Conquering the Curse of Dimensionality in Gene Expression Cancer Diagnosis: Tough Problem, Simple Models

Minca Mramor[1], Gregor Leban[1], Janez Demšar[1], and Blaž Zupan[1,2]

[1] Faculty of Computer and Information Science
University of Ljubljana, Tržaška 25, Ljubljana, Slovenia
[2] Department of Molecular and Human Genetics
Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, U.S.A.

Abstract. In the paper we study the properties of cancer gene expression data sets from the perspective of classification and tumor diagnosis. Our findings and case studies are based on several recently published data sets. We find that these data sets typically include a subset of about 100 highly discriminating features of which predictive power can be further enhanced by exploring their interactions. This finding speaks against often used univariate feature selection methods, and may explain the superior performance of support vector machines recently reported in the related work. We argue that a much simpler technique that directly finds visualizations with clear separation of diagnostic classes may be used instead. Furthermore, it may perform better in inference of an understandable classifier that includes only a few relevant features.

1 Introduction

Carcinogenesis is a multi step process in which genetic alterations drive the progressive transformation of normal human cells into malignant derivates. Gene expression microarrays can be used to identify specific genes that are differentially expressed across different tumor types. Classification of clinically heterogeneous cancers using gene expression profiles is an emerging application of microarray technology. Several recent studies of different cancer types, including acute leukemias [1], lymphomas [2], and brain tumors [3] have already demonstrated the utility of gene expression profiles for cancer classification, and reported on the superior classification performance when compared to standard morphological criteria. More accurate classifications based on molecular phenotype can improve and individualize treatment, influence the development of targeted therapeutics, and help in the identification of biomarkers for diagnosis and prognosis.

From the viewpoint of inference of classification models, gene expression data sets are rather peculiar. They most often include thousands of attributes (genes) and a small number of examples (patients). Another problem is a substantial component of noise, resulting from numerous sources of variation affecting expression measurements.

In conquering the curse of dimensionality, the prevailing modelling approaches include gene filtering in the data preprocessing phase, and gene subset selection often coupled with a modelling technique. For instance, in the work reported by Khan *et al.* [4] to classify four types of tumors in childhood, authors first ignored genes with low expression values throughout the data set, then trained 3750 feed-forward neural networks on different subsets of genes as determined by principal component analysis, analyzed the resulting networks for most informative genes and in this way obtained a subset of 96 genes of which expression clearly separated different cancer types when using multi-dimensional scaling. Other approaches, often similar in complexity of data analysis procedure, include k-nearest neighbors with weighted voting of informative genes [1] and support vector machines (SVM, [5–7]). In most cases, the resulting prediction models include complex computation over a set of gene expressions (*e.g.*, neural networks, SVM, or principal components models) which are hard to interpret and can not be communicated to the domain experts in a way that would easily reveal the role genes play in separating different cancer types.

While different approaches have been used for selecting marker genes [5, 1, 4, 8] (for a review see [9]), the prevalent approach is based on univariate studies which examine the value of a gene in absence of context, that is, disregarding the expressions of other observed genes. Such an approach, most often used in practice, is for instance signal-to-noise statistics [1]. Since genes interact, one would expect that more information can be gained by observing a set of genes as a group, rather than summing their individual effects. This is confirmed experimentally through a success of non-linear modelling methods that can account for gene interactions [5]. In this respect, the use of univariate scoring for gene subset selection is questionable.

In the paper, we provide a simple alternative to rather complex gene expression diagnostic modelling approaches mentioned above. Namely, we show that a subset of two to five genes can most often provide sufficient information to clearly separate the diagnostic classes when their expression data is visualized either in scatterplot (two genes) or radviz (three genes or more) planar geometric graphs. The paper first introduces gene scoring, visualization, and visual projection search methods we apply in our studies. Our experimental study, together with the data sets used, is described next. The paper finishes with the discussion of results and concluding remarks.

2 Methods

2.1 Gene Ranking

In the paper, we use two in essence very different methods for gene ranking. Signal-to-noise (S2N) [1] is a univariate statistics for scoring of attributes that is derived from the standard parametric t-test statistic and is computed as $\frac{\mu_0-\mu_1}{\sigma_0+\sigma_1}$, where μ and σ represent the mean and standard deviation of gene's expression, respectively for each class. To score genes in multi-class problems, we have taken the data for each pair of class values, computed the statistics, and then averaged it across all possible class value pairs.

