Intelligent Data Analysis in Medicine and Pharmacology: A Position Statement

Riccardo Bellazzi [1] and **Blaž Zupan** [2]

Abstract. Intelligent data analysis methods support information extraction from data by exploiting domain's Background Knowledge. We address several issues regarding definition, use and impact of these methods, and investigate for their acceptance in application domains of medicine and pharmacology by a MEDLINE search. The authors of the paper believe that the basic philosophy of IDA is to be application driven: its goal is to develop, adapt, or re-use existing methods to solve a specific problem. Sticking to application driven approach may help to prove the points on cost-effectiveness and may increase the awareness and acceptance of these methods in medical community.

1 Introduction

A crucial problem in Medical Informatics (MI) is to provide final users (physicians, patients, researchers) with instruments for interpreting data, so that final decisions (diagnosis, therapy) will be better *with the help of computers* than *without* [20]. In other words, a central role of MI is to help users in transforming *data into information*. This process is usually heavily mediated by the expert knowledge, and its automatization requires the application of techniques that should be able to integrate data analysis and knowledge representation.

Intelligent Data Analysis (IDA) refers to all methods that are devoted to automatically transform data into information exploiting the Background Knowledge on the domain.

Background Knowledge (BK) is the domain knowledge obtained from the literature or from domain experts. In general, BK refers to the meta-information available when analyzing the data [9]. For example, a typical kind of meta-information is the contextual knowledge that allows for a proper interpretation of the data, taking into account the particular context in which they are collected. For instance, the same value of blood glucose level for diabetic patients bears different meaning to a physician when measured in different time periods of the day, or different period of the year (glucose level may be high during Christmas holidays). IDA researchers should address the problems of acquisition, encoding, and exploitation of BK. IDA, of course, does not exclude the user intervention during the information extraction process, but aims at reducing interaction through the use of BK, thus reducing the costs of data analysis, both in terms of users' time and resources.

The primary goal of IDA is to provide methods that support data understanding; so, although its immediate goal is not the construction of data models for prediction or classification, the relations and principles discovered by IDA must be designed to provide support for the decision making process.

It should be hence clear that IDA is neither KDD nor Data Mining, although it may be part of the KDD process and some of its methods are also Data Mining methods. The focus of IDA is on the marriage of data analysis and BK for providing the user with information.

Medicine and pharmacology are "natural" application domains for IDA methods. In bio-medical data analysis, "human reasoning is pivotal" (and therefore BK is pivotal, too), since the "object and the subject of decision making is human being" [8]. This means that the "value" of each datum may be higher than in other contexts: experiments can be costly, due to personnel and instrumentation involvement and/or to the patients discomfort; the data set can be small, reporting non-reproducible situations. Nevertheless, the data may also be affected by several sources of uncertainty, from measurements errors to missing data, or from data coding errors to information buried into textual reports. Physicians and researchers deal with such difficulties by exploiting their knowledge on the domain. IDA methods can help them, by integrating BK in the data analysis process.

The chance for IDA to have an impact in MI research and clinical practice is great, but the number of "in-use" applications can be counted with one-hand fingers. In the following, we will try to understand why, and what could be future research directions for IDA exploitation.

2 Current approaches, applications, and related problems

Intelligent data analysis in medicine and pharmacology (IDAMAP) relies on variety of different approaches. A proposal of the classification of IDAMAP techniques and in-deep discussion of several of them can be found in [13]. This section presents our opinion on several issues related to IDAMAP and its applicability to real medical and pharmacological problems and concludes with the attempt to asses which IDA methods are the most often used within these application fields.

2.1 Elicitation of Background Knowledge

One important problem that is still to be solved with IDAMAP is the elicitation of BK. Following the statements of the introduction, the philosophy of IDA reflects the one of Bayesian statistics: prior knowledge (BK) is modified by data to obtain posterior knowledge (data analysis results). Nevertheless, a well known problem is that physicians want to know posteriors, but they are not able to provide priors. Traditionally, Bayesian statistics overcome this problem using flat priors, that are unable to really express BK. This poses serious limitations to the application of Bayesian approaches in typical IDA settings. Some solutions to this problem have been proposed by the Bayesian community: priors are elicited through ranges or intervals, that are updated on the basis of data collection [19, 17]. Other solutions refuse the Bayesian statements, and first elicit and then update BK using different approaches, like fuzzy logic, Inductive Logic Programming and hierarchical structuring. We believe that IDA researcher may select the proper method to be applied for a certain application also in relation with the easiest way of obtaining BK.

