf* values, the feature set with the maximum *Cluster FS Conf* is chosen and the results of that clustering undergo external cluster validity investigation. This investigation is performed by a medical expert.

7 Feature Selection

The problem of cluster validity and choice of features for clustering are interrelated. The result of clustering with one set of features can diverge from the result with a second set since the distance measures are different.

Not only little research was done on feature selections for unsupervised data [Dash and Liu, 2000; Kim *et al.*, 2002] but additionally the methods developed for this purpose assume that the goal, when searching for the best feature subset, is to obtain a set of attributes that gives the clustering as similar as possible to the one obtained with all the features [Boudjeloud and Poulet, 2005]. Other methods assume that the set of features is constant. In our case, the set of features is not constant, the clustering obtained with all the features is not necessarily the best, and the goal is to find the best clustering possible without the medical experts having to go through all the care models. Our goal is to devise a methodology that uses clustering techniques and metrics, and results in a reasonable clustering from a medical standpoint while the experts have to look only at a small subset of the patients' records.

8 Data Set

The data set contains the anonymized medical records of 458,346 patients and covers over one year of visits. As a proof-of-concept, we concentrated on patients with SA or MRSA infections. MRSA strains, referred to as 'superbugs' because of their resistance to most antibiotics, have become an important concern for hospital and public health personnel. The Centers for Disease Control and Prevention (CDC) estimates that in 2005 there were 100,000 cases of invasive MRSA causing about 18,650 deaths. MRSA can be a cause of skin infections and invasive infections among adults and children [Gorwitz *et al.*, 2006]. Historically, two types of MRSA have been described: community acquired (CA) and healthcare acquired (HA). In the past, the bacterial strains and course of the disease caused by CA MRSA were felt to be distinct from HA MRSA. Increasingly the distinction between CA and HA infections has blurred.

We extracted from the data set 9984 patient records that had lab tests performed that identified the presence of SA or MRSA. This represented 2743 patients. The rest of this research is performed on this data set. To reduce the sparseness of the data set, we collapsed 1195 features into 883. This was performed by adding the values of similar attributes. As an example, the attribute DX-C-Spine was created as the sum of the following features: DX-C-Spine-1-View, DX-C-Spine-2-View, DX-C-Spine-2-or-3-Views, which are different types of cervical spine X-rays.

9 Distance Measures

The next step was to define a distance measure between patient care descriptors for clustering. We used the Euclidean distance with weights set to 1 for chosen features. We excluded certain features from the distance computation because they had nothing to do with MRSA and related illnesses; this included most types of radiology tests: DX Abdomen, DX Knee, etc.

We used 13 different sets of features for computing the distance. They range from all the features (883) to a subset of only 9 MRSA-related features. Most of the feature sets included: Deceased Flag, HVisit, Lab Order, Lab Result, Radiology Order, Radiology Result, and Rx Order. Many of the sets included syndromes (Botulism-like, Fever, Gastrointestinal, Hemorrhagic Illness, Localized Cutaneous Lesion, Neurological, Rash, Respiratory, Severe Illness or Death). Some of the subsets included several types of lab tests such as: Blood Culture, Isolator Blood Culture, Respiratory Culture, Resp-Viral Culture, Urine Culture, Antibody Screen, and MRSA PCR Screen. Some of the sets included microorganisms such as SA or MRSA. Other sets included certain radiology tests: DX Chest, PX Chest, DX Feeding Tube Placement, or PX Central Line Placement. Some sets included subsyndromes that could be related to MRSA: Specific Infection, Lymphadenitis, Lymphadenopathy, Skin infection, Urinary tract infection, or Pneumonia.

The two unique feature sets for clustering were all the features, and MRSA-only features (respectively, FS1 and FS9 in Table 2). The other feature sets were chosen based on information related to the severity of illness, general types of lab tests, information from the literature on MRSA symptoms, medical expert input, and results of clustering with all the attributes (FS1). When means and standard deviations for a feature of the two clusters identified with FS1 were substantially different, a given feature was included at least in one feature subset. An example is Lab Order (its mean in cluster 1 was 5.77, and its mean in cluster 2 was 21.36). In comparison, the feature *Achromobacter Xylosoxidans* was never included in a feature subset, since cluster 1 mean was 0.0018 and cluster 2 mean was 0.0067.

Table 2 Number of Clusters Determined by Different Validation Methods for Each Feature Set

Feature Set	Semi-Partial R Sq	RMSSTD	Pseudo F	Pseudo T Sq
FS1	2	6	2	2 or 7
FS2	3 or 6	5 or 10	3	3 or 6
FS3	6	5 or 10	2 or 6	6 or 10
FS4	3	2 or 11	3	12
FS5	3	6 or 11	3	3 or 12
FS6	2	4	2 or 4	2 or 5
FS7	3	5 or 10	3	3 or 6 or 11
FS8	2	3 or 5	3 or 6	6 or 8
FS9	2	2 or 4	2 or 6	5 or 11
FS10	2 or 3	4 or 6	2	?
FS11	2	2 or 10	2	3
FS12	3 or 5	5 or 11	2	3 or 6
FS13	6	5	2	3 or 6

10 Results

Table 2 describes the number of clusters identified by each of the four cluster validation methods for each of the subsets. The number of clusters identified ranges from 2 to 12. Usually one or two equiprobable numbers of clusters are determined by a metric for a given feature subset.

Table 3 lists the *Cluster FS Conf* values for the number of clusters varying from 1 to 12. The *Overall Clust Conf* is shown in Fig. 2. Its value is much higher for 2, 3, and 6 clusters than for others. We find from Table 2 for 2, 3, and 6 clusters which feature set gave the highest *Cluster FS Conf* in each case. This is FS1 for 2 clusters, FS5 for 3 clusters, and FS3 for 6 clusters. These three clustering results underwent external evaluation by the medical expert. During the external evaluation, a medical doctor received 88 (3% of the total) patient care instances chosen in such a way that for each of the three clusterings there were at least 10 PCIs in each cluster. He reviewed the records in a masked fashion. His goal was to come up with medical criteria and determine how he thought the care instances should be clustered. He came up with subcategories such as hospital acquired (HA) infection, community acquired (CA) infection, immunosuppressed, invasive disease, cutaneous disease, trauma, and death, among others. The categories are not mutually exclusive. Once he categorized all the 88 patient care models, we gave him the results of the three clusterings to determine how much sense, from a medical standpoint, each one of them made.

Table 3 Cluster Feature Set Confidence Values for Number of Clusters 1 to 12

	Cl1	Cl2	Cl3	Cl4	Cl5	Cl6	Cl7	Cl8	Cl9	Cl10	Cl11	Cl12
FS1	0	*2.5*	0	0	0	1	0.5	0	0	0	0	0
FS2	0	0	2	0	0.5	1	0	0	0	0.5	0	0
FS3	0	0.5	0	0	0.5	*2*	0	0	0	1	0	0
FS4	0	0.5	2	0	0	0	0	0	0	0	0.5	1
FS5	0	0	*2.5*	0	0	0.5	0	0	0	0	0.5	0.5
FS6	0	2	0	1.5	0.5	0	0	0	0	0	0	0
FS7	0	0	2.33	0	0.5	0.33	0	0	0	0.5	0.33	0
FS8	0	1	1	0	0.5	1	0	0.5	0	0	0	0
FS9	0	2	0	0.5	0.5	0.5	0	0	0	0	0.5	0
FS10	0	1.5	0.5	0.5	0	0.5	0	0	0	0	0	0
FS11	0	2.5	1	0	0	0	0	0	0	0.5	0	0
FS12	0	1	1	0	1	0.5	0	0	0	0	0.5	0
FS13	0	1	0.5	0	1	1.5	0	0	0	0	0	0

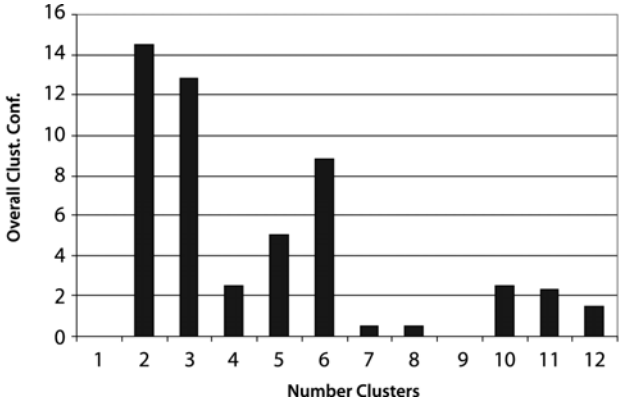


Figure 2 Overall *Cluster Conf* value.

