STUDIA INFORMATICA 2012
Volume 33 Number 2A (105)

Marcin MAZUREK
Wojskowa Akademia Techniczna, Instytut Systemów Informatycznych

PATTERNS OF MULTIRELATIONAL DATA TRANSFORMATION IN DATA MINING PROCESS

Summary. Multirelational data mining requires complex preprocessing of data. Identification of transformation patterns and implementation of reusable components lead to more robust data-mining flow construction process. In this paper concept of implementation of selected transformation patterns is presented. Rapid Miner environment is used to build transformations, which can be later used in predicting customer behavior.

Keywords: muti-relational data mining, time-series, propositionalization

WZORCE TRANSFORMACJI DANYCH WIELORELACYJNYCH W PROCESIE EKSPLORACJI DANYCH

Streszczenie. Eksploracja danych wielorelacyjnych w dostępnych środowiskach eksploracji danych wymaga złożonego wstępnego przetwarzania danych. Identyfikacja wzorców przetwarzania oraz ich implementacja w postaci komponentów wielokrotnego użytku prowadzi do zwiększenia efektywności konstrukcji przepływów danych. W artykule przedstawiono koncepcję implementacji w środowisku Rapid Miner wybranych transformacji, które znajdują zastosowanie w prognozowaniu zachowań klientów.

Słowa kluczowe: eksploracja danych wielorelacyjnych, szeregi czasowe, propozycjonalizacja

1. Introduction

1.1. Multirelational data mining

Data mining is integral part of data warehouse reporting software suites. Data mining allows discovering patterns in data, hidden from user because of multidimensionality and volume of data. Classification models, based on logistic regression, neural networks and decision trees are successfully deployed in business to predict customer behavior, scoring, fraud detection and many others applications.

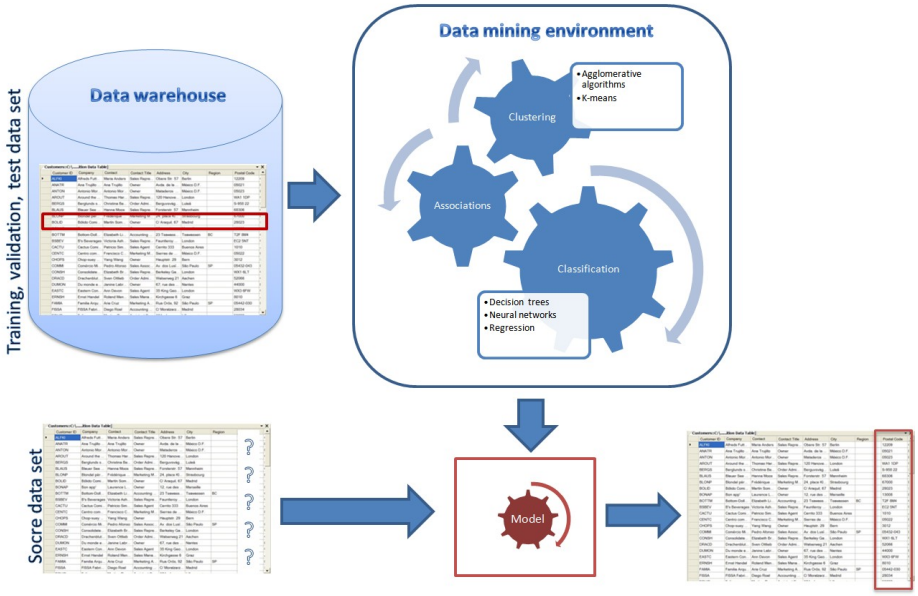

Fig. 1. Data flow in data mining environment
Rys. 1. Przepływ danych w środowisku eksploracji danych

The architecture of data mining tools is presented on Fig. 1. Data mining algorithms develop classification or predictive models from training data set, which contains data objects with assigned values of target variable. Target variable represents class identifier (in classifications task) or interval value (in prediction task). The output from learning process is a model, which can be in the form of decision tree, regression equation, trained neural network, Bayesian classifier or other. This model is then used to score new data set, which has the same structure as training data set, but with null target variable column.

Existing data mining tools require that all information about object, which is subject to analysis both in learning and scoring data set, is saved in exactly one row of the database table. Input database schema consists of only one table.

This attribute-value learners approach seems to be inadequate since it ignores the fact, that in most cases data objects are related to other entities in the database. Types of relations and attributes of associated objects are additional input to the knowledge discovery process, which is omitted in classical data mining approach.

Current practices to overcome this limitation rely mostly on extending object attribute list stored in target table. The additional attributes should be "a summary" of the knowledge contained in the remaining database tables. However, in many cases adding extra attributes is done in an intuitive way rather than as a result of systematic approach.