[1] Dipartimento di Informatica e Sistemistica, Universitá di Pavia, Italy
[2] Jožef Stefan Institute, Ljubljana, Slovenia

2.2 Experimental design and its effect to IDAMAP performance

There were several recent performance comparison studies which confirmed the superiority of a naive Bayesian classification method over more elaborate and sophisticated IDAMAP methods [14]. The experiments were usually run over the medically-related "benchmark" datasets that reside at UCI Machine Learning repository [15]. We believe it may be the experimental design under which these data were prepared that fosters the use of naive Bayes.

For instance, one of the authors recently attempted to explain the reasoning and motivation about decision trees to one of his friend physicians. After working for several hours on his dataset, physician admitted that although he understands the "logic" and "working principles" about the decision trees, it is uncommon for him to think this way. Namely, in the medical school, he learned to collect the data so that to include only the possibly independent features, from which he could devise, say, a diagnosis such that a feature would contribute either in a positive or negative way and the diagnosis (decision for or against some pathologic stage) then being related to a simple weighted sum of the effects of these features.

It may be that such selection of independent features that can be linearly combined was also the present in the setup under which the medical data at the UCI ML repository were gathered, thus contributing to the success of naive Bayes over other methods. However, with the emerging technology of data warehousing in modern hospitals and health systems, data are collected having no special purpose (or hypothesis to check) in mind. It may well be that the time has come were more sophisticated data analysis tools that properly treat nonlinear features and their interdependencies will find their comparative advantage.

2.3 Probabilistic classification

The standard classification tools (like decision trees and rules) that use "black-and-white" classifications in the sense that, using the machine induced model, a data instance (a case) is classified to a single class appear to be a good candidate for IDAMAP. But although these tools, at least in methodological terms, include the mechanisms for probabilistic reasoning, these are rarely used in both model induction and verification phases.

Let us illustrate the insufficiency of such single class classification on an example. One of the authors was recently involved in the analysis of prediction data for prostate cancer pathologic stage. Previous statistical analysis of the same data show that for a lymph node involvement there is not a single combination of values of (discrete) prediction features that data instances classified to this stage would be in majority [16]. Similar was true for the lymph node involvement. Thus the resulting decision trees could never predict any of these two stages. Yet, it is exactly for these two stages that the prediction is the most interesting for physicians — for example, they would highly value, given some clinical patient data, if the prediction model would be able to identify an increased probability (say of about 10%) for these stages, as such situation would trigger a more comprehensive work-up.

Many medical domains may require similar treatment. The standard classification-oriented data analysis tools would therefore need to be modified in order to induce rules that predict class probabilities rather than classes. The verification of prediction models would then require some other methods than those based on classification accuracy. Whereas a good candidate is a Harrell's c index (e.g., concordance index, see [4, 6]), it only measures the quality of differential prediction. We therefore believe that in the field of machine learning classification methods for IDAMAP there may be substantial room for improvement both in devising methods that use probabilities and methods to test for the quality of induced models.

2.4 IDAMAP in use

Although the above points identify some of the potential limitations of current IDA techniques and require further investigation, they are not the main reason for the relative failure of IDA (and AI in general) as a routine instrument in medicine and pharmacology. The basic motivations can be found taking a look at the literature.

At the beginning of the nineties, several papers addressed the problem of the failure of expert systems in clinical practice. In particular, a couple of papers of Heatefield and Wyatt [5] and Musen [7] pointed out that one of the most important reasons for the AI-based systems failure was a lacking of comprehension of the real nature of the medical decision problems, and a substantial misunderstanding of the final user needs.