10.1 Results for FS1

The clustering performed with all features (FS1) determines 2 clusters. The dendrogram (Semi-Partial R Square) for this set of hierarchical clusterings is shown in Fig. 3 and points to two clusters. The biggest drop in RMSSTD (Fig. 4) points to 6 clusters. Pseudo F and Pseudo T squared are depicted in Fig. 5. The first points to 2, while the second points to 2 or 7 clusters.

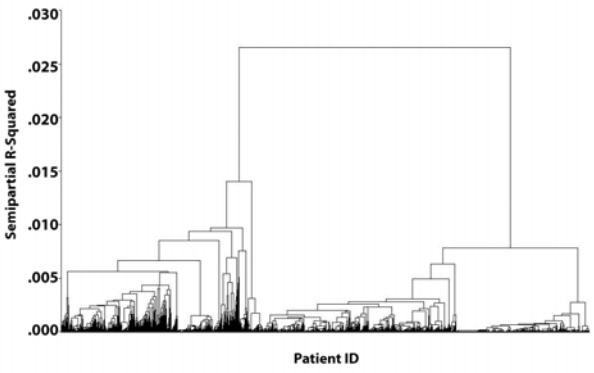


Figure 3 Dendrogram with Semi Partial R Square for FS1.

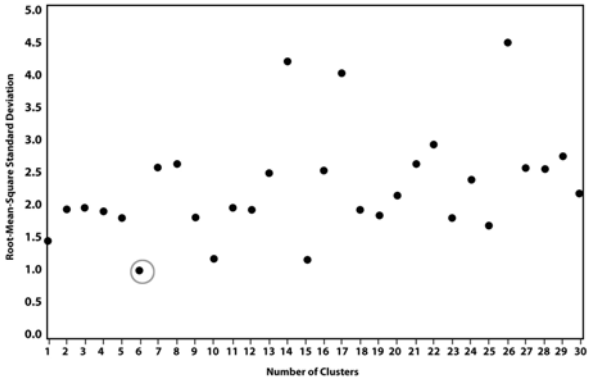


Figure 4 RMSSTD for FS1.

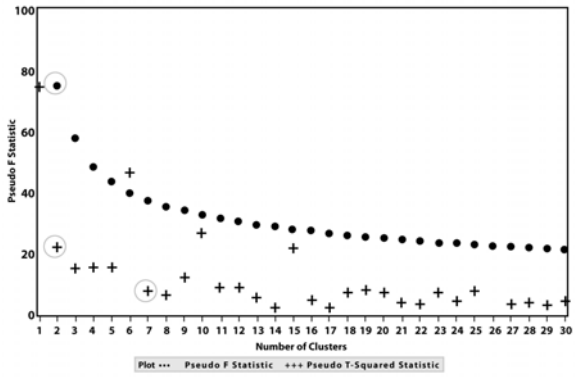


Figure 5 Pseudo F and Pseudo T Square for FS1.

Cluster 1 patients had mainly cutaneous disease and very few died (0.06%). About half of this cluster had evidence of CA disease, although there was more evidence of post-operative complication disease than in Cluster 2. Patients were more likely to have a history of trauma preceding SA disease. Patients from Cluster 1 had a much lower number of lab orders, lab results, radiology orders, radiology results, and Rx orders than the patients from Cluster 2 (see Table 4). Cluster 2 patients were much more severely ill from a variety of disease processes (e.g., HIV infection). They overwhelmingly had invasive disease and more of them died (7.77%). Very few had evidence of CA disease and they were more likely to have HA, but not post-operative HA infection.

Table 4 Mean Characteristics for 2 Clusters Obtained with FS1

Cluster Number	Num Elements	Deceased Flag	HVisit	Lab Order	Lab Result	Rad Order	Rad Result	Rx Order
1	1700	0.0006	3.60	5.77	6.58	3.32	3.09	4.97
2	1043	0.0777	5.37	21.36	21.53	13.04	11.52	14.90

10.2 Results for FS5

The second highest *Overall Clust Conf* (Table 3) points to 3 clusters and Table 2 points to FS5 in this case. This feature subset includes Deceased Flag, HVisit, Lab Order, Lab Result, Radiology Order, Radiology Result, Rx Order, all 9 syndromes, all MRSA-related tests and results, certain cultures and subsyndromes.

Essentially all Cluster 1 patients had cutaneous disease, many with a preceding history of trauma. Greater than 65% had evidence of CA disease. There were no deaths. These patients had the smallest number of lab orders, lab results, radiology orders, radiology results, and Rx orders (see Table 5). Cluster 2 patients had moderate disease with some deaths (0.12%). MRSA infection was split between invasive and cutaneous. About 25% was CA. These patients represented a wide variety of disease etiologies. Cluster 3 was similar to Cluster 2 in that the patients had a variety of disease etiologies, albeit much more severely ill and all of them died. About half of them had invasive disease. None had evidence of CA disease or MRSA from post-operative complications.

Table 5 Mean Characteristics for 3 Clusters Obtained with FS5

Cluster Number	Num Elements	Deceased Flag	HVisit	Lab Order	Lab Result	Rad Order	Rad Result	Rx Order
1	1048	0	2.92	4.10	4.90	1.90	1.70	2.95
2	1615	0.0012	5.31	16.54	16.90	10.11	9.12	12.56
3	80	1	1.24	13.39	14.96	11.46	9.43	7.69

10.3 Results for FS3

The third highest *Overall Clust Conf* (Table 3) points to 6 clusters and Table 3 points to FS3 in this case. This feature subset includes Deceased Flag, HVisit, Lab Order, Lab Result, Radiology Order, Radiology Result, Rx Order, all 9 syndromes, all MRSA-related tests and results, SA, Blood Culture, Isolator Blood Culture, Respiratory Culture, Resp-Viral Culture, Urine Culture, Antibody Screen, DX Chest, PX Chest, certain subsyndromes such as Specific Infection, Lymphadenitis, Lymphadenopathy, Skin infection, and Urinary tract infection.

Cluster 1 contained primarily patients with CA cutaneous MRSA infections. No patients died. There were cases with both cutaneous disease and invasive disease. Histories of mental disease, immunosuppression, and trauma were common. These patients had the smallest number of lab orders, lab results, radiology orders, radiology results, Rx orders (see Table 6). Cluster 2 patients appeared to be moderately ill. There were about equal numbers of CA and HA disease. Cutaneous MRSA was more common. Only one person died (0.1%). There were multiple reports of positive SA cultures that were not identified as MRSA. There was a range of presentations, patient encounter processes and underlying etiologies. Cluster 3 had a predominance of patients with positive urine cultures, which suggests patients with indwelling urinary catheters. HA infections were more common than CA, which is also consistent with long-term patients. There was some cutaneous disease. Nobody died. These patients had the second highest number of urine cultures and blood cultures. Cluster 4 patients were very seriously ill but nobody died. Several patients were presumed to have HIV infection or other serious immunodeficiency. There was a strong infectious disease feel to these patients. All Cluster 5 patients died. Some deaths were from MRSA, most were not. Coronary artery disease was the most common cause of death. SA infection, though almost always invasive, often seemed to be of little significance. This group exhibited the fewest number of hospital visits. Cluster 6 patients had almost entirely invasive MRSA disease. 1.7%

of them died and most were very seriously ill. These were very sick patients who were ill from infectious causes as opposed to Cluster 5 patients who were sick from non-infectious causes who happened to get a positive SA or MRSA culture. These patients had the most lab orders, lab results, radiology orders, radiology results, Rx orders, MRSA Cultures, MRSA-related tests, Blood Cultures, respiratory cultures, urine cultures, DX Chest, PX Chest, Isolator Blood cultures, and Resp-viral cultures.

Table 6 Mean Characteristics for 6 Clusters Obtained with FS3

Cluster Number	Num Elements	Deceased Flag	HVisit	Lab Order	Lab Result	Rad Order	Rad Result	Rx Order
1	1058	0	2.95	3.85	4.54	1.90	1.81	3.12
2	1001	0.001	5.11	14.32	15.13	9.33	8.38	10.37
3	392	0	5.22	14.82	14.42	7.82	7.06	7.24
4	38	0	6.37	13.63	14.24	7.26	7.84	7.32
5	78	1	1.22	12.77	13.27	11.36	9.42	7.88
6	176	0.017	6.30	36.09	36.68	20.80	17.94	37.40

The best clustering in terms of the external criteria is the partitioning into 6 clusters obtained with feature set FS3. From the point of view of cluster validity metrics, this partitioning has the third highest *Overall Cluster Conf* value. This grouping is superior from a clinical / epidemiological standpoint because it not only grouped patients with similar health care processes, but also separated them out by underlying disease processes. This clustering separated those who died *from* MRSA from those who died from other causes *with* MRSA infection.

