Rough Set Based Feature Extraction for Medical Data

Dominik Ślęzak, Piotr Synak, Jakub Wróblewski
Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland

Abstract

Rough-set-based KDD tools are applied to analysis of PKDD'2001 Discovery Challenge data set `thrombosis`. We focus on the phase of the feature extraction, aiming at finding new attributes which enable better classification of new cases.

1 Introduction

KDD in medical domain is very promising and potentially useful, but difficult. Medical data are often in the form of relational database with useful information distributed over many tables. Many features are time-dependent. There is a wide spectrum of medical examinations available and – in most cases – only a few of them are actually needed (and applied) to a patient. Creation of a data table, containing a limited set of features, usually needs advanced preprocessing. Final information system consists of many null values and distribution of decision classes is often very unbalanced.

PKDD'2001 Discovery Challenge data set `thrombosis` [9] is an example of relational database in medical domain. The crucial point in the analysis of such a type of data is to select the most appropriate features, enabling construction of effective and understandable rules applicable to forecasting the development of disease. One can apply various methods of the new feature extraction (cf. [4]). In this paper we focus on rough-set-based techniques, since – in our opinion – they provide medical expert with the most intuitive description of the extraction process [12, 15, 16].

In case of the `thrombosis` data set one can distinguish several possibilities of new feature construction. First one assumes that the decision problem is formulated within a single decision table. Then one could apply some standard algorithms for constructing classification models (see e.g. [2]). In [12] we proved for empirical data that such techniques are worth additional support, e.g., by applying very simple and intuitive method for automatic extraction of new features from data, as linear combinations of conditions (see also [13, 17]).

The second possibility relates to the multi-table architecture of data. One can use expert knowledge to create new aggregation features which seem to have an influence on concepts we want to approximate. Procedure of adding such features to a decision table is based on relations between tables. One can combine a manual, SQL-based extraction method with a framework developed in [16], enabling automated way of searching for potentially most appropriate aggregation attributes within distributed data environment.

Finally, since the `thrombosis` data set includes some time-related aspects, we also deal with temporal analysis. To obtain self-contained framework for the feature extraction, we focus on rough-set-based methods developed in [15]. We propose to search for temporal patterns that can potentially be specific for the

occurrence of disease and its development. If there exist such patterns, they can be used as new descriptors. Thus, new time-related features can be created.

It is worth noting that mentioned approaches are based not only on appropriate usage of already given attributes but also on searching for completely new means of expression. Presented algorithms optimize quality measures which have strong rough-set-based theoretical background. New attributes are applicable not only to rough-set-based methods. We obtain possibility of an intelligent preprocessing of data information rather than a final classification system.

2 Reasoning with data

One of the main goals of data analysis is to properly classify objects (described by data) to some classes. Objects (measurements, observations, records) are usually described by some features (attributes) and divided into a number of decision classes (in case of training data the decision is given as an additional attribute). Reasoning with data can be stated as a classification problem, concerning prediction of values of a decision attribute under information provided by conditional attributes. For this purpose, one stores data within decision tables, where each training case drops into one of predefined decision classes. The aim of data analysis is to construct a classifier – an algorithm which is able to classify a previously unseen object into proper decision class.

A data set can be defined as a decision table $\mathbb{A} = (U, A \cup \{d\})$, where each attribute $a \in A$ is identified with a function $a : U \to V_a$ from the universe of objects U into the set V_a of all possible values of a and d defines partition of U to mutually disjoint decision classes.

In the most practical cases a decision table is not given in advance and should be constructed in a preprocessing stage of the classifier construction process. The decision table is then modeled by classification algorithms (classifiers). Methods for construction of classifiers can be regarded as tools for data generalization. These methods include rule-based classifiers (e.g. rough-set-based [8]), decision trees, k-NN classifiers, neural nets etc.

3 Experimental medical data

PKDD'2001 Discovery Challenge data set `thrombosis` [9] is an example of relational database in medical domain. Figure 1 presents a diagram of tables (characterized by number of objects and names of attributes) and relations.

Thrombosis is an important and severe complication in collagen diseases, and one of the major causes of death [9]. It relates to an increased coagulation of blood which clogs blood vessels. Usually its attacks last several hours and can repeat. It is very important to predict the possibility of the attack occurrence. One of the main problems concerned with `thrombosis` database is thus to capture temporal patterns specific and sensitive to attacks.

The database contains the following information:

1. Patient's personal data (sex, birthdate).
2. Diagnoses and diseases characteristics.
3. Thrombosis attacks (date, symptoms).
4. Results of laboratory examinations.