Heatefield and Wyatt reported the results of a study published by Haynes [3], in which the topics of 346 MEDLINE searches were compared with the decision problems addressed by the expert systems developed from 1973 to 1992. This study showed that while the 41% of the MEDLINE searches was on therapeutic problems, only the 19% of the expert systems dealt with therapy; moreover, while only the 6% of the MEDLINE searches was on diagnosis, the 53% of expert systems was a diagnoser. On the other hand, the analysis carried out by Musen, stressed the lacking of contextualization of expert systems: medical knowledge have been designed without considering the real environment in which they were supposed to work. AI-based systems should instead take into account the information flows of the medical setting, thus providing an instrument to health professionals for improving their daily work. Exploiting the network and knowing the work-flow of physicians may help in designing tools that really address the user needs. AI should be hence intended as a commodity for user, fully integrated within the set of information technology tools nowadays available, from data bases to web browsers.

Although in the following years the situation has been slightly improved, at least from the technological point of view, in the last Artificial Intelligence in Medicine Conference, in Grenoble (AIME97) [12], physicians were still criticizing the AI community that too much emphasis is put on diagnostic problems. Another criticism at the same Conference was that there was not a single contribution in the area of outcomes analysis, which is, especially in the USA, gaining an increasing interest among the physicians and may substantially gain from computerized decision support and analysis tools.

The real problem is that, rather surprisingly, there is still a dichotomy between the researcher goals and the user needs. From one side, researchers would like to solve complex diagnostic problems with complex methodologies, in competition with physicians minds. From the other side, physicians would like to have help in their routine activity, handling problems in a better way (e.g. considering more data than what they can do) or in a faster way (e.g. handling more patients in the same period of time). As a consequence, physicians prefer simple tools that can help them in managing more quickly easy problems, with respect to sophisticated tools able to handle complex problems, but that in any case requires physicians expertise. Following the physicians point of view, only few diagnostic problems are now perceived as "real problems"; the improvement in clinical diagnostic examinations move the attention of health care providers to the patients management, in its broadest sense, includ-

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
AI*	71	76	144	148	93	166	181	207	157	133
expert systems*	92	156	186	173	121	130	172	182	129	108
neural networks* 3)	0	0	2	62	160	274	317	388	448	373
fuzzy logic*	0	0	0	0	6	21	33	60	51	45
genetic algorithms	0	1	2	0	3	9	9	27	34	30
regression trees	0	1	2	5	2	8	2	13	5	16
classification tree	3	1	10	5	3	4	6	8	4	3
c4.5	0	0	0	0	0	0	0	1	2	0
CART	0	1	0	2	2	5	2	7	5	5
rule induction	1	2	1	1	3	2	3	2	4	4
rough sets	1	0	1	3	0	2	3	5	1	3
ILP	0	0	0	0	2	1	4	0	2	0
causal networks	0	0	0	3	1	2	6	2	1	6
temporal logic	0	2	0	0	0	0	1	1	0	1
CBR	0	0	1	0	1	5	7	5	3	0
data mining	0	0	0	0	0	0	0	1	1	7
machine learning	3	0	3	5	7	12	15	22	18	8
IDA	0	0	0	0	0	0	0	0	1	0

Table 1. MEDLINE references by year and keyword or keyword combination (those with * appear in MEDLINE's subject index). Where needed, results were manually checked so that to include accidental match-ups (say CART as a machine learning technique and as a hospital equipment).

ing cost analysis. This means that the IDA community should focus on providing support where it is needed, like for example in defining new health care protocols, by investigating if data analysis may provide new evidence on how to design or to optimize diagnostic or therapeutic protocols.

In his excellent article on the adolescence of AI in Medicine, Shortliffe [18] stated the following critical conditions to be fulfilled for a successful integration of AIM systems into patient-care settings and their widespread use:

- enhancement of *training* in medical informatics,
- development of communication, data exchange and information retrieval *infrastructure*,
- setting the *standards* for communication, data and knowledge exchange.

Shortliffe's observations were made in early '90s. We believe that also in late '90s, it is the fulfillment of the same set of conditions that the success of IDAMAP techniques depend.