11 Conclusions and Future Work

We describe a novel data mining methodology for clustering MRSA patients according to the patterns of care they experienced, the subtype of MRSA they are sick with, and the clinical outcome. The first step is extracting patterns of care from the EMR data, the second step is building Patient Care Descriptors that summarize them, and the final step is the clustering of those descriptors. When performing clustering, we faced a number of challenges: the number of underlying clusters is unknown, the number of attributes is very large, the clustering obtained with all the attributes is not the best from a medical standpoint, and the number of patients is prohibitively large for a medical expert to go through all of them. In our approach to cluster validation we employed a combination of internal and external criteria. We defined the *Overall Cluster Conf* and other metrics so we could choose the most meaningful clusterings. The external evaluation criteria were employed on those results. The resulting 6 clusters are meaningful from a medical standpoint and very useful for the generation of synthetic medical records for people with any form of MRSA.

In future work, we will further refine this methodology to obtain medically meaningful clusterings for any type of disease present in the data. This will require the development of an automatic technique for choosing the features for the distance measure used in clustering.

Acknowledgments

This research was supported by Grant Number P01 HK000028-02 from the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC.

References

[Boudjeloud and Poulet, 2005] Lydia Boudjeloud and Francois Poulet. Attribute selection for high dimensional data clustering. *ASMDA 2005 Conference*, pages 387-395, May 2005.

[Buczak *et al.*, 2008] Anna L. Buczak, Linda J. Moniz, John Copeland, Henry Rolka, Joseph Lombardo, Steven Babin, Brian Feighner. Data-driven hybrid method for synthetic electronic medical records generation. In: *Proc. of IDAMAP*, pages 81–86, 2008.

[Calinski and Harabasz, 1974] R.B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[Dash and Liu, 2000] Manoranjan Dash and Huan Liu. Feature selection for clustering. *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 110–121, 2000.

[Gorwitz *et al.*, 2006] Rachel J. Gorwitz, Daniel B. Jernigan, John H. Powers, and John A. Jernigan. Strategies for clinical management of MRSA in the community: summary of an experts meeting. March 2006. http://www.cdc.gov/ncidod/dhqp/ar_mrsa_ca.html.

[Halkihi *et al.*, 2001] Maria Halkihi, Yannis Batiskis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[Kim *et al.*, 2002] YongSeog Kim, W. Nick Street, and Filippo Menczer. Evolutionary model selection in unsupervised learning. *Intelligent Data Analysis*, 6(6):531–556, 2002.

[Kovacs *et al.*, 2005] Ferenc Kovacs, Csaba Legany, Attila Babos. Cluster Validity Measurement Techniques. In: Proc. Sixth Int'l Symp. Hungarian Researchers on Computational Intelligence (CINTI), 2005.

[Theodoridis and Koutroubas, 1999] Sergios Theodoridis and Konstantinos Koutroubas. *Pattern Recognition*. Academic Press. 1999.

[Ward, 1963] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Using Pseudo Time-Series Trajectories to Explore Disease Regions in Glaucoma

Yuanxi Li[1], **David Garway-Heath**[2*] **and Allan Tucker**[1]

1 School of Information Systems, Computing and Mathematics, Brunel University, West London, UK

2 NIHR Biomedical Research Centre for Ophthalmology,

Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK, London, UK

{yuanxi.li,allan.tucker}@brunel.ac.uk

Abstract

Previously, we have developed methods to build *pseudo time-series* from cross-sectional medical data. In this paper we extend these ideas to automatically identify disease regions of interest at the extremes of trajectories, and explore how trajectories differ between data generated from different glaucoma tests.

1 Introduction

Progressive loss of the field of vision is characteristic of a number of eye diseases such as glaucoma, a leading cause of irreversible blindness in the world. There are considerable data being collected on patients who suffer from eye disease such as Visual Field (VF) test data and the Heidelberg Retina Tomograph (HRT) data. Many of these datasets are cross-sectional and the time dimension is not measured due to the expensive nature of such studies across an entire population, despite the inherently temporal nature of eye diseases. Previously, we have developed methods to build *pseudo time-series* from cross sectional data using a combination of distance metrics, graph theoretical operations and bootstrapping [Tucker and Garway-Heath, 2009]. This results in trajectories through the dataspace starting at healthy data and ending at cases of advanced disease. Here we extend some of these ideas in order to automatically identify disease regions of interest along these trajectories as well as the likely transitions between them. We also explore how HRT and VF data interact within the disease process.

2 Methods

The VF test assesses the sensitivity of the retina to light. It is typically measured by automated perimetry, a technique in which the subject views a dim background as brighter spots of light are shone onto the background at various locations in a regular grid pattern. The brightness at which the subject sees the spots of light is related to the retinal sensitivity. For this paper, the VF data are aggregated into average values based upon their association with one of 6 *nerve fibre bundles* based upon the mappings in [Garway-Heath *et al.*, 2000]. The other type of data that we explore are HRT data [Kamal *et al.*, 2000] and involves generating images of the retina in order to calculate certain measurements associated with the three dimensional shape of the optic nerve head. These include neuro-retinal rim area which are used for the experiments in this paper. The measurements are calculated for 6 different segments: nasal, nasal inferior and superior, temporal, temporal inferior and superior. Figure 1 shows an example of a VF test and an image from which HRT data are collected. Here we combine the VF and HRT datasets (we have data for HRT and VF on each of 180 patients) to see if trajectories that are identified capture the interaction between the two data types as glaucoma progresses. The data are classified into healthy or glaucomatous based upon the VF data using an AGIS algorithm [AGIS, 1994].

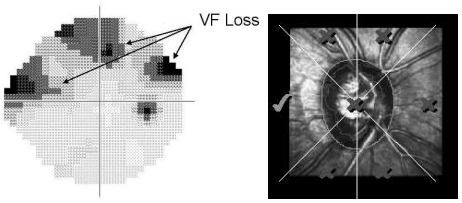


Figure 1: VF data from someone suffering typical glaucomatous field loss and an HRT image

Building pseudo time-series involves plotting trajectories through cross-sectional data based upon distances between each point using prior knowledge of healthy and disease states. These trajectories can then be used to build temporal models such as Hidden Markov Models to make forecasts. The temporal bootstrap involves resampling data from a cross-sectional study and repeatedly building these trajectories through the samples in order to build more robust time-series models [Tucker and Garway-Heath, 2009]. In this paper, we use the following algorithm to generate and explore the transitions of the different states within the trajectories that are discovered from the combined VF and HRT data. Essentially, it starts by searching for 3 hidden states, h (one more than the original 'healthy' and 'disease'), whilst learning the HMM with the EM algorithm [Bilmes, 1997]. This is then repeated for increasing values for h until more than one stable end state is found in the transition matrix of the HMM.

*David Garway-Heath is part funded by the Department of Healths National Institute for Health Research Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health.

Algorithm 1 Disease Region Identification

1: Input: Pseudo time-series generated from the combined VF and HRT data
2: Unlabel the binary disease class states
3: h=3
4: **repeat**
5: Train a HMM on the PMTS with h hidden states using EM
6: **until** transition table in the paramterised HMM has more than one stable disease region (or 'end state')
7: h = h + 1
8: Output: HMM with new multiple hidden 'end states'

3 The Experiments and Results

The 'end states' represent expected values for data based upon the HMM learnt from the unlabelled pseudo time-series. These are likely to represent stable stages of glaucoma data at the extreme of each trajectory found from the pseudo time-series. Firstly we look at the state transition diagram generated from the transition matrix of the learnt HMM in the top of Figure 2. Note that 4 states were found that resulted in two end states. Full lines represent probabilities of over 0.15 and dashed represent probabilities between 0.05 and 0.15. The full transition matrix is given in Table 1 where h_t represents the hidden state h at time t. This table and diagram show that

$h_{t-1} \backslash h_t$	1	2	3	4
1	0.9747	0.0253	0.0000	0.0000
2	0.0509	0.7461	0.0897	0.1133
3	0.2749	0.0000	0.6827	0.0424
4	0.0000	0.0700	0.0990	0.8310

Table 1: Transition matrix for discovered states

there appear to be three relatively stable states: 1, 2 and 4 (where the probability of moving to another state are less than 0.15). 4 coincides with the starting healthy state, 1 and 2 appear to represent relatively stable end states and 3 appears a transitory state. We now explore these states further by calculating the expected values of the variables associated with each state as shown in the bottom of Figure 2 (plotted as VFs and rim measurements spaced anatomically). These values are calculated using a junction tree inference algorithm on the HMM where the evidence entered is simply the end state [Murphy, 2002]. As expected state 4 shows a normal rim width and the healthiest VF sensitivity, whereas state 1 shows marked diffuse rim narrowing, especially in temporal, temporal superior and temporal inferior sectors, and moderate loss of retinal sensitivity, especially in inferior arcuate and superior arcuate regions. This is what would be expected, based on known anatomical relationships. State 3 displays mild diffuse narrowing of rim and mild diffuse reduction of retinal sensitivity. State 2 shows moderate narrowing of rim, especially in nasal and temporal superior sectors and temporal sector. Here one would expect temporal rim sector loss.