Data mining, which recognizes the structure of input observations and possible associations, is described as structured data mining or multirelational data mining [2][3]. Prefix "multi-" differentiates from relational data mining, which describes single-table algorithms that work on data stored in relational data table.

As each relation between tables in database schema may represent different business relationships, there is no universal rule, applicable to all schema, how to transform data. Nevertheless, one can identify most common data object dependency cases and propose method to analyze the data.

The problem of the multirelational data mining can be formulated as follows: given the relational data schema with table storing target objects, build such a predictive model, that calculates target variable for each object or classifies objects into predefined groups, taking into account both target object attributes and another database objects associated with the target object.

2. Propositional approach

Because of highly varied business logic there are different methods dealing with the problem of multiple relations in the data mining tasks, integrating techniques from graph modeling, logic programming and machine learning [4]. Basically, there are two approaches to the problem:

- transformation of data to single-table schema (propositionalization) and application of traditional data mining methods,
- utilization of multi-table representation of data by inductive logic programming, multirelational decision trees, graph mining methods and others. They are not implemented in available commercial and open-source data mining suites and thus are beyond the scope of the paper.

The first approach transforms variable number of records into fixed-length attribute list [5, 6]. After child record information is captured in a fixed-length attribute list, commercial off the shelf tools (COTS) can be used.

Basically there are two possible approaches to transformation of multi-table data into single relation schema with fixed set of attributes:

- join of all tables resulting in universal relation,
- transformation by creating new attributes, that summarizes or aggregates information from other tables.

The output of the first transformation can be extremely large and in practice impossible to process. It results in multi-instance data mining problem.

The latter technique requires extensive domain understanding to propose relevant attributes. There is no single algorithm for propositionalization, because of the semantical variety of data stored in non-target tables. If data is distributed between more than two tables, and each subsequent table is joint by one-to-many relationship, aggregation can be applied recursively.

3. Problem statement

3.1. Data mining task

The primary goal of the data mining process considered in this paper is classification of the customers according to the odds of performing particular action or their transition to particular state. Examples of such actions are churn, bankruptcy, breaking the contract, funds withdrawal. They result in reductions of revenue. Model should be developed on training data coming from relational data warehouse, which holds all the data from customer relationship management system.

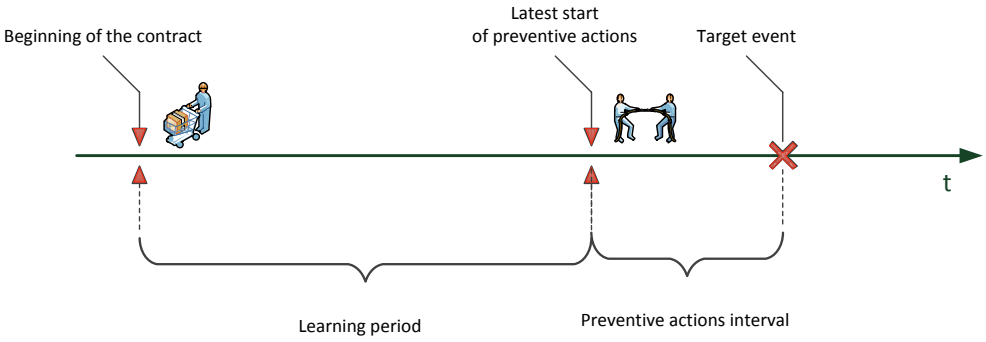

Fig. 2. Timing of events in customer contract history
Rys. 2. Umiejscowienie zdarzeń w historii kontraktu

There is a limited period of time, when any of the preventive actions can be successfully undertaken. After some particular moment in customer relationship history, corrective actions is economically unjustified, even when it is clear that customer falls into target group (see Fig. 2). This has implications on selection of the training data set records, which should be limited to the records gathered during the learning period.

3.2. Explanatory data

Explanatory analysis and classification models, built for solving the problem presented above, should have possibility to handle following input data schemas:

- static, practically not changing over time social and demographical characteristics of the customer,
- time series data,
- aggregation schema,
- associations,
- external, unrelated directly time-series and events.

Data mining models, widely employed by business analysts to predict odds of such event for individual customer, rely on static features. Such features can be easily stored in single table. Classification models can be directly developed on such data sets.

Additional, and in some businesses the only one, data about the customer is data collected by the CRM systems, describing his contacts, activities and orders since the moment he signs the contract. Analyzing the history of customers' actions in the period preceding occurrence of target event, can boost predictive power of the model. Recorded events form univariate or multivariate time-series.

