A Unified Framework for Data Modeling on Medical Information Systems

José Neves Paulo Cortez Miguel Rocha António Abelha José Machado
Vitor Alves Sousa Basto Henrique Botelho João Neves
Universidade do Minho, Largo do Paço, Braga, Portugal

Abstract. *Medical Information Systems (MIS)* are seen as a way of optimizing the use of existing health-care infrastructure, without resorting to new and costly hospital (re)construction. The qualitative (re)design of such an environment requires a basic understanding of patient and doctors related characteristics and capabilities. Patient care, patient education, medical education, and clinical research need to be considered to meet the basic requirements on the level of services desirable, determined on the basis of the patient's length of stay; i.e., used for modeling the significant entities of such a world. The aim is to extract conclusions for the level of services provided to the users. One's concept will capture, as well as will integrate, the basic design principles under which *MIS* may be set.

1 Introduction

Medicine has been for some years a very attractive domain for *Computer Science (CS)* researchers, in general. There is a great potential for information automation, and a lot remains to be done. *Medical informatics (MI)* is indeed becoming an issue of study in which *Medicine* and Computing overlap. Another reason for this increasing interest is costs. Today's strained health-care economics makes it necessary for expensive resources to be efficiently used, and requires a balanced management.

CS researchers have long used *Medicine* to elaborate on their own work. The field is probably one of the most knowledge intensive ones, loaded with human reasoning, with most of the procedure relying exclusively on the clinical experts. This makes health-care a perfect target for *CS*, since conventional systems are naturally bounded by their lack of rich knowledge representation and proof schemes. *Medicine* allows for the testing and exposing of new ideas and techniques by *CS* scientists, which creates a certain complicity between the two communities. Early knowledge-based systems like MYCIN, a medical consultation system for diagnosis of blood infectations (e.g., meningitis) and the recommendation of drug (e.g., antibiotic) treatment, or CADUCEUS, an expert system for medical diagnosis developed by Jack Mayers and Harry Pople at the University of Pittsburgh, 1985, which is an enhancement of INTERNIST, in that it incorporates causal relationships in its diagnosis process, shows the benefits of using *CS* in complex tasks. The proliferation of such systems during the eighties helped,

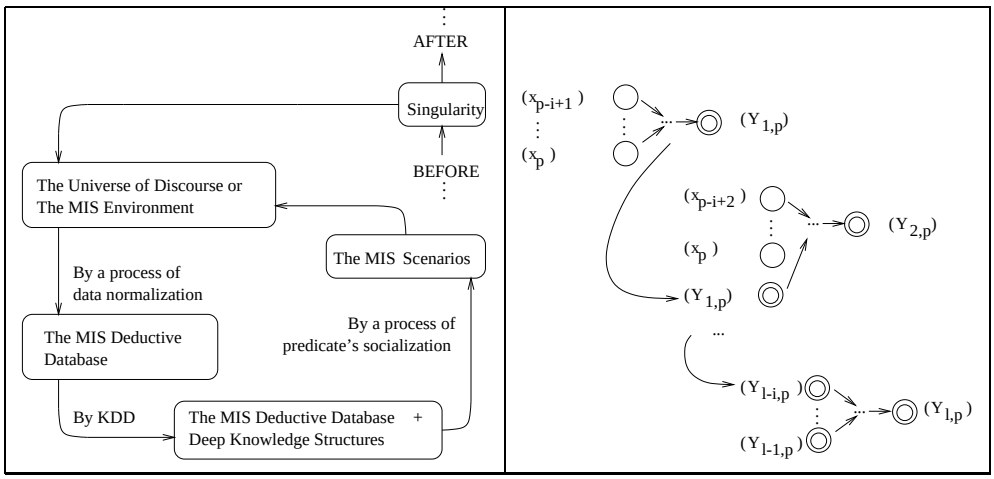

Figure 1: A Cosmic View of the Design Process of *MIS* and Getting Y

in a certain way, to bridge the gap between theoretical, academia *CS*, and real world applications.

What is the best approach to help doctors and patients? It is believed *Artificial Intelligence (AI)* enters at this point. A system intended to be a support tool for clinicians must have its focus on knowledge representation and reasoning. Such system should be able to explain and justify its conclusions. On the other hand it has to be flexible enough to allow for simulation, training of novices and maintenance.

2 The Methodology

Any complex software system can be viewed as a collection of independent cooperating units or entities, each of which implements a subset of an application's total functionality. In some sense a cooperating unit will be called an agent, where relationships create bridges and methods among them [5]. The underlined cognitive architecture will be defined as the portion of the system that provides and manages the primitive resources of such an entity; i.e., the substrate upon which a physical symbol system will be realized. Such a system, part symbolic, part sub-symbolic, will make use of *Neural Networks (NNs)*, understood as autonomous agents in their own right. Indeed the concept of intelligence is keen to many different interpretations. In a this setting it may amalgamate the best of two worlds:

- the *symbolic AI* or *deliberate thinking* paradigms, where an agent is assumed to be driven by logic [7], and
- the *nouvelle AI*, *behavior-based AI*, or *situated agents* paradigms, where intelligence comes into being a property exhibited by the system's agents [4][3].

For instance, if one is working in areas such as detection of pathological occurrences in images of cervical smears or breasts images for automatic screening, soon realize that his/her experience and knowledge cannot be easily expressed using exclusively conventional symbolic systems [2].

In the design phase of a *MIS*, strategies need to be developed in order to apply some mitigation measures in the most cost effective manner. The strategy for each

situation has to be based on the results of the cost effectiveness of the improvement options, the level of potential state investment required to achieve positive returns, and the potential long term benefits for both the state and the user. The improvements to be implemented may range from simple device enhancements to major reconstruction of facilities. The objective is to improve the quality of health-care, and, consequently, to reduce the loss of lives.