4 Conclusions

In this paper, we explore how to build time-series models from cross-sectional glaucoma data, how two different

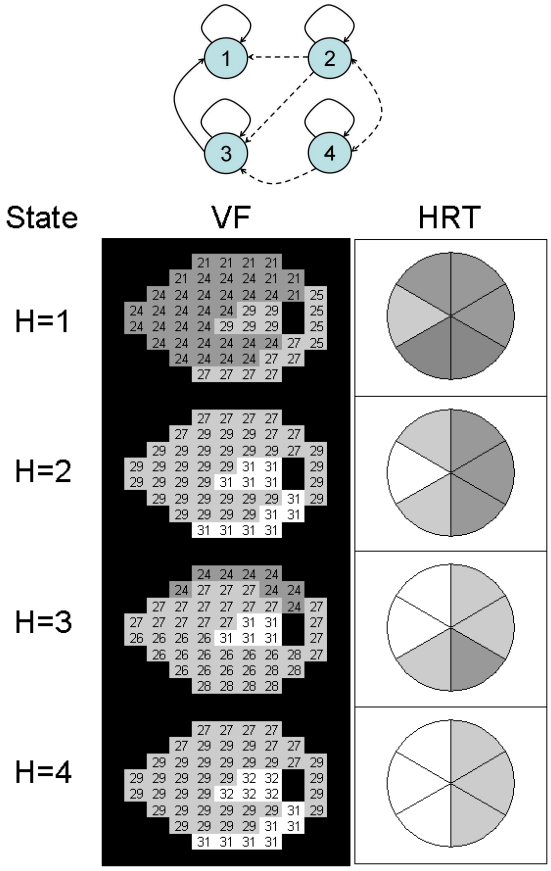


Figure 2: (top) The state transitions for the combined VF and HRT data and (bottom) their expected data.

types of this data interact as the disease progresses, and finally how to automatically identify different disease states at the end of the discovered trajectories. We have carried out a small study with promising results, being able to identify stable states with abnormal VF sensitivity and marked rim narrowing and transitory states with moderate narrowing of rim, especially in the nasal and temporal superior sectors and temporal sector, and subtle loss of retinal sensitivity in the central macula. We intend to expand this research to considerably larger longitudinal glaucoma studies in order to test the methods more thoroughly.

References

[AGIS, 1994] AGIS. Advanced glaucoma intervention study. 2. visual field test scoring and reliability. *Ophthalmology*, 101(8):1445–1455, 1994.

[Bilmes, 1997] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Technical Report TR-97-021, ICSI*, 1997.

[Garway-Heath *et al.*, 2000] D. Garway-Heath, F. Fitzke, and R. Hitchings. Mapping the visual field to the optic disc. *Ophthalmology*, (107):1809–1815, 2000.

[Kamal *et al.*, 2000] D. Kamal, D. Garway-Heath, R. Hitchings, and F. Fitzke. Use of sequential heidelberg retina tomograph images to identify changes at the optic disc in ocular hypertensive patients at risk of developing glaucoma. *British Journal of Ophthalmology*, page 993998, 2000.

[Murphy, 2002] K. Murphy. Dynamic bayesian networks: Representation, inference and learning. *PhD Thesis, UC Berkeley, Computer Science Division*, 2002.

[Tucker and Garway-Heath, 2009] A. Tucker and D. Garway-Heath. The pseudo temporal bootstrap for predicting glaucoma from cross-sectional visual field data. *IEEE Transactions on IT in Biomedicine*, 2009.

Determining Useful Sensors for Automatic Recognition of Activities of Daily Living in Health smart home

François Portet[1], Anthony Fleury[2], Michel Vacher[1] and Norbert Noury[2]
[1] Laboratoire d'Informatique de Grenoble, UMR CNRS/UJF 5217, FRANCE
[2] Laboratoire TIMC-IMAG, UMR CNRS/UJF 5525, Faculté de Médecine de Grenoble, FRANCE
{Francois.Portet,Michel.Vacher}@imag.fr, Fleury_Anthony@hotmail.com

Abstract

To face the rapid growth of the world elderly population, *health smart homes* with sensing technology are emerging to automatically detect early loss of autonomy using objective criterion such as the Activity of Daily Living grid. The paper presents data mining techniques to classify seven activities in a health smart home using only the most relevant attributes. The evaluation has shown that a correct classification of 84.5% can be reached with a dataset reduced to 16% related to less than 34% of the current sensors. Results also showed the importance of microphones as complementary data source.

1 Introduction

One of the important goals of health smart homes is to assess how a person copes with her handicap and to detect a loss of autonomy as early as possible. But the technologies involved in health smart home, to be set-up in a large number of flats and institutions, need to be robust, scalable and affordable. Most of the researches related to health smart home is focused on sensors, network and data sharing [Chan *et al.*, 2008], but a fair number of laboratories started to work on reliable activities detection. However, to the best of our knowledge, only few approaches have determined which sensors are important for robust classification. Thus, the domain still needs a robust method to determine the most informative sensors for smart homes. In this paper, we present an evaluation of several data mining state-of-the-art techniques applied to the problem of robust ADL recognition from a minimal set of sensors. One of the originality of the approach is to consider microphones which is a modality not much exploited in this domain.

2 Telemonitoring Data

An experiment has been run to acquire telemonitoring data in the Health smart home of the TIMC-IMAG lab located in the Faculty of Medicine of Grenoble. Thirteen healthy participants were asked to perform 7 activities, at least once, without condition on the time spent. The 7 activities were defined based on the ADL scale: (1) Sleeping; (2) Resting; (3) Dressing and undressing; (4) Feeding; (5) Eliminating; (6) Hygiene activity; and (7) Communicating.

The flat contains 18 sensors which get different information about the inhabitant. Presence infra-red (PIR) sensors bring information about the location and agitation of the person, doors contacts reflect the use of some furniture and a weather station indicates the temperature and hygrometry of the bathroom. The last sensors are a set of seven microphones distributed all around the flat (hidden in the ceiling) analysed in real-time by the AuditHIS system [Vacher *et al.*, to appear] to extract sounds and speech. From these sensors, 38 numerical and boolean attributes have been derived. Data has been annotated by cutting down each ADL interval into 3-minute windows labelled with the name of the activity. The final dataset was thus composed of 232 examples of 3-minute activity described by 38 attributes.

3 Method

In this work, selection of attributes is used to find out what the sensors of interest for ADL automatic recognition are and the impact of this selection on the learning performance. The induction algorithms used were: *Decision Tree (C4.5)*, *Decision Table Majority (DTM)*, *Naïve Bayesian Network (NBayes)* and *Support Vector Machine (SVM)*. They have been chosen for their popularity in data mining applications and because they represent quite different approaches to learning. Although most of the chosen algorithms can handle numerical attributes, the data has been discretized using supervised discretization. Attribute selection techniques are generally divided into two families *filter* and *wrapper*. To test the impact of attribute selection on the learning performance a small subset of each method type has been chosen. *Correlation-based Attribute Selection (CorrFA)* searches for subsets of attributes that are highly correlated with the class but with minimal intercorrelation with each other. This method is thus well suited to discover non redundant attributes sets. *Consistency-based Attribute Selection (ConsFA)* searches for subsets of attributes that are consistent with the class. An attribute subset is inconsistent if there are more than one instance with same attribute values but associated with different classes. *Wrapping* method is an attribute subset selection method that uses the targeted learning algorithm score at each node as evaluator. This method is more time consuming but leads generally to higher performance than the previous described as it fits the learning algorithm.