Temporal patters can involve information gathered in tables ANTIBODY_EXAM, ANA_PATTERN and LAB_EXAM. The first two of them are devoted to relatively sophisticated antibody tests, which are usually applied to patients with the thrombosis attacks already registered. This type of information enables to correlate thrombosis with other collagen diseases and thus – to gather more knowledge about its characteristics. It corresponds to a slightly different problem, related to prediction of occurrence and correlation of occurrence of particular diseases registered for particular patients in DIAGNOSIS table. This table is shaded in Figure 1 to mark that information gathered

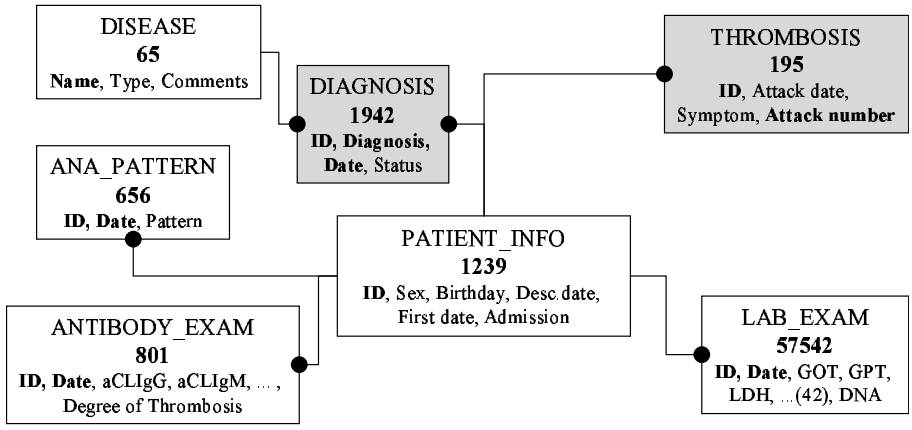

Figure 1: Medical database: `thrombosis`. Tables used to define decision are shaded.

within it can be used directly to formulate decision classes – the target concepts which should be approximated in terms of the other features, based possibly on information from the other tables.

In order to obtain a complete framework for the attack prediction, LAB_EXAM table seems to be of the greatest importance. It consists of more regularly stored data about standard laboratory examinations of particular patients and thus provides better source for the training process. On the other hand, information provided by those examinations is assumed to be at least partially redundant with respect to thrombosis occurrence. Thus, one can apply appropriate reduction and transformation procedures to get the most relevant subspace of features describing patients in time.

One can combine information from LAB_EXAM and THROMBOSIS tables to construct a single data table consisting of records corresponding to examination/attack characteristics of particular patients along particular days. Such a table may be defined in terms of the following features:

- Results of particular laboratory examinations for a particular patient during particular day, as well as during a number of days in the past.
- Information about occurrence of thrombosis attacks in the past, as well as information about attacks in the nearest future (related to the time stamp of a particular patient/day record).

Given such features, one can distinguish the attribute corresponding to intensity of the future attacks as the decision and try to approximate particular decision classes by using the rules involving information from the past. Just like in case of DIAGNOSIS table described before, THROMBOSIS is shaded as well to distinguish it as a straightforward source of decision classes definition. One should remember that although it is the source of definition of decision classes, the calculations should run over a completely new decision table, with records corresponding to patients in time, and with features corresponding to both attacks and examinations.

One can also enrich the list of features with columns corresponding to occurrence – for particular patients in particular time – of:

- Previously derived, complex patterns of examination results, potentially correlated to occurrence of thrombosis attacks in the nearest future – Such patterns, possibly involving trends of examination results in time, provide new attributes being transformations of original data vectors (sequences of vectors in time) into more relevant dimensions.
- Non-temporal features concerned with particular patients, like those from PA-

TIENT_INFO table, or those from DIAGNOSIS table, if treated just in terms of discovering particular diseases for particular patients – it may happen that for particular clusters of patients one can extract different temporal patterns. They can achieve characteristics varying with respect to age or gender. One can even consider separate forecasting temporal patterns for distinguished groups of patients with various levels of the illness intensity (obviously, in such a case, we would have to be able to classify patients into particular groups of intensity first).

4 Automatic feature extraction

Many data mining tools operate on information systems – simple data representations, consisting of one table of objects described by fixed number of attributes (numeric or symbolic). A medical database like `thrombosis` should be prepared for further analysis by transformation of its original structure. All useful information hidden in temporal coincidences and trends, distributed in many tables bounded by relations, should be encoded into a set of new features, often defined over a completely new data table, with semantics of records differing from all original tables. Such a transformation (preprocessing) in most cases is being performed manually and requires knowledge of an expert.