ReliefF [10, 11] is a feature scoring function that is, in principle, sensitive to feature interactions by being able to detect features that may not provide much information on their own, but could be very useful when used together with some other features. ReliefF scores features according to how well their values distinguish among instances that are similar to each other. Since the similarity is computed based on all features in the data set, they define the context for the feature's score thus providing grounds for revealing the interactions.

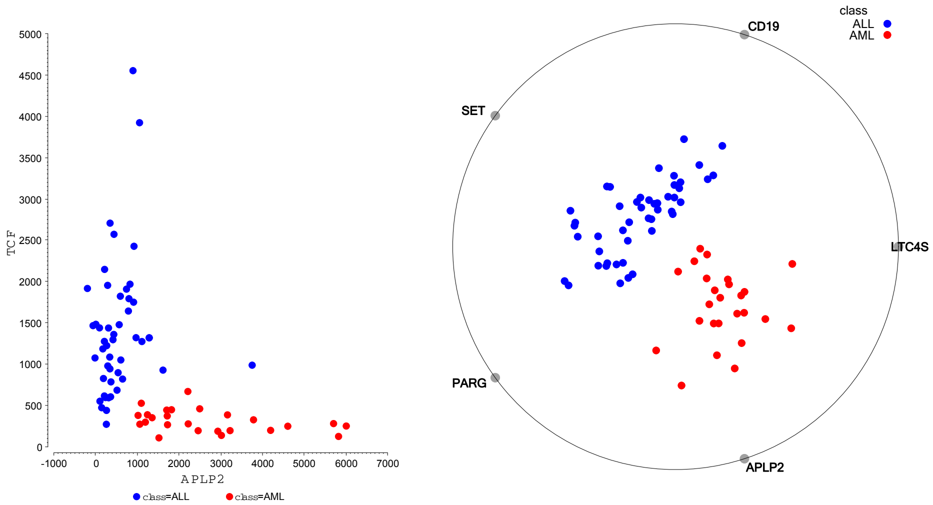
2.2 Two-Dimensional Geometric Visualization Methods

In the paper we propose to visualize gene expression data by plotting the examples from the data sets in a two-dimensional graphs. Depending on how many genes we use in the plot, we draw either a scatterplot (two genes) or a radviz (three or more genes). By selecting a suitable set of genes for the plots (see next section), we aspire that either of the two visualization methods would provide for a clear visual separation of diagnostic classes.

For a scatterplot, an example of such a graph is given in Figure 1.a. Figure 1.b shows a radviz with five genes, represented as anchors that are equally spaced around the perimeter of a unit circle. The examples are visualized as points inside the unit circle, where their position depends on gene expression value: the higher the value for a gene, the more the anchor attracts the corresponding point. Finding an attraction equilibria for a visualized set of genes determines the placement of each of examples in the data set [12]. Examples with approximately equal expressions of genes that lie on the opposite sides of the circle will lie close to the center of the circle. On the other hand, if the expression of a single gene in a visualized set prevails, the point will lie close to the corresponding anchor. The radviz projection is defined with the gene subset being visualized, and with the placement of gene anchors. While enabling the visualization of several genes in a single graph, radviz has some deficiencies. For instance, placing two highly correlated genes that are good at discriminating between classes on the opposite side of the circle will make them useless in the visualization, since there joint effect will be cancelled out. On the other hand, they might generate a projection with well separated classes if their anchors are placed adjacently. The "correct" placement of feature anchors was for instance crucial for a nice separation of classes in Figure 1.b, where the two anchors (genes) on the top of the circle (SET and CD19) attract data points from the ALL class and genes APLP2 and LTC4S attract points with AML class value.

2.3 Visualization Scoring and Ranking and Projection Search

In Figure 1 we have shown two data visualizations that utilize only a small number of attributes (genes) but provide a good separation of diagnostic classes. Cancer microarray data includes thousands of genes, and it is therefore not trivial to find a useful projection, that is, a subset of genes to visualize. In fact, we have



(a) scatterplot (b) radviz

Fig. 1. Two projections for the leukemia data set (see Section 3.3).

observed that, for a typical cancer data set, there are only a few among millions of possible projections that exhibit a clear separation of diagnostic classes, making the manual search for a good visualization impossible.