4 Evaluation

4.1 Attribute Selection Results

For the filter selection method, the number of retained attributes varies with the method employed (CorrFA, ConsFA and Global Filtering) but a global trend appears. Four to five PIR attributes are always in the top of the list (bathroom, living room, bedroom and kitchen) followed by two microphone attributes (sound in bedroom and speech). This is not surprising as each room is related to several ADLs and thus the presence of someone in a room has a high predictive value about what s/he doing in it. With the wrapper methods, the only noticeable change with the filter method is the rank of the dresser state attribute which has a low entropy which is true (i.e., open) only in case of dressing activity, making it quite interesting for classification. Global Filtering (**GF**) and Global Wrapping (**GW**) are composed from the attributes that have been selected more than 50% of the time over stratified 10-fold cross-validation and the selection methods.

4.2 Impact on the Supervised Learning

The impact of the attribute selection has been assessed by learning from the data sets composed from subsets of attributes using a 10-fold stratified cross-validation repeated 10 times. Table 1 summarises the results.

Performance with the whole set is the reference for corrected paired student t-test. This data set gives the lowest performances for DTM (80.0%). NBayes and C4.5 have significantly higher performance than DTM ($p < .05$) but no significant difference is observed between them nor with SVM. Data set composed of PIR attributes (i.e., 'PIR only' data set) only gives significantly reduced performances but still reasonable which emphasizes the location information impact for the classification. When the sound attributes are removed from the whole set (i.e., 'No sound' data set), the performances are significantly lower. This shows that the sound processing does present essential information to perform activity recognition. However, the Signal-To-Noise ratio of the sound signal must be improved to reach satisfying performance in this flat which has poor noise insulation [Vacher *et al.*, to appear].

The data set composed of the attributes selected by Global Filtering (GF) attribute selection method leads to performances that are not significantly different from the ones with the whole set. This data set contains less than 29% of the original data (using only 7 sensors). No learning scheme is significantly better than the others. The data set composed of the attributes selected by Global Wrapping (GW) attribute selection method leads to performances that are significantly better for the DTM learning scheme ($p < .05$) but not for the other schemes when compared with the whole set. No significant difference is observed when compared to the GF performances. This data set contains less than 16% of the original data (using only 6 sensors). No learning scheme is significantly better than the others. Overall the GW method leads to higher average performance (83.9%) than with the whole set (82.9%) and than the GF method (82.9%) but this is not significant and is mainly led by the DTM learning scheme.

method	Whole set	No sound	PIR only	GF	GW
C4.5	83.3	76.8^{**}	71.7^{**}	82.5	83.4
DTM	80.0	75.7^{**}	71.3^{**}	82.6	83.0^{*}
NBayes	85.3	77.4^{**}	72.9^{**}	85.1	84.5
SVM	82.9	78.9^{*}	75.0^{**}	81.3	84.6
average	82.9	77.2	72.7	82.9	83.9

Table 1: Correctly classified Instances (%) for different learning algorithms and data sets ($^{*}p < .05$; $^{**}p < .01$).

5 Conclusion

The main result of the study is that it is possible to keep high performance for automatic classification of ADL when selecting a relevant subset of attributes. About 33% of the sensors (and less than 16% of the attributes) are enough to classify ADL with same (and sometime superior) performance as with the whole data set. But the retained sensors are of different nature (location, sound, contact door) and thus complement each other. The selected attributes were mainly related to PIR sensors and microphones. While these sensors seem to be the most informative, contact door attached to the dresser was essential for classifying dressing activity. Indeed, the chosen ADLs were all related to a location (e.g.: sleeping in the bedroom, eating in the kitchen ...) and when two activities are usually done in the same room (e.g., sleeping and dressing) a strict location sensor is not enough to distinguish them. Thus, realistic ADLs should include activities in unusual location (e.g., eating while watching TV, sleeping on the sofa ...) to challenge the learning process and acquire more accurate models. This is illustrated by the eliminating and hygiene activities which, due to their natural interrelation (e.g., WC and then washing hand) and the flat configuration, challenged the learning.

Globally, sound sensors attributes have a good predictive power and the study shows that this information is essential for ADL classification. But the study also showed the limit of the current audio processing. Indeed, the sound attribute should deliver the same information as the PIR sensors while adding higher semantic level attributes (speech, footstep ...) but the very hostile sound conditions of the experiment shows that the robustness of the current audio processing needs to be improved. However, the presented results confirmed the information power of this modality at least for support to classical smart home sensors.

References

[Chan *et al.*, 2008] Marie Chan, Daniel Estève, Christophe Escriba, and Eric Campo. A review of smart homes- present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1):55–81, Jul 2008.

[Vacher *et al.*, to appear] Michel Vacher, Anthony Fleury, François Portet, Jean-François Sérignat, and Norbert Noury. Complete sound and speech recognition system for health smart homes: Application to the recognition of activities of daily living. In *Recent Advances in Biomedical Engineering*. InTech, Croatia, to appear.

Personalized Feedback based on Automatic Activity Recognition from Mixed-Source Raw Sensor Data

Harm op den Akker
Roessingh Research and Development
h.opdenakker@rrd.nl

Val Jones
Telemedicine Group
University of Twente
v.m.jones@utwente.nl

Hermie Hermens
Roessing Research and Development
h.hermens@rrd.nl

Keywords: Wireless sensor networks, data mining, sensor data fusion, activity monitoring, COPD, biosignals.

Abstract

We present a data set consisting of multiple wireless sensors that monitor movement and various types of bio signals, recorded from patients that suffer from Chronic Obstructive Pulmonary Disease (COPD). From this data, the goal is to derive appropriate feedback to the patient that will motivate them to achieve a healthy lifestyle and a good phyisical condition.

1 Introduction

The AAL[1] project IS-ACTIVE (Inertial Sensing Systems for Advanced Chronic Condition Monitoring and Risk Prevention) started in April 2009. The project addresses continuous monitoring of activities and health status of patients, suffering from Chronic Obstructive Pulmonary Disease (COPD), in their daily environment. The goal is to promote a healthy lifestyle by providing personalized feedback on daily life activities taking into account the limitations for the patient caused by his chronic condition. To achieve this goal, we need to know what the patient is doing, and what the condition of the patient is throughout the day.

The patient will be equipped with a series of smart wirelessly networked sensor nodes. The final selection of sensors to be used has not yet been made but will likely include MEMS accelerometers, tilt switches, gyroscopes and magnetic compasses. Each sensor node will also include a microcontroller which takes care of sampling and networking, but resources must also be reserved for intelligent processing of the sampled data on the microcontroller itself. Besides the motion sensors needed for activity recognition, the patient will also wear biosensors to monitor his health status. Physiological parameters of interest include heart rate, some measure of respiration and oxygen saturation. In addition, analysis of audio recordings may be used to detect respiratory difficulty indicated by coughing and heavy breathing.

The resulting dataset will include sensor data outputs captured while performing a wide range of movements like walking, nordic walking (with sensors on the sticks), cycling, and any physiotherapy exercises that are commonly prescribed to COPD patients. Continuous series of sensor data will be hand-annotated with descriptions of the activities performed. If possible, the corpus will include video recordings of some basic activities, such as walking on a treadmill while wearing all the different sensors, so that these activities can be studied in greater detail afterwards. Because initially all raw sensor outputs are saved to the corpus, and the sampling frequency of the sensors will be set to a high level (100Hz or higher), the total quantity of data collected will be very large. The use of multiple movement sensors, such as the Inertia ProMove sensor[2], which will feature 9 degrees of freedom from three sensors: 3D accelerometer, 3D magnetic compass and 3D gyroscope, will result in a data set comprising as many as 40-50 layers of data.

The data mining challenge will be to automatically classify time segments of sensor data as belonging to one of the identified activities, and at the same time to calculate an estimate of the amount of strain that is put on the patient while performing that activity. The definition of strain in this context depends heavily on the individual patients, and can be seen as physical- or psychological strain, stress or a combination thereof. An important constraint on the algorithms to be designed is that they should run on the wireless sensor network nodes in a distributed way. Data transmission should be kept to a minimum to preserve battery lifetime, while processing power on the nodes themselves is limited. For examples of a distributed activity recognition approache see [Marin-Perianu *et al.*, 2008; Amft *et al.*, 2007].