With a recent growth in number and quality of informations systems within health care and medicine, the required information infrastructure is now or will be soon available. In addition, data warehousing and data marts (data warehouses of a smaller scale) are being introduced for the purposes of wide-range collection and managing the medical data. The medical community is becoming aware that long terms general purpose data collection may be useful to extract to extract information for different goals and in different times. From technological point of view, the Web based systems and development tools are available, and represent the best instrument for collaborative efforts of the community. Examples of successful use of such technology include Human Genome Project [10], Human Brain Project [11], G-7 project on Cardiology, etc.

Recently, several textbooks in medical informatics have been published that include IDAMAP techniques [20, 2]. We hope that this will reflect the course programs, and increase the level of awareness and understanding of the IDAMAP technologies in medical community.

For IDAMAP, it is the standards which, of the three Shortliffe's conditions, will be the hardest or the least likely to develop. Being a part of decision support tools within information infrastructure, IDAMAP tools rely on existing underlying medical data communication and patient records standards. It is unrealistic to expect that IDAMAP community will develop its own specific standards, however the researchers and users of IDAMAP should be engaged in setting and publishing a set of guidelines that recommend a set of approaches for a specific analysis problems.

In respect with IDAMAP as a commodity for the users, there is a lack of tools that are readily available and easily adaptable to be included in medical data infrastructures. Open systems, reusability, and component based approach are the keywords. We believe that in its present state the technology is now mature to support such an approach.

2.5 IDAMAP on MEDLINE

In order to attempt to estimate which and to what extend the IDAMAP techniques are being used, we performed a search on a MEDLINE reference database. The MEDLINE was chosen because it is a primary source where a physician would look for a existing solution of a certain problem. The purpose of this search was only to have a grasp of current trends and not to provide an extensive and systematic evaluation. The results (Table 1) are presented as number of papers that include specific keyword in either title, abstract, or subject index.

Of all techniques that IDAMAP potentially includes, neural networks, fuzzy logic, genetic algorithms are most referred to. It is a matter of discussion how much these techniques focus on IDAMAP problems as stated in the Introduction. For neural networks, for example, there is an increasing interest in symbolic rule induction from learned networks and for the use of background knowledge to help determining the nets architecture and weights [1] — however, this is barely reflected to medical applications as there are only a few related references. To a surprise, the tools that are considered a state-of-the-art in machine learning and are widely used and cited within this field, are scarcely used in medicine. For example, C4.5 — a widely used and referred tool within machine learning for decision tree induction — is reported to be used in only a few papers. The MEDLINE scores are somehow more favorable for CART, another state-of-the-art decision tree inducer, probably also due to its inclusion in else popular S-PLUS statistical data exploration package.

Overall, apart of three exceptions mentioned, IDAMAP techniques seem to be rarely used in the work reported in medical and pharmacological articles. The further verification of references counted in Table 1 revealed that the majority of the IDAMAP related articles are methodological and appear in journals such as *Artificial Intelligence in Medicine*, *Neural Computation* and *Methods of Information in Medicine*. The rare use of IDAMAP methods is in a way surprising, since these techniques are considered well developed and documented within computer science community. The results may indicate that in general the IDAMAP methods are not very known in medical community, with only few applications documented to send an encouraging message to medical experts and convince them of IDAMAP applicability.

3 Some proposals for future directions

In order to foster the usage of the IDA techniques in medicine and pharmacology, we believe that the community should prove that IDA, where applied, may enhance the cost-effectiveness of health-care.

In order to achieve this goal, several research directions could be undertaken; among them, we highlight the following topics:

Time: IDA must help users in saving *time*. This can be pursued by:

- exploiting information technology to its limits, providing services to all users involved in health care management (physicians, patients, nurses, managers), and hence implementing complex techniques as commodities, fully integrated within the information system
- reducing physicians involvement by using BK, and requiring their intervention only in the critical phases of the data analysis process

Cost: the use of IDA techniques may be really helpful in reducing the cost of treatment. A crucial emphasis of IDA application must be put on providing information for the selection of cost/effective treatment, for the decision of cost/effective therapeutic protocols, and so on.

Quality: the results of IDA research must be disseminated by highlighting their quality through:

- Publication on journals that has impact on the bio-medical community, such as Medical Decision Making or the Journal of American Medical Informatics Association.
- Design and implementation of IDA-based software components to be integrated in existing information systems in health-care and medicine; in this sense, Web-based technological solutions may represent an important vehicle for the spreading out of IDA methods.