To automate the search for a good projection, we first need to define how to evaluate them. The algorithm we use, VizRank [13], scores a particular projection by training a k-nearest neighbor classifier on two attributes – the x and y positions of points in the projection. The classification accuracy of the classifier is then assessed using 10-fold cross validation and provides for a projection score. When classes in the projection are well separated, the classification accuracy of the k-NN classifier will be high and the projection will be highly ranked. In projections where some points from different classes overlap, the accuracy of the classifier and with it the value of the projection will be accordingly lower.

Although VizRank's visualization scoring can be very efficient (more than 2000 projections can be evaluated per minute on a 2.4GHz computer) evaluating all projections in the data sets with several thousands of attributes is not feasible. Instead, VizRank uses an efficient heuristic that first scores the attributes using ReliefF [11], ranks the subsets of attributes to be considered by the sum of ReliefF scores, and evaluates the projections starting with the most likely candidates using this heuristics. Our experiments show that by using this heuristic only a few percents of possible projections have to be assessed in order to find the most interesting ones.

Table 1. Cancer-related gene expression data sets used in our study. Columns report on number of examples, diagnostic classes and genes included in a data set, and proportion of examples in the majority diagnostic class. Last two columns show the average probability of correct classification ($\overline{P}$) for the top-ranked scatterplot and radviz projection.

	Number of			Major	$\overline{P}$ for top projection	
Data set	Samples	Classes	Genes	class	Scatterplot	Radviz
Leukemia	73	2	7074	52.8%	98.04%	99.55%
MLL	72	3	12533	38.9%	94.82%	99.75%
SRBCT	83	4	2308	34.9%	87.69%	99.74%
Prostate	102	2	12533	51.0%	91.76%	98.27%
DLBCL	77	2	7070	75.3%	96.82%	99.71%

3 Experimental Study

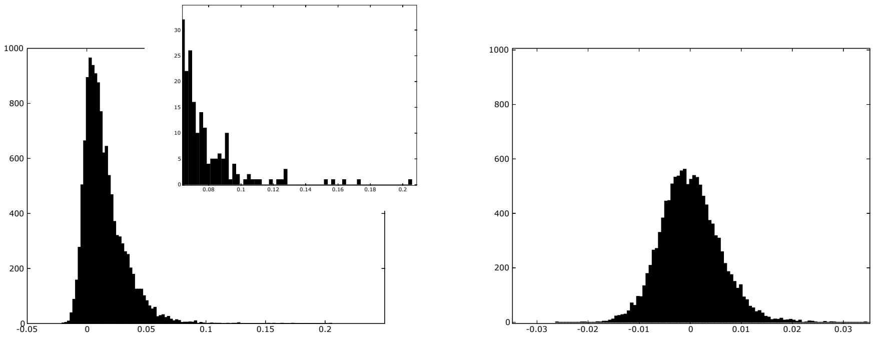
3.1 Data Sets

For experimental analysis reported in this paper we use five publicly available data sets with information on gene expression profiles in different human cancer types (Table 1). Three data sets, leukemia [1], diffuse large B-cell lymphoma (DLBCL) [2] and prostate tumor [14] have two categories. The leukemia data includes 48 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples, each with 7074 gene expression values. The DLBCL data set includes 7070 gene expression profiles for 77 patients, 58 with DLBCL and 19 with follicular lymphoma (FL). The prostate tumor data set includes 12533 genes measured for 52 prostate tumor and 50 normal tissue samples. The data for these three data sets and the mixed lineage leukemia (MLL) data set were produced from Affymetrix gene chips and are available at http://www-genome.wi.mit.edu/cancer/.

We additionally analyzed two multi category data sets. The MLL [15] data set includes 12533 gene expression values for 72 samples obtained from the peripheral blood or bone marrow of affected individuals. The ALL samples with a chromosomal translocation involving the mixed lineage gene were diagnosed as MLL, so three different leukemia classes were obtained (AML, ALL and MLL). The SRBCT data set [4] consists of four types of tumors in childhood, including Ewing's sarcoma (EWS), rhabdomyosarcoma (RB), neuroblastoma (NB) and Burkitt's lymphoma (BL). It includes 2308 genes and 83 samples derived from tumor biopsy and cell lines. The data for the SRBCT data set were obtained from cDNA microarrays. The data set is available at http://research.nhgri.nih.gov/microarray/Supplement/.