The challenge in designing the automatic activity recognizers can be described by three requirements. First, the algorithms should use as little data as possible from the sensors in order to minimise the number of sensors actually needed and to enable reduction of each individual sensor's sampling frequency. Second, processing of individual sensor outputs should be done on the wireless sensor node itself, as far as possible, in order to reduce the need

[1] Ambient Assisted Living

[2] http://www.inertia-technology.com/

for wireless transmission between sensor nodes. And third, the part of the algorithm which combines the various sensor outputs should be as simple as possible so as to be able to run on one of the (resource poor) nodes.

2 Approach

Because of the distributed nature of the task, we propose a layered approach with well defined subtasks that can be performed on specific nodes in the network. At the lowest layer, feature extraction from the wireless sensor data will take place. Once it is clear which features are needed for the activity recognition task, these features should be extracted on the sensor nodes themselves so that network transmission can be kept to a minimum. Then, the feedback device, which will most probably be some sort of PDA, is charged with the task of collecting relevant features from the nodes and doing the actual activity recognition. Note that with state-of-the-art PDA devices, the resources available for this part of the algorithm might not be that limited at all, but battery usage remains an issue. A similar approach is required for the biosensors, which will send their data (e.g. heartrate) to the feedback device on a previously defined minimum need basis. A second algorithm running on the feedback device will then combine biosensor and activity data and generate appropriate feedback for the patient.

This feedback is meant to help patients to be as active as possible, while preventing attacks of breathlessness. In order to provide each patient with the optimal feedback, the system will adapt to the behaviour and health status of the user. If, for example, a patient repeatedly chooses to ignore advice from the system to take it easy, with no serious health consequences for the patient, the system should become less cautious and allow the user to be more active. More importantly, if the system fails to warn a patient in time to lower his/her activity level, the system should issue its warnings more quickly. This general activity monitor is one of the envisioned applications, one that requires only a rather broad measure of activities. A second application is to aid patients in performing their daily physiotherapy excercises in a correct way. This requires a better accuracy from the activity recognition algorithm, because it has to correctly detect, for example, short series of arm or leg movements. The feedback device can then take on the task of personal coach, by keeping track of the exercise schedule while giving motivational feedback.

These different applications impose different requirements on the classification tasks. On the broad scope, the system should never mistake running for lying in bed, but mistaking slowly riding a bike for walking might not be a huge problem. On the other hand, the need for accuracy greatly increases when trying to detect all the actions that are performed in a physiotherapy excercise session. These differences have to be taken into account when collecting the training data for the algorithms. For detecting a bicycle ride, it may be sufficient to annotate a 15 minute trip from home to work as "riding a bike" (without indicating a stop for a red light, or the speed at every moment) and use it for training. For the excercise patterns, it is probably a good idea to make video recordings of various sessions, and letting each phase of the movements be annotated by multiple annotators according to a previously agreed upon annotation manual. The inter annotator agreement then needs to be high overall, but small inconsistencies near the boundaries of movements may be acceptable.

For the annotation schema we propose a layered approach. On the highest layer, at least five different classes will be distinguished including riding a bike, walking, jogging, doing excercises and *non-active*. Then in the second layer, more detailed activities like the exact excersises can be annotated. These might include up to 10 different at home excersises for COPD patients. If necessary for the classification algorithms, a third layer may contain annotations of specific arm- or leg movements.

3 Feedback

As stated earlier, the final goal of the research is to promote a healthy lifestyle for COPD patients. We attempt to achieve this by providing feedback that motivates each individual patient to improve their physical condition to the maximum of their abilities. This raises the question of when and how to provide feedback, which is a non-trivial and not well understood issue. That is why an important part of our research will focus on using the recognized activity patterns and bio-signal data as input to a feedback system. This system can be seen as a sort of Clinical Decision Support System that will also have to adjust its 'decisions' (i.e. feedback responses) to how the patient reacts to them. At this point however, the details of the development of such a system are largely unclear.

To conclude, the goal of this article is to start a discussion on how to use data mining or machine learning techniques to eventually derive appropriate patient feedback from a large set of raw sensor data.

References

[Amft *et al.*, 2007] Oliver Amft, Clemens Lombriser, Thomas Stiefmeier, and Gerhard Tröster. Recognition of user activity sequences using distributed event detection. In *EuroSSC 2007: Proceedings of the 2nd European Conference on Smart Sensing and Context*, volume 4793 of *Lecture Notes in Computer Science*, pages 126–141. Springer, October 2007.

[Marin-Perianu *et al.*, 2008] Mihai Marin-Perianu, Clemens Lombriser, Oliver Amft, Paul Havinga, and Gerhard Tröster. Distributed activity recognition with fuzzy-enabled wireless sensor networks. In *Proceedings of the International Conference on Distributed Computing in Sensor Systems*, pages 296–313, 2008.

Analyzing Episodes of Care in Hospital and Outpatient Settings

Kirk T. Phillips
Iowa Health System
Des Moines, Iowa USA
phillikt@ihs.org

Abstract

Traditional studies of health care quality center upon single events of hospital inpatient treatment, measuring treatment processes and outcomes for patients grouped by disease or unit of service. This paper seeks to refine models for analyzing episodes of care with temporal data mining methods. We aim to characterize patterns of patient condition, treatment processes, and outcomes, based upon administrative data from outpatient and inpatient settings to be considered in the design and monitoring of quality improvement interventions.

1 Introduction

We are collaborating with the Institute for Healthcare Improvement, who recently commented that "Reducing rates of re-hospitalization has attracted attention from policymakers as a way to improve quality of care and reduce costs. However, we have limited information on the frequency and patterns of re-hospitalization in the United States to aid in planning the necessary changes." (Jencks 2009). Re-hospitalization is one among many patterns of treatment for a given disease, as shown in Figure 1. In this model, cases are identified as "index events" for hospitalization, and their records from other sites of care eg emergency room, are added to the data set Quality improvement questions to be addressed are – What patient conditions lead to re-hospitalization and other outcomes? How can we optimize the patients' quality and cost of care?

2.1 Methods for Episodes of Care

In current practice, hospitalized patients are tabulated by their source of admission and discharge for a specified condition ie. index event. These tables provide a limited view of their care, without an understanding of comorbidities and other sites of care during an episode of illness.

A better approach is to select patient records that have been hospitalized for the condition (index event), and in a second pass through the data set, select all records for those patients representing care before and after their hospitalization. Simple sorts of the data set illustrate treatments for individual patients, useful for clinical decision support. Temporal data mining methods reveal patterns of disease and treatment, to be used in decisions about quality improvement and financing of health care.

2.2 Discussion

Concaro et al. applied sequential pattern mining algorithms with a national inpatient sample to find frequent sequences of disease such as pneumonia, hypertension, and diabetes, with a market basket form of analysis. They recommend further research to "specify the temporal evolution and potential cause-effect relationships between healthcare events. "

Tudor et al. applied logistic regression models, accounting for daily organ failure scores in the prediction of survival for intensive care patients. Prediction of hospital treatment outcomes addresses a part of our episode of care scenario, aimed to improve the quality and cost of care in all settings

This paper is submitted to IDAMAP with an initial description of a problem and data set that may benefit from an analysis through intelligent data analysis or data mining. We look forward to discussing alternative methods for this application.

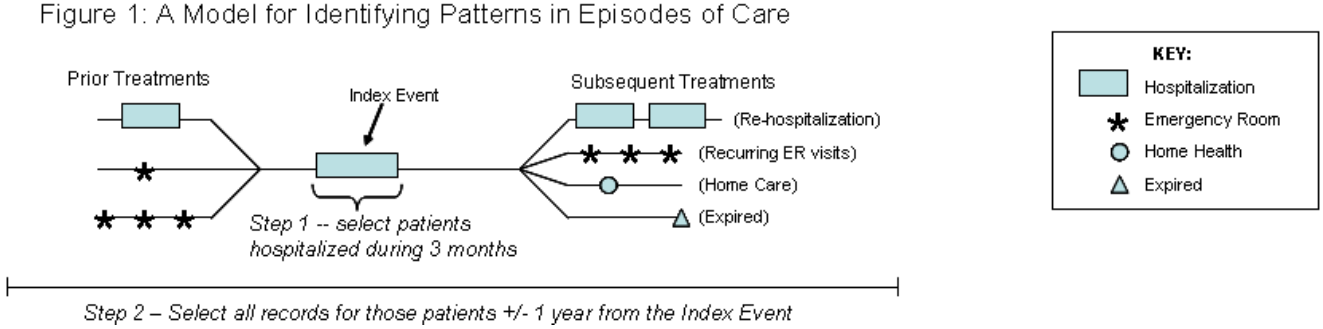
References

[Concaro et al, 2008] Stefano Concaro, Lucia Sacchi, Carlo Cerra, Pietro Fratino, Riccardo Bellazzi. *Temporal Data Mining for the Analysis of Administrative Healthcare Data* IDAMAP 2008 Intelligent Data Analysis in Biomedicine and Pharmacology. Artificial Intelligence in Medicine 2008, Washington DC.