Our idea of construction and selection of new features is based on *wrapper approach* [4]: a result (transformation) found by algorithm is then evaluated basing on its potential usefulness in classification of new cases. The process of the new feature optimization is adaptive and leads to more efficient set of attributes. In [16] and [17] a set of tools for selecting and evaluating new attributes derived from the set of tables and relations is presented. Experimental results presented in [16] (financial domain) and [14] (multimedia data) suggest, that these methods can be successfully used also on `thrombosis` database.

Let us consider the task of thrombosis occurrence prediction. To describe the problem in terms of decision table (information system), one should define a notion of object (record) and decision values. As stated in the previous section, decision attribute can be derived from THROMBOSIS table and refers not to a patient itself, but to a patient-in-specific-time-moment. The latter should be regarded as a definition of an object (record). The same patient may be treated as positive example (when observed in the day before thrombosis occurrence) as well as negative one (when no thrombosis occurs in the several next days), constituting more than one object in final decision table. Patient's data take the form of relational database containing time series (results of examinations etc.) and other information distributed in several tables. Due to the framework presented in [16] we start with the *base table* containing only straightforward, static information about patient (e.g. age, gender) and try to extend it with new attributes (defined as aggregation operations on various data tables) found by adaptive algorithm.

Efficiency of classifier based on a given set of attributes depends not only on domain-dependent information provided by its values, but also on its granularity, i.e. level of data generalization. Proper granularity of attributes' values depends on knowledge representation and generalization techniques used in classification algorithm. In case of numerical features, such techniques as discretization, hyperplanes, clustering, and principle component analysis (see e.g. [4]), are used to transform the original domains into more general or more descriptive ones. One can treat the analysis process over transformed data either as a modeling of a new data table (extended by new attributes given as a function of original ones) or, equivalently, as an extension of model language. The latter means, e.g., change of metric definition in k-NN algorithm or extension of language of rules or templates.

In our approach the original data set is extended by a number of new attributes defined as a linear combination of existing ones. Let $B = b_1, ..., b_m \subseteq A$ be a subset of attributes, $|B| = m$, and let $\alpha = (\alpha_1, ..., \alpha_m) \in \mathbf{R}^m$ will

be a vector of coefficients. Let $h : U \to \mathbf{R}$ be a function defined as $h(u) = \alpha_1 b_1(u) + ... + \alpha_m b_m(u)$. Usefulness of new attribute defined as $\overline{a}(u) = h(u)$ depends on proper selection of parameters B and α. The new attribute $\overline{a}$ is useful, when the model of data (e.g. decision rules) based on discretized values of $\overline{a}$ becomes more general (without loss of accuracy). Evolution strategy algorithm optimizes $\overline{a}$ using quality function based on intuition that a model with lower number of (consistent) decision rules is better than the others (cf. [2], [8]). For further details refer to [12].

5 Time-related features

Most of medical applications contain data, where time domain plays an important role. That includes patient examinations, drug taking, disease history, etc. Taking advantage of time dependencies seems to be a key problem. In our approach we propose to search for temporal patterns in time related data. Presence of one temporal pattern and absence of another one can be specific for given type of disease. In the same way a sequence of temporal patterns occurring in data can be disease specific. Though, temporal patterns or their sequences can be widely used as new attributes.

In case of `thrombosis` database we propose to search for temporal patterns in LAB_EXAM table. It contains information about history of performing 42 tests on patients. One patient is usually examined several times - data contain information about date of examination.

We propose to create a number of information systems, each containing results of tests performed before attack of thrombosis. For example, we can create k=2 systems $\mathbb{A}_0$, $\mathbb{A}_{-1}$, where $\mathbb{A}_0$ ($\mathbb{A}_{-1}$) would contain results of last (one before last) tests performed before attack of thrombosis. Thus, $\mathbb{A}_0$ ($\mathbb{A}_{-1}$) would have number of rows equal to number of all attacks of thrombosis; and as many columns as tests. Moreover, in each system we propose to add additional attributes specifying degree of change of each test from previous one. In such data tables we can search for regularities, that can be next used for creation of temporal patterns.

A regularity can be understood as *template* T, i.e. set of *descriptors* ($a \in V$):

$$T = \{(a \in V) : a \in A' \subseteq A,\, V \subseteq V_a\},$$

The notion of template was intensively studied in literature (see e.g. [1], [6], [15]).

Having found templates of high quality for some of k information systems we can construct temporal patterns over them as combinations preserving time order. Time order is determined by sequence of systems (responding to historical results of tests before attack of thrombosis).

