Generalized Net Model of the Knowledge Discovery in Medical Databases

Orozova D.[1*], Sotirova E.[2], Chountas P.[3]

[1] Burgas Free University
6, Stefan Karadja Str., 8000 Burgas, Bulgaria
E-mail: orozova@bfu.bg
[2] Burgas University "Prof. A. Zlatarov"
8000 Burgas, Bulgaria
[3] School of Electronics and Computer Science
University of Westminster, London, UK

Summary: A generalized net model has been developed of the process of discovery of hidden knowledge patterns in databases, which defines the necessary consequence of activities in order to extract knowledge from the input data. The suggested approach is applied to systems for storage and processing of medical data.

Keywords: Knowledge discovery in databases, Generalized nets, Instrumental tools, Medical databases.

1. INTRODUCTION

Information technologies (IT) assist the overall functioning of healthcare units. One of the well developed aspects of application of IT in medicine is the usage of electronic patient files that collect and store data of the patients' health status, consultations and treatments, medical investigations, etc. Such information systems maintain and provide access to enormous nomenclatures of ICD codes, drug lists, etc. These allow for the automatic creation and maintenance of registries, such as Dispensary registry, Maternal care registry, and other registries necessary for particular healthcare units.

Provision of fast and easy access to such accumulated data vastly improves healthcare. IT penetrate medicine in various directions like automated collection, processing and storage of increasingly bulky results of clinical and paraclinical investigations and observations with specialized equipment, as well as processing of the patient-centric information from hospital and home treatment. The patients' information accessible via Internet is stored in distributed databases.

[*] *Corresponding author*

Systems for decision making support are utilized for data analysis, models development and evaluation of the alternatives. The so produced analyses provide clear argumentation for the decisions made. Such decisions are usually case specific, quickly adaptable and the factors for their evaluation are not easily determined and are rarely revised.

This class of systems exhibit vast analytical possibilities. The decision support techniques used there may be:

- model oriented techniques: simulation, optimization,
- techniques from the field of artificial intelligence: expert systems, neural networks, fuzzy logic, intelligent agents,
- data oriented techniques: OLAP, data mining.

The enormous quantities of accumulated data drastically exceed the human's possibility for their effective utilization without powerful specialized tools for data analysis. When heavy databases have to be used, certain questions arise like: How may these data be analyzed in reasonable time?, How may characteristic data representatives be outlined?, and How to overcome the problematic situation "Rich in data, poor in knowledge"? The increasing divide between data and knowledge has led to research of the process of knowledge discovery that aims at turning the "tombs of data" into "goldmines of knowledge" [3-7].

The process of knowledge discovery in databases includes the stages of data preparation, selection of informative indicators, data cleaning, application of the data mining method, data processing and interpretation of the results.

A major accent in this process in the approach of data mining, which enables the retrieval of information in the form of rules, describing the relation between data properties, results of classification and clustering of the data. Data mining software is used for analysis of the data, but it further includes technologies enabling the discovery of hidden patterns and interrelations within various data excerpts. Data, derived in the process of data mining, may be utilized in the Executive Information Systems for determining the strategies of development of a given organization or a process.

Oftentimes, expert systems take part in the data processing. They solve difficult and experience demanding problems. Main elements of these systems are the knowledge base, the inference mechanisms and the user interface. The knowledge base contains not only data from the applied area, but also expert knowledge derived from long-term practitioners and specialists working in the area. This kind of knowledge is recorded in the database where using the development area tools they may be incremented, modified and updated. The inference mechanism uses the knowledge and heuristics to present recommendatory solutions to the problems formulated in front of the system. The course of system's reasoning that leads to devising the solution are documented and submitted for analysis, if needed.

2. MODEL OF THE KNOWLEDGE DISCOVERY PROCESS

The generalized net (GN, [1, 2]) model of the process of knowledge discovery in databases is presented on Fig. 1. It is built of six transitions and seventeen places, with the transitions describing the processes listed below.

Initially, the α- and β-tokens stay in places l_3 and l_{16}. They will be in their own places during the whole time during which the GN functions. While they may split into two tokens, the original token will remain in its own place the whole time. The original tokens have the following initial and current characteristics:

- token α: "*Data warehouse*" (in place l_3),
- token β: "*Systems for decision making*" (in place l_{16}).