[Jencks et al., 2001] Stephen F. Jencks, Mark V. Williams and Eric A. Coleman. *Rehospitalizations among Patients in the Medicare Fee-for-Service Program*
N Engl J Med 360;14 nejm.org, April 2, 2009.

[Toma et al., 2007] Toma, Tudor, Ameen Abu-Hanna, Robert-Jan Bosman. *Discovery and inclusion of SOFA score episodes in mortality prediction* JBI 40:6; December 2007.

Figure 1: A Model for Identifying Patterns in Episodes of Care



IDAMAP 2009 Student Challenge

Description and Aim

The purpose of this challenge is to encourage students to study and apply best knowledge discovery approaches to a real medical data mining problem. The goal is to extract knowledge from the National Ambulatory Medical Care Survey (NAMCS) data set (see description of NAMCS data).

The contributions will be judged by the IDAMAP program committee members. The winner will receive a cash award of 300 EUR and will have the opportunity to present his/her work at the IDAMAP workshop. No funds for travel or workshop registration will be provided.

Eligibility

The challenge is open to all students. Individual participants as well as groups are welcomed.

Participation

Participants are invited to report on interesting findings and knowledge obtained applying data mining approaches to the NAMCS data set. There are no restrictions on the data mining technique to be used (supervised or unsupervised learning are allowed).

The data is publicly available at: www.cdc.gov [look for NAMCS]. The data are organized by year, but there are no restrictions or requirements to use data from any specific year. The participants are free to use data from one or multiple years.

The contributions must be submitted as a written report and must include:

- names, academic degree(s), affiliations of all authors (advisors should be listed as authors),
- short abstract,
- introduction, methods, results, discussion and conclusion.

The entry must be formatted using the Word template for regular IDAMAP submissions (up to six pages, 4500 words, see instructions for submission in the IDAMAP call for papers). Entries must be submitted by email to the workshop chairs.

Winners

Congratulations to Lan Umek and Minca Mramor from the University of Ljubljana, Slovenia, and their supervisor Blaz Zupan.

They have applied a subgroup discovery approach to analyze the records on outpatient visits from the year 2006. Their method discovered a number of significant subgroups of similar patients in terms of characteristics, reasons for visiting the physician, diagnosis and treatment. The authors comment on the somewhat unexpected properties of three such subgroups and suggest that each may deserve further biomedical research.

Subgroup Discovery in Data Sets with Multi–Dimensional Responses: Application to a Medical Domain

The IDAMAP 2009 Student challenge

Lan Umek, Minca Mramor
supervisor: **Blaž Zupan**
Faculty of Computer and Information Science
University of Ljubljana, Slovenia
{lan.umek, minca.mramor}@fri.uni-lj.si

Abstract

Medical data sets often include a large number of features. When inferring a machine learning model, the features can be divided into the input (description of patients, symptoms, other diseases, ...) and output (diagnosis, treatment, drug prescription, ...) variables. To analyse such data sets we used the technique that can treat many output features simultaneously [Umek *et al.*, 2009]. The method uses a combination of k-medoids clustering and analysis of contingency tables and aims to find subgroups of patients that are similar both in the input and output space with significant dependencies between both sets of variables. We applied this method on the *The National Ambulatory Medical Care Survey* data and discovered several meaningful patterns.

1 Introduction

The aim of our analysis for the student challenge was the application of a subgroup discovery approach [Umek *et al.*, 2009] on the records from the outpatient departments in the year of 2006 from *The National Ambulatory Medical Care Survey (NAMCS)* data [NAMCS, 2009]. Due to the high diversity of medical data we applied the method on several different subsets of the patients. Our goal was to group patients into subgroups that would reflect reasonable relations between their individual characteristics, vital signs, and reasons for visiting the physician and the corresponding diagnosis and treatment procedure.

2 Data

We used the data from the National Ambulatory Medical Care Survey (NAMCS) conducted in the year of 2006. The aim of NAMCS is to gather reliable information about the provision and use of ambulatory medical care services in the United States. The survey first started in the year of 1973. The information about the patient visits is collected from the non-federally employed office-based physicians. Data are obtained on the demographic characteristics of patients, patients symptoms, physicians diagnoses, diagnostic procedures, medications ordered or provided, patient management, and planned future treatment. The data is freely available online at `ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHAMCS/`.

The 2006 data included the information on 35104 patient visits described with 416 variables. In our analysis we used 55 variables that we divided into the following 6 sections: patient characteristics (*e.g.* age, sex, race) reason for visit (symptoms), associated diseases (*e.g.* diabetes, obesity, cancer), vital signs (*e.g.* blood pressure, temperature), diagnosis, medications, and diagnostic screening services. These variables were used to find the subgroups in the data, as shown in Table 1.

3 Models

3.1 Search Algorithm

Our analysis was based on a subgroup discovery approach [Umek *et al.*, 2009] which requires a data set with two sets of variables at the input and output. The method's main idea is to reduce the complexity of the possible links between the two sets of variables (X and Y) by summarizing the variability among X and Y. This is done by inferring cluster systems L_x and L_y and afterwards studying the relation between the input and output variables by analysing the cross–table $L_x \times L_y$. For this purpose we extracted a subset of variables from the original data and divided it to inputs and outputs (Table 1).

Table 1: Feature sets, the number of features included (n_{var}) and a list of names for a selection of representative features from the set.

feature set	n_{var}	representative features
INPUT	32	
basic	3	sex, age, race
vital signs	7	height, BMI index, blood pressure
other diagnosis	21	arthritis, asthma, cancer, ...
+ use of tobacco		
reason for visit	1	main reason for visit
OUTPUT	23	
diagnosis	1	principal diagnosis
diagnostic tests	22	total number of tests
+ number of medications		EKG, urine test, ...

The used method discovers subgroups of patients that are similar both in inputs and outputs and reflect the local

dependencies between both sets of variables. The search algorithm can be briefly described in the following steps:

1. Perform k_1–medoids clustering [Kaufman and Rousseeuw, 1990] of n patients in X–space and k_2–medoids clustering in Y–space. Label each instance with the corresponding cluster indices L_x and L_y.
2. Present the results of clustering in a contingency table for L_x and L_y:

		L_y				
		1_y	2_y	$\cdots$	k_2	
	1_x	n_{11}	n_{12}	$\cdots$	n_{1k_2}	n_{1+}
	2_x	n_{21}	n_{22}	$\cdots$	n_{2k_2}	n_{2+}
L_x	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	k_1	$n_{k_1 1}$	n_{k1}	$\cdots$	$n_{k_1 k_2}$	n_{k_1+}
		n_{+1}	n_{+2}	$\cdots$	n_{+k_2}	n

   where n_{ij} represents the number of patients in cluster i in the input space and cluster j in the output space.
3. Under null hypothesis

   $$H_0 : \textit{variables } L_x \textit{ and } L_y \textit{ are independent}$$

   the contribution

   $$c_{ij} = \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}}. \quad (1)$$

   of the subgroup (cell) ij approximately follows χ^2 distribution with one degree of freedom [Agresti, 2001]. The estimated p–value p_{ij} is then assigned to each subgroup. Subgroups with small p–values show significant evidence against the null hypothesis and are reported to the expert.

In the search algorithm we tested several different combinations of the numbers of clusters k_1 and k_2. This consequently produces a very large number of overlapping subgroups and increases the danger of false discoveries. The set of discovered subgroups was therefore reduced in the following steps:

- the p–values were adjusted using *false discovery rate* [Benjamini and Hochberg, 1995],
- we measured the similarity between subgroup and reported only the most representative ones

For expert's interpretation of the results we described each subgroup of patients with the most frequent value of *reason for visit* and most frequent *principal diagnosis*. The distribution of other variables has been reported only in case of significantly different distribution compared to the reference set.