For example, suppose we found template $T_1 = \{(test2 \in [0,5]), (test4 \in \{high\})\}$ in $\mathbb{A}_0$ and $T_2 = \{(test1 \in [1,2]), (test2 \in \{medium\}), (\Delta test3 \in \{high\})\}$ in $\mathbb{A}_{-2}$. We can construct temporal pattern $< (T_2, -2), (T_1, 0) >$, which can be interpreted as "results of last examination match descriptors of T_1 and results of second before last examination match descriptors of T_2".

Constructed temporal patterns can be used as new binary attributes - for a given patient-in-time object (see Section 4) we set value of such attribute to *true*, if results of tests performed so far match responding temporal pattern.

6 Conclusions

We discussed possibilities of application of rough-set-based KDD tools to analysis of PKDD'2001 `thrombosis` data. We focused on the phase of the feature extraction, aiming at finding new attributes which enable better classification of new cases.

One can formulate a number of tasks concerned with the analysis of `thrombosis` data. One of the most important challenges is to extract an efficient model for forecasting the occurrence of the thrombosis attacks, by basing on observations of the recent results of laboratory examinations for particular patients. We

proposed a framework for creation of a new decision table, with records corresponding to information about particular patients in particular days. We described some methods for automatic searching for potentially relevant features of such records.

Acknowledgements

Supported by Polish National Committee for Scientific Research (KBN) in the form of PJIIT Project No. 1/2001 and KBN grant No. $8T11C02417$.

References

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, I.: Fast Discovery of Association Rules, *Proc. of the Advances in Knowledge Discovery and Data Mining.* AAAI Press/The MIT Press, CA (1996) pp. 307–328.

[2] Bazan, J.G., Nguyen, H.S., Nguyen, S.H, Synak, P., Wróblewski, J.: Rough Set Algorithms in Classification Problem. In: Polkowski, L., Tsumoto, S., Lin, T.Y. (eds), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Physica-Verlag (2000) pp. 49–88.

[3] Berthold, M., Hand, D.J.: Intelligent Data Analysis. An Introduction. Springer-Verlag (1999).

[4] Liu, H., Motoda, H. (eds): Feature extraction, construction and selection – a data mining perspective. Kluwer Academic Publishers, Dordrecht (1998).

[5] Mannila, H., Toivonen, H.,Verkamo, A.I.: Discovery of frequent episodes in event sequences. *Report C-1997-15*, University of Helsinki, Finland (1997).

[6] Nguyen, H.S.: Discretization od Real Value Attributes: Boolean Reasoning Approach. Ph.D. thesis. Institute of Mathematics, Warsaw Univ., Poland (1997).

[7] Nguyen, S.H.: Regularity analysis and its applications in data mining. Ph.D. thesis. Institute of Informatics, Warsaw University, Poland (1999).

[8] Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991).

[9] PKDD'2001 Discovery Challenge homepage (data and description), http://lisp.vse.cz/challenge/pkdd2001/

[10] Polkowski, L. , Skowron, A. (eds): Proc. of The First International Conference on Rough Sets & Current Trends in Computing (RSCTC'98). Springer-Verlag, Berlin, Heidelberg (1998).

[11] Polkowski, L., Skowron, A. (eds): Rough Sets in Knowledge Discovery vol. **1**, **2**. Physica-Verlag, Heidelberg (1998).

[12] Ślęzak, D., Wróblewski, J.: Classification algorithms based on linear combinations of features. In: Proc. of PKDD'99. Praga, Czech Republik, LNAI 1704, Springer, Heidelberg (1999) pp. 548–553.

[13] Ślęzak, D.: Approximate decision reducts (In Polish). Ph.D. thesis. Institute of Mathematics, Warsaw University, Poland (2001).

[14] Ślęzak, D., Synak, P., Wieczorkowska, A., Wróblewski, J.: KDD-based approach to musical instrument sound recognition. Accepted to ISMIS'02, Lyon, France (2002).

[15] Synak, P.: Temporal templates and analysis of time related data. In: Ziarko, W. , Yao, Y.Y. (eds), Proc. of The Second International Conference on Rough Sets & Current Trends in Computing (RSCTC'00), Banff, Canada (2000).

[16] Wróblewski, J.: Analyzing relational databases using rough set based methods. In: Proc. of IPMU'00. Madrid, Spain (2000) 1, pp. 256–262.

[17] Wróblewski, J.: Adaptive methods of the object classification (In Polish). Ph.D. thesis. Institute of Mathematics, Warsaw University, Poland (2001).