3.2 Gene Ranking by ReliefF and Signal-to-Noise Statistics

We started our experiments with a comparative study of ReliefF and S2N scores and associated gene ranking. For all five data sets, histograms of ReliefF and



(a) histogram for actual attribute values; the part with the best ranked attributes is magnified in the upper right corner.

(b) histogram for permuted data

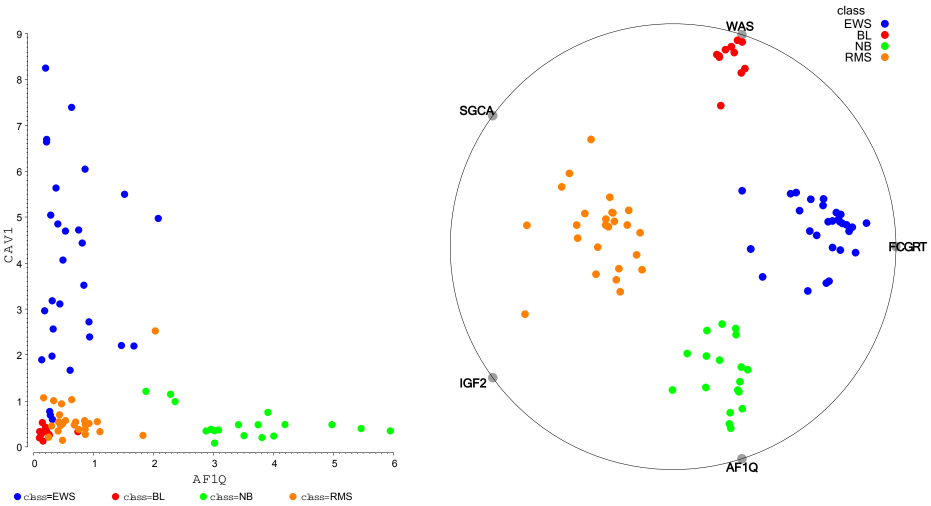
Fig. 2. Histograms of ReliefF on actual and permuted values of attributes (Section 3.2)

S2N scores were qualitatively similar, being skewed to the right, with a group of about 50 to 100 most discriminating genes in the right tail. A permutation test was used to verify if these highly discriminatory genes were assigned high scores by chance. We used permutation analysis and calculated ReliefF scores after random permutation of expression values for each of the attributes. In the interest of brevity, we here only show a histogram for ReliefF scores on MLL data set and its corresponding histogram on randomly permuted data (Figure 2). Note that the part with the highest scored genes (the magnification in Fig 2.a) is far outside the normal-shaped distribution computed on permuted data (Fig 2.b).

The association between the gene ranks obtained by the univariate S2N and multivariate ReliefF gene ranking methods was obtained by computing the non-parametric Spearman correlation coefficient. The Spearman rank correlation coefficient varied importantly depending on which data set we analyzed. The correlation between ReliefF and S2N was highest (0.89) in the MLL data set but as low as 0.24 in the DLBCL data set, indicating that these two scoring functions would typically yield very different ranking and providing grounds for hypothesis that data includes much interactions between genes.

3.3 Results for the Cancer Gene Expression Data Sets

On all the data sets, VizRank found either scatterplot or radviz visualizations with clear separation of diagnostic classes. If let run for an hour on a standard PC, the number of such projections was in the range of ten to twenty, but most importantly, most of them were found in the first few seconds. The last two columns in Table 1 show Vizrank's scores, the average probability of correct classification, for the top-ranked scatterplot and radviz projections. The best scatterplot projections were scored from 87.64% to 98.04%, with the lowest



(a) scatterplot

(b) radviz

Fig. 3. Projections for the SRBCT data set (Section 3.3).

score assigned to the multiclass SRBCT data set. We found that as the number of classes exceeds two, scatterplot becomes less appropriate and radviz with a number of genes not exceeding five provides for excellent separations. For radviz visualizations the scores shown (from 98.27% to 99.75%) hold for the top-ranked projections using only five genes; increasing the number of genes may yield better separation of instances form different classes and thus a higher score. For reasons of space, we here show the visualizations and provide a corresponding discussion only for one two-valued and one multi-valued class problem.