Below we shall omit these characteristics in descriptions of the separate transitions. If v is one of these tokens that can be split, then the new tokens will be noted by v', v'' and so on.

New α-tokens enter the net from place l_1 with characteristic "*New data*". New β-tokens will enter the net via place l_{14} with characteristic "*New system for decision making*".

All tokens that enter transitions Z_1 and Z_6 will unite with the corresponding original token. All information generated by the respective subject (Data warehouse, Systems for decision making) will be put as an initial characteristic of a token, generated by the respective original token.

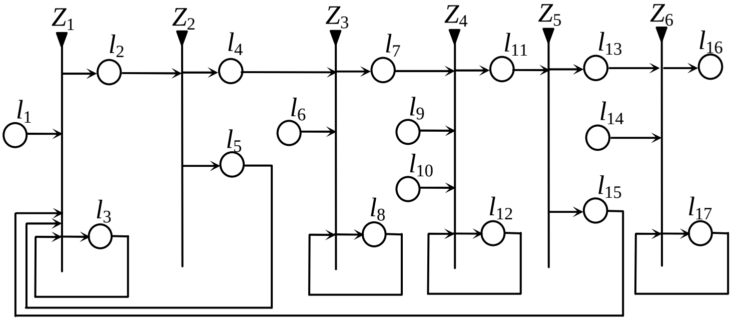

Fig. 1 Generalized Net Model of Knowledge Discovery in Medical Databases

Transition Z_1 represents the choice of data from the warehouse, which are relevant to the analysis. On this stage of the process, the collection of data from various sources is prepared, as well as the choice of the training data sample. For this reason, well developed tools for access to various data sources are needed here.

$$Z_1 = <\{l_1, l_3, l_5, l_{15}\}, \{l_2, l_3\}, \begin{array}{c|cc} & l_2 & l_3 \\ \hline l_1 & \mathit{false} & \mathit{true} \\ l_3 & W_{3,2} & \mathit{true} \\ l_5 & \mathit{false} & \mathit{true} \\ l_{15} & \mathit{false} & \mathit{true} \end{array}>,$$

where $W_{3,2}$ = "*Data are extracted from the warehouse*".

The characteristic of the original α-token has been described above. In some moment of time this α-token splits into two tokens. Once of them, namely the original α-token, remains in its initial place, and the new α-token enters place l_2 with the characteristic "*Data, extracted from the warehouse*".

Transition Z_2 represents the processing and cleaning of data retrieved from the data warehouse. These data may exhibit lapses, noise, anomalies, they may be inappropriate, insufficient, unusable, etc. Some tasks require for supplementing the data with some a priori information. However, with respect to the utilized method for data mining, data have to be correct and qualitative. Sometimes, this implies reduction of the output space, utilizing special algorithms.

$$Z_2 = <\{l_2\}, \{l_4, l_5\}, \begin{array}{c|cc} & l_4 & l_5 \\ \hline l_2 & W_{2,4} & W_{2,5} \end{array}>,$$

where $W_{2,4}$ = "*Data are processed*" and $W_{2,5} = \neg W_{2,4}$.
The α-tokens that enter respectively places l_4 and l_5 obtain characteristics: "*Processed data*" and "*Query for additional extraction of data from the warehouse*".

Transition Z_3 represents the process of transformation or normalization of data. Data are converted to a format that is adequate for the consecutive analysis. Procedures like data type unification, quantification, etc., take place. Some methods for analysis demand that output data are presented in a strictly predefined way. For instance, neural networks work only with numerical data, hence data have to be normalized.

A φ-token enters the model via place l_6, with the characteristic: "*Conditions for data transformation*".

$$Z_3 = <\{l_4, l_6, l_8\}, \{l_7, l_8\}, \begin{array}{c|cc} & l_7 & l_8 \\ \hline l_4 & \mathit{false} & \mathit{true} \\ l_6 & \mathit{false} & \mathit{true} \\ l_8 & W_{8,7} & \mathit{true} \end{array}>,$$

where $W_{8,7}$ = "*Data are transformed*".
The α- and φ-tokens that enter place l_8 do not obtain any new characteristics.

On the next step of model functioning, a new α-token enters place l_7 (from place l_8), and obtains the characteristic "*Transformed data*".