To quantify not only this parameter, and claim cost savings, but also to characterize the improvements to be implemented, and the fact that the data involved in the process has far out paced one's ability to interpret and digest it, it is felt a need for a new generation of tools and techniques for automated and intelligent construction and analysis of *MIS*. These tools and techniques make a new field of *Knowledge Discovery in Databases (KDD)*, and also make the core of one's approach (Figure 1).

KDD is concerned with the identification and description of data patterns [6], which, in this work, are interpreted as classes or concepts. In order to obtain such patterns it was followed the work by Agrawal [1] on association rules, expressions of the form $X \Longrightarrow Y$, where X and Y are sets of items (the relations' attributes). The intuitive meaning of such a rule is that database transactions which contain X tend to contain Y.

Definition: Let $I = \{i_1, i_2, ..., i_n\}$ be a set of terms, a set of the database's relations attributes. Let D be a set of transactions, where each transaction Tr is an attribute set such that $Tr \subseteq I$; i.e., such that I is a set of attributes over the binary domain $\{0, 1\}$. Let a tuple Tu of the database D be represented by identifying its attributes with the binary value 1. Let TID be a unique identifier associated with transaction Tr. Let a set of attributes $X \subset I$ be called an attribute set. It may be claimed that:

- A transaction Tr contains an attribute set X, if $X \subseteq Tr$.
- An *association rule* is an implication of the form $X \Longrightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$ (the empty set).
- The rule $X \Longrightarrow Y$ holds in the transaction set D with confidence C, if $C\%$ of transactions in D that contain X also contain Y.
- The rule $X \Longrightarrow Y$ has support S in the transaction set D, if $S\%$ of transactions in D contain $X \cup Y$.

The set D stands for the extension of the relations (or predicates):

- *binary-identify-characteristics()*;

...

- *binary-healthy-and-happy-community()*.

where *binary-*()* stands for a relation whose attributes are defined over the binary domain $\{0, 1\}$, and * stands for itself. To get X (from the rule $X \Longrightarrow Y$), are used the procedures presented at [1]. To get Y (from the rule $X \Longrightarrow Y$), it is followed the *Artificial Neural Network (ANN)* approach to *Time Series Forecasting* presented at [8]. The *ANN* will take the form $L_i - L_h - 1$, for a network with L_i input nodes (or X), L_h hidden nodes, and *one* output node. The training cases, given by X, will take the form:

$$\begin{array}{lcl} x_1, x_2, ..., x_i & \rightarrow & x_{i+1} \\ ... & \rightarrow & ... \\ x_{p-i}, ..., x_{p-1} & \rightarrow & x_p \end{array}$$

for a set Y of cardinality p. After training, one-step ahead forecast (i.e., getting $Y_{1,p}$), is produced be feeding the *ANN* with the last known values for set X, namely $Y_{1,p} = ANN(x_{p-i+1}, ..., x_p)$. where $ANN()$ stands for the function modeled by the *ANN* and $Y_{r,s}$ for the forecast in the s period to r periods ahead of s. Multi-step ahead forecasts (looking at different Ys) are done by feedback of the previous ones (Figure 1).

3 Conclusions

For several decades, *CS* scientists and physicians have been building computer programs to diagnose medical illness and to recommend therapy. The resulting work helped to define and seed the development of expert and decision support systems in the area. On the other hand medical diagnosis and patient management procedures assisted to ensure that domain specific assertions and extensive knowledge about a problem domain are more crucial to problem solving, that are domain specific principles of reasoning. Thus, early generation of hypotheses seems to provide leverage for the diagnostician. Building on these results, an approach to data modeling on a *MIS* environment was set. It explores new techniques for the discovered of rules and data, driven by both deliberate thinking and behaviour based *AI*. It makes possible to dig in knowledge without generalization or background knowledge, a step forward in attribute oriented induction in data mining. Besides further advances on *KDD* on the attribute-oriented induction area, we are also investigation other data mining methods potentially applied to *MIS* design, with a particular emphasis on non-destructive worlds with incomplete information.

References

[1] Agrawal, R.; Manila, H.; Srikant, R.; Toivonen, H; and Verkamo, A. I. 1996. Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining*, Fayyed, U.M.; Shapiro, G. P.; Smyth, P.; and Uthurusamy, R. eds. The MIT Press, Menlo Park, California, USA.

[2] Anthony, D. *The Use of Artificial Intelligence in Medicine*, Warwick University, September 11, 1994.

[3] Bond A. A Computational Model for Organizations of Cooperating Agents, *SIGOIS Bulletin II*, 1990.

[4] Carbonell, J. *Machine Learning Paradigms and Methods*, The MIT Press, England, 1990.

[5] Cavedon, L. and Tidhar G. A Logical Framework for Multi-Agent Systems and Joint Attitudes, In *Proceedings of the First Australian Workshop on Distributed Artificial Intelligence*, University of New South Wales, Australian Defense Force Academy, Australia, 1995.

[6] Frawley, W.; Piatetsky-Shapiro, G. ; and Mathews, C. Discovery in Databases: An Overview. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley, Eds. The AAAI Press, Menlo Park, California, USA, 1991.

[7] Maes, P. Situated Agents Can Have Goals, In P.Maes ed. *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*, The MIT Press, London, England, 1990.

[8] Neves, J.; and Cortez, P. 1998. Combining Genetic Algorithms, Neural Networks and Data Filtering for Times Series Forecasting. In *Recent Advances in Circuits and Systems*, Mastorakis, N. E. ed. The World Scientific Press, London, UK.