For the leukemia data set the best scatterplot obtained a score of 98% and included genes APLP2 and TCF (Figure 1.a). Notice a clear separation of instances from the ALL and AML classes with only a few outliers. An example of the radviz visualization with good separation of examples from different classes and a Vizrank score of 98% is shown in Figure 1.b. The five genes used in this projection are CFTR, CD19, LTC4S, IL8 and RNS2.

In Table 2 we show genes appearing in the best scatterplots and the ReliefF and S2N ranks belonging to these genes. Out of the 20 genes represented, two were found to be cancer genes and eleven are cancer related according to the Atlas of genetics and cytogenetics in oncology and haematology (http://www.infobiogen.fr/services/chromcancer/index.html). In the table those genes are marked as C and CR, respectively.

Figure 3 shows the best rated scatterplot and radviz visualization with clear separation of instances from different classes for the multicategory SRBCT data set. While the scatterplot does not provide a good separation between all diagnostic classes, separation in radviz is clear. In Table 2 we show twenty genes

Table 2. Twenty best genes from the scatterplot visualizations for the leukemia data set (a) and from the radviz visualization for the SRBCT data set (b). In column 3 and 4 the ReliefF and S2N ranks are shown, respectively. In the cancer column: C = cancer gene, CR = cancer related gene, N = not related to cancer.

Gene	name	Rlf	S2N	cancer
L09209	APLP2	3	3	N
M31523	TCF3	49	0	C
U77948	KAI1	414	26	CR
X85116	EPB72	42	62	CR
M91592	ZNF76	578	112	N
U22376	MYB	50	9	CR
M62982	ALOX12	143	872	CR
Y12556	PRKAB1	123	3082	CR
U82759	HOXA9	9	25	C
U27460	UGP2	588	59	N
L08895	MEF2C	172	72	N
U46499	MGST1	0	2	CR
X03663	CSF1R	856	1189	CR
U26312	CBX3	1152	108	CR
X04741	UCHL1	792	4140	CR
M31211	MYL	57	4	CR
D16469	ATP6AP1	249	341	CR
U68063	SFRS10	1070	134	N
U51240	KIAA0085	104	208	N
U09087	TMPO	100	42	N

Gene	name	Rlf	S2N	cancer
770394	FCGRT	0	1	CR
236282	WAS	26	3	CR
796258	SGCA	23	104	N
207274	IGF2	92	39	CR
812105	AF1Q	5	0	N
183337	HLA-DMA	29	14	N
784224	FGFR4	2	44	CR
866702	PTPN13	19	54	CR
786084	CBX1	10	79	CR
814260	FVT1	35	135	CR
325182	CDH2	66	63	CR
244618	EST	106	25	
377461	CAV1	4	23	CR
296448	IGF2	58	49	CR
629896	MAP1B	117	13	CR
624360	PSMB8	356	18	N
745019	EHD1	12	5	N
1435862	CD99	3	12	CR
383188	RCV1	13	2	CR
767183	HCLS1	21	22	N

appearing in the 41 radviz visualizations with VizRank score of 100%, their Relieff and S2N ranks and report on whether they are related to carcinogenesis according to the Atlas of genetics and cytogenetics in oncology and haematology (http://www.infobiogen.fr/services/chromcancer/index.html). One of the genes from this list is an expressed sequence tags (ESTs). Out of the remaining 19 gene products, 13 are cancer related, which is 68%.

4 Discussion

The results from Table 2 show that signal-to-noise and ReliefF measures rank genes very differently. This was expected, as most of our best-scored visualizations show at least a degree of interaction. For instance, in radviz from Figure 3 neither of the genes can provide a clear separation of classes alone, yet when combined all of the cancer categories are perfectly separated.

Contrary to our expectations, though, we would expect a better match between ReliefF ranking and genes included in best set of visualizations. We hypothesize that the problem is in the number of attributes in the data set ReliefF considers as a context. When ReliefF evaluates an attribute it selects some reference examples and searches for the most similar examples from the same class and from the other classes. The context with too many genes can mask the effect of genes in interaction with the estimated gene, thus disabling ReliefF to appropriately account for interactions.