The authors of this paper believe that the basic philosophy of IDA is to be application driven: its goal is to develop, adapt, or re-use existing methods to solve a specific problem. To clarify this point, it is more easy to describe what IDA is not: develop a new method for classification, get several medical dataset from Irvine repository, prove that classification accuracy is increased with respect to a number of other methods, and claim this is why the methods is suited for medical applications. We believe that sticking to application driven approach may help to prove the points on cost-effectiveness and may increase the awareness and acceptance of these methods in medical community.

REFERENCES

[1] Andrews, R., Diederich, J., and Tickle, A. B., “A survey and critique of techniques for extracting rules from trained artificial neural networks,” *Knowledge Based Systems*, vol. 8, no. 6, pp. 373–389, 1995.

[2] Coiera, E., *Guide to medical informatics, the Internet and telemedicine*, Chapman & Hall Medical, London, 1997.

[3] Haynes R, McKibbon K., Walker C et al, Online access to MEDLINE in a clinical setting. *Ann. Intern. Med.*, 112, 78-84, 1994.

[4] Harrell, F. E., Lee, K. L., Mark, D. B., Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, 15:361–387, 1996.

[5] Heatefield H.A., Wyatt J. Philosophies for the Design and Development of Clinical Decision-Support Systems. *Methods of Information in Medicine*, 32, 1-8, 1993.)

[6] Kattan, M.W., Ishida, H., Scardino, P.T., Beck, J.R., Applying a neural network to prostate cancer survival data, in (N. Lavrač, E. Kervanou, B. Zupan, eds.) *Intelligent data analysis in medicine and pharmacology*, Kluwer, Boston, 295–306, 1997.

[7] Musen M. Architectures for Architects. *Methods of Information in Medicine*, 32, 12-13, 1993.

[8] J.H. Van Bemmel Handbook of Medical Informatics. Bohn Stafleu Van Logum, Houten/Diegem, 1996.

[9] Hand, D.J., Intelligent Data Analysis: Issues and Opportunities, in *Advances In Intelligent Data Analysis- Lecture Notes in Computer Science*, (Eds: X Liu, P Cohen and M Berthold), Springer Verlag, Berlin, 1–14, 1997.

[10] The Human Genome Organisation. [http://hugo.gdp.org/]

[11] The Human Brain Project. [http://www-hbp-np.scripps.edu/Home.html]

[12] Keravnou, E., Garbay, C., Baud, R., and Wyatt, J. (eds.) *Artificial Intelligence in Medicine*, Springer, Berlin, 1997.

[13] Lavrač, N., Keravnou, E., Zupan, B. (eds), *Intelligent data analysis in medicine and pharmacology*, Kluwer, Boston, 1997.

[14] Kononenko, I., Bratko, I., Kukar, M., Application of machine learning to medical diagnosis, In (Michalski, R.S., Bratko, I., and Kubat, M., eds.) Methods and Applications of Machine Learning, Data Mining and Knowledge D iscovery, John Wiley & Sons Ltd., 1997.

[15] Murphy, P.M., and Aha, D.W., UCI Repository of machine learning databases [http://www.ics.uci.edu/˜mlearn/mlrepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1994.

[16] Partin, A. W., et al., Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. *JAMA*, 277(18) 1445–1451.

[17] Ramoni, M., and Sebastiani, P., The use of exogenous knowledge to learn Bayesian Networks for incomplete databases. in *Advances In Intelligent Data Analysis- Lecture Notes in Computer Science*, (Eds: X Liu, P Cohen and M Berthold), Springer Verlag, Berlin, 537-548, 1997.

[18] Shortliffe, E.H., The adolescence of AI in medicine: will the field come of age in the ’90s?, *Artificial Intelligence in Medicine*, 5 (1993) 93-106.

[19] Spiegelhalter D., Dawid A., Lauritzen S., Cowell R., Bayesian Analysis in Expert Systems. *Statistical Science*, 8 (1993) 219-283.

[20] van Bemmel, J. H. (editor) *Handbook of Medical Informatics*, Bohn Stafleu Van Loghum, Houten/Diegem, 1996.