Transition Z_4 describes the process of data mining that comprises various techniques and tools for discovery of hidden patterns in databases [3], like neural networks, decision trees, clustering algorithms, associative rules, etc. The patterns discovered are evaluated by various coefficients for measuring their applicability, and as a result of this evaluation they are determined as knowledge, or not.

Via places l_9 and l_{10} respectively, γ- and δ-tokens enter the net with characteristics: "*Data mining tools*" and "*Target of the process of data mining*".

$$Z_4 = <\{l_7, l_9, l_{10}, l_{12}\}, \{l_{11}, l_{12}\}, \begin{array}{c|cc} & l_{11} & l_{12} \\ \hline l_7 & \mathit{false} & \mathit{true} \\ l_8 & \mathit{false} & \mathit{true} \\ l_{10} & \mathit{false} & \mathit{true} \\ l_{12} & W_{12,11} & \mathit{true} \end{array}>,$$

where $W_{12,11}$ = “*Data models (patterns) are retrieved*”.
The α-, γ- and δ-tokens entering place l_{12} do not obtain any new characteristics.

On the subsequent step of the model functioning, a new α-token enters place l_{11} (from place l_{12}), where it obtains the characteristic: *“Data models (patterns)”*.

Transition Z_5 presents the interpretation of the obtained results and the utilization of the derived knowledge in various applications. This step includes visualization of the discovered knowledge, which aims at facilitating the user to understand and interpret the obtained results.

$$Z_5 = <\{l_{11}\}, \{l_{13}, l_{14}\}, \begin{array}{c|cc} & l_{13} & l_{14} \\ \hline l_{11} & W_{11,13} & W_{11,14} \end{array}>,$$

where $W_{11,13} = W_{11,14}$ = “*Data are processed*”.

The α-tokens entering places l_{13} and l_{14} obtain the characteristic “*Knowledge*” .

$$Z_6 = <\{l_{13}, l_{14}, l_{17}\}, \{l_{16}, l_{17}\}, \begin{array}{c|cc} & l_{16} & l_{17} \\ \hline l_{13} & \mathit{false} & \mathit{true} \\ l_{14} & \mathit{false} & \mathit{true} \\ l_{17} & W_{17,16} & \mathit{true} \end{array}>,$$

where $W_{17,16}$ = *“Knowledge based solution is obtained”*.

The α-token entering place l_{16} obtains the characteristic *“Knowledge based solution”*.

3. APPLICATION OF THE MODEL TO MEDICAL DATABASES

Data warehouses and distributed federated databases store and manage the primary and derived biomedical data, such as genetic data. The proposed knowledge discovery GN model can be used for representing cases of biodata analysis such as distributed bio-medical databases, genome databases and proteome databases. Some interesting applications in this direction are illustrated as follows:

- One of the most important search problems in bio-data analysis is similarity search and comparison among bio-sequences and structures. For example, gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the different classes of genes.
- Identification of correlated bio-sequences or other correlated patterns. Most diseases are not triggered by a single gene but by a combination of genes acting together. Association and correlation analysis can be used to help determine the kinds of genes or proteins that are likely to co-occur in target samples.
- Path analysis; linking proteins to different stages of disease development. While a group of proteins may contribute to a disease, different proteins may become active at different stages of the disease.
- Visual data mining. Complex structures and sequencing patterns of genes and proteins are most effectively presented in graphs, trees, cubes. Such visually structures and patterns facilitate pattern understanding, and interactive data exploration. GN models for visual data mining therefore will play an important role in biomedical data mining.

In future it will be important to develop new GN models for scalable data mining methods for expandable and effective bio-data analysis. We believe that the active interactions and collaborations between these two fields have just started and a lot of exciting results will appear in the near future.

REFERENCES

1. Atanassov K., Generalized Nets, World Scientific, Singapore, 1991.
2. Atanassov K., On Generalized Nets Theory, Prof. M. Drinov Academic Publ. House, Sofia, 2007.
3. Chattamvelli R., Data Mining Methods, Narosa Book Distributors, Pvt, Ltd, 2008.
4. Cios K., W. Pedrycz, R. Swiniarski, L. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer, 2007.
5. Fayyad U. M., G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT, 1996.
6. Sumathi S., S. N. Sivanandam, Introduction to Data Mining Principles and its Applications, No. 29 in series *Studies in Computational Intelligence*, Springer, 2006.
7. Witten I., E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, 2005.