We compared our selection of "marker" genes in the leukemia data set to the genes selected by other methods. The best 20 genes from the scatterplot visualizations are different from the best ranked genes by Golub *et al.* [1] except for HOXA9 and EPB72. The primary reason is the use of univariate feature selection in their study. Also, in our study, we joined the test and training sets used by Golub *et al.* and performed the gene ranking and visualization methods on the combined set.

In contrast to the leukemia data set, the selection of our "marker" genes for the SRBCT data set compared with the genes ranked best by other methods is very similar. We found a very high consensus between our selection and the selection based on artificial neural networks [4]. 19 genes from the best scatterplot visualizations and 16 out of 20 from the best radviz visualizations were also selected by the ANN method. Interestingly, there is also a very high consensus (11 and 12 genes out of 20 included in best ranked scatterplot and radviz visualizations) between our method and the method of Fu [7] based on support vector machines.

5 Conclusion

The most striking result from our work is how easy it is to find a simple two-dimensional scatterplot or radviz visualization that clearly, non-ambiguously separates cancer diagnostic classes based on expression measurements for a few selected genes. This holds for all five data set studied, but the same observations also applies to about fifty other publicly available data sets we have studied but not reported in this paper. VizRank, a method we used to find good visualizations, often identifies the best ones within seconds of runtime. This is a significant achievement, especially when compared to hours of required runtime reported in a recent study that uses support vector machines combined with a set of other machine learning and feature selection approaches [5], and considering a clear presentation of results offered by these visualizations.

To a surprise came a relatively poor performance of ReliefF, which was expected to outshine the univariate gene scoring, but instead performed similarly. We have yet to fully understand why this is so, as the finding does not match with those from the systematic study of ReliefF on data sets that contain much fewer attributes [16].

Gene expression data visualizations reported here provide evidence that cancer diagnostic classes can be clearly separated when using the expression data from only a few genes. VizRank also provides a way for robust selection of genes without the need for a particular scoring function. Our future work aims at using top rated visualizations for probabilistic classification, thus also providing grounds for comparison with other, much more complex but nowadays prevailing computational methods in gene expression cancer diagnosis area.

References

1. Golub, T.R., Slonim, D.K., Tamayo, P.*et al.*: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
2. Shipp, M.A., Ross, K.N., Tamayo, P.*et al.*: Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine **8** (2002) 68–74
3. Nutt, C.L., Mani, D.R., Betensky, R.A.*et al.*: Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res **63** (2003) 1602–1607
4. Khan, J., Wei, J.S., Ringnr, M.*et al.*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. 7 **6** (2001) 673–679
5. Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics (2004) 33 – 46
6. Su, A.I., Welsh, J.B., Sapinoso, L.M.*et al.*: Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res **61** (2001) 7388–7393
7. Fu, L.M., Fu-Liu, C.S.: Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. FEBS Letters **561** (2004) 186–190
8. Gamberger, D., Lavrac, N., Zelezny, F., Tolar, J.: Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. Journal of Biomedical Informatics **37** (2004) 269–284
9. Wang, Y., Tetko, I.V., Hall, M.A.*et al.*: Gene selection from microarray data for cancer classification–a machine learning approach. Computational Biology and Chemistry **29** (2005) 37–46
10. Kira, K., Rendell, L.: A practical approach to feature selection. In: Proceedings of the Ninth International Conference on Machine Learning. (1992) 249–256
11. Kononenko, I., Simec, E.: Induction of decision trees using relieff. In: Mathematical and statistical methods in artificial intelligence. Springer Verlag (1995)
12. Brunsdon, C., Fotheringham, A.S., Charlton, M.: An investigation of methods for visualising highly multivariate datasets. Case Studies of Visualization in the Social Sciences (1998) 55–80
13. Leban, G., Bratko, I., Petrovic, U., Curk, T., Zupan, B.: Vizrank: finding informative data projections in functional genomics by machine learning. Bioinformatics **21** (2005) 413–414
14. Singh, D., Febbo, P.G., Ross, K.*et al.*: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1** (2002) 203–209
15. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R.*et al.*: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics **30** (2001) 41–47
16. Sikonja, M.R., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Machine Learning **53** (2003) 23 – 69