3.2 Parameters Used in the Algorithm

The crucial part of every clustering algorithm is the definition of the dissimilarity measure. We used weighted Manhattan distance where each feature set from Table 1 had the same weight. Within each feature set all the variables were equally important. With such definition we stressed the importance of variables from smaller feature sets, especially the *reason for visit* in X–space and *principal diagnosis* in Y–space.

Based on the variable *reason for visit*, we divided the data into blocks and were interested in specific subgroups within these predefined blocks. The reasons for visit are classified into 7 modules, from which we used the Symptom module, Disease module and Injuries and Adverse effects module as defined in the appendix II of the NAMCS Micro-data file documentation. The modules are additionally divided into groups based on the organ system the symptoms or diseases refer to (*e.g.* Symptoms Referable to the Respiratory System or Diseases of the Nervous System, see Table 2).

4 Results and Discussion

For most of the patients from different groups in the modules described in the previous paragraph, we have discovered several interesting and significant subgroups of patients. Table 2 shows the number of subgroups that our method identified as significant.

Table 2: Overview of the results. The block of symptoms defined the data set for analysis, # patients is its size. The number # subgroups presents the number of interesting subgroups after filtering.

block	# patient	# subgroups
symptoms referable to		
skin, nails, and hair	919	9
nervous system (excluding sense organs)	1033	10
eyes and ears	1090	9
digestive system	1252	3
cardiovascular and lymphatic systems	82	2
poisoning and adverse effects	88	2
neoplasms	524	14
mental disorders	219	9
diseases of		
skin and subcutaneous tissue	287	5
respiratory system	267	7
musculoskeletal system and connective tissue	208	2
genitourinary system	401	6
eye	253	8
digestive system	211	4
circulatory system	533	3
blood and blood-forming organs	115	2
congenital anomalies	150	2
injury by type and-or location	1385	8

4.1 Selected Results

We here explain three of the discovered subgroups, showing the methods output, our description and the experts explanation of the subgroup. The adjusted p–values of all subgroups are less than 0.01.

1. A subgroup of 11 out of 115 patients from block "Diseases of the Blood and Blood-forming Organs" ($p < 0.001$)

Methods output:
Main symptom: Anemia (probability=1)

Other interesting patient's properties:
Distribution of BMI
On the entire data set:
mean = 23.9, standard deviation = 6.9
On the subgroup:
mean = 17.4, standard deviation = 2.1

Distribution of RACE
On the entire data set:
P(RACE=Asian Only)=0.06
P(RACE=Black/African American Only)=0.32
P(RACE=More than one race reported)=0.02
P(RACE=Native Hawaiian/Oth Pac Isl Only)=0.01
P(RACE=White Only)=0.59
On the subgroup:
P(RACE=Black/African American Only)=1

Diagnosis
Sickle-cell anemia, unspecified (probability = 0.909)
Hb-S disease with mention of crisis (probability = 0.09)

Description of the subgroup:
All of the patient from this subgroup had anaemia. From the input variables, this subgroup significantly differs from the entire group of patients with "Diseases of the Blood and Blood-forming Organs" in the body mass index (BMI) and the race variables. Specifically, the patients from the subgroup have lower BMI (average 17.1) and are all Black American/African by race. From the output variables only the Diagnosis is reported, which is for all 11 patients Sickle-cell anaemia (one of them having a Sickle-cell anaemia crisis).

Expert's explanation:
Sickle-cell disease, or sickle-cell anaemia, is a lifelong blood disorder characterized by red blood cells that assume an abnormal, rigid, sickle shape. Sickling decreases the cells' flexibility and results in a risk of various complications. The sickling occurs because of a mutation in the hemoglobin gene. Sickle cell disease is the most common inherited blood disorder in the United States. The prevalence of the disease in the United States is approximately 1 in 5,000, mostly affecting African Americans, according to the National Institutes of Health [NIH, 2009].

The discovered subgroup points out this known characteristic of patients with sickle cell anaemia (mostly African Americans). However, the patients in the subgroup also have a significantly lower BMI than other patients with diseases of the blood and blood-forming organs.

2. Subgroup of 143 out of 533 patients from block "Diseases of the Circulatory System" ($p = 0.0156$)

   Methods output:
   Main symptom: Hypertension (probability = 0.889)

   Other interesting patient's properties:
   Distribution of CASTAGE
   On the entire data set:
   P(CASTAGE=In situ)=0.83
   P(CASTAGE=Local)=0.17
   On the subgroup:
   P(CASTAGE=Local)=1

   Diagnosis = Unspecified essential hypertension (probability = 0.74)

   Description of the subgroup:
   This is a subgroup of more than 20% of patients with the reason for visit "Diseases of the Circulatory System". 89% of these patients had the main symptom of hypertension. All of these patients were also diagnosed with cancer at the local stage. The diagnosis of most of the patients from the subgroup was unspecified essential hypertension.

   Expert's explanation:
   This is an interesting subgroup because all the patients with hypertension also have cancer at the local stage. Hypertension is one of the known cardiovascular side effects of cancer treatment[Yeh and Bickford, 2009]. However, hypertension in this large group of patients is most likely not only due to chemotherapy. The association between hypertension and cancer in this subgroup is intriguing and would be appealing for further research.

3. A subgroup of 19 out of 287 patients from block "Diseases of the Skin and Subcutaneous Tissue" ($p < 0.001$)

   Methods output:
   Main symptom: Psoriasis (probability = 1)

   Other interesting patient's properties:
   Distribution of BPDIAS
   On the entire data set:
   mean = 72.3, standard deviation = 13.3
   On the subgroup:
   mean = 85.5, standard deviation = 7.5

   Distribution of BMI
   On the entire data set:
   mean = 29.0, standard deviation = 10.5
   On the subgroup:
   mean = 46.1, standard deviation = 15.0

   Diagnosis = Other psoriasis (probability = 1)

   Description of the subgroup:
   This is a smaller subgroup of 19 patients with diseases of skin and subcutaneous tissue. All of them had the symptoms and diagnosis of psoriasis. The significant characteristics of this group are also high diastolic blood pressure (average value of 85) and a very high body mass index (BMI, average value 46.1, normal values from 20 to 25).

Expert's explanation:
Obesity is a significant risk factor for psoriasis and body mass index (BMI) correlates with disease severity. The reason might be in the higher levels of adipokines such as leptin and resistin in overweight individuals that correlate with higher levels of inflammatory cytokines and lead to a more severe course of disease[Johnston *et al.*, 2008].

5 Conclusions

With the chosen subgroup discovery approach several interesting subgroups of patients were discover from the *The National Ambulatory Medical Care Survey (NAMCS)* data. However, it is important to stress, that the definition of the metric has a large impact on the results. In our first attempt all the variables were treated equally and consequently the resulting subgroups did not reflect reasonable relations. It is senseless to expect that patient characteristics, for example sex or race, have the same impact on the physician's diagnosis as the major reason for visit. On the other side, we have to consider all the relevant information. Therefore, the choice of suitable weights in the definition of the metric plays a crucial part in the analysis of medical data.

References

[Agresti, 2001] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2 edition, 2001.

[Benjamini and Hochberg, 1995] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[Johnston *et al.*, 2008] A. Johnston, S. Arnadottir, JE. Gudjonsson, A. Aphale, AA. Sigmarsdottir, SI. Gunnarsson, JT. Steinsson, JT. Elder, and H. Valdimarsson. Obesity in psoriasis: leptin and resistin as mediators of cutaneous inflammation. *British Journal of Dermatology*, 159(2):342 – 350, 2008.

[Kaufman and Rousseeuw, 1990] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

[NAMCS, 2009] NAMCS. National center for health statistics. `http://www.cdc.gov/nchs/about/major/ahcd/namcsdes.htm`, 2009.

[NIH, 2009] National institute of health, sickle cell anemia. `http://www.nhlbi.nih.gov/health/dci/Diseases/Sca/SCA_Summary.html`, 2009.

[Umek *et al.*, 2009] Lan Umek, Blaž Zupan, Marko Toplak, Annie Morin, Jean-Hugues Chauchat, Gregor Makovec, and Dragica Smrke. Subgroup Discovery in Data Sets with Multi–Dimensional Responses: A Method and a Case Study in Traumatology. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine, AIME09*, Lecture Notes in Computer Science, pages 265–274. Springer Verlag, 2009.

[Yeh and Bickford, 2009] Edward T.H. Yeh and Courtney L. Bickford. Cardiovascular complications of cancer therapy: Incidence, pathogenesis, diagnosis, and management. *Journal of the American College of Cardiology*, 53(24):2231 – 2247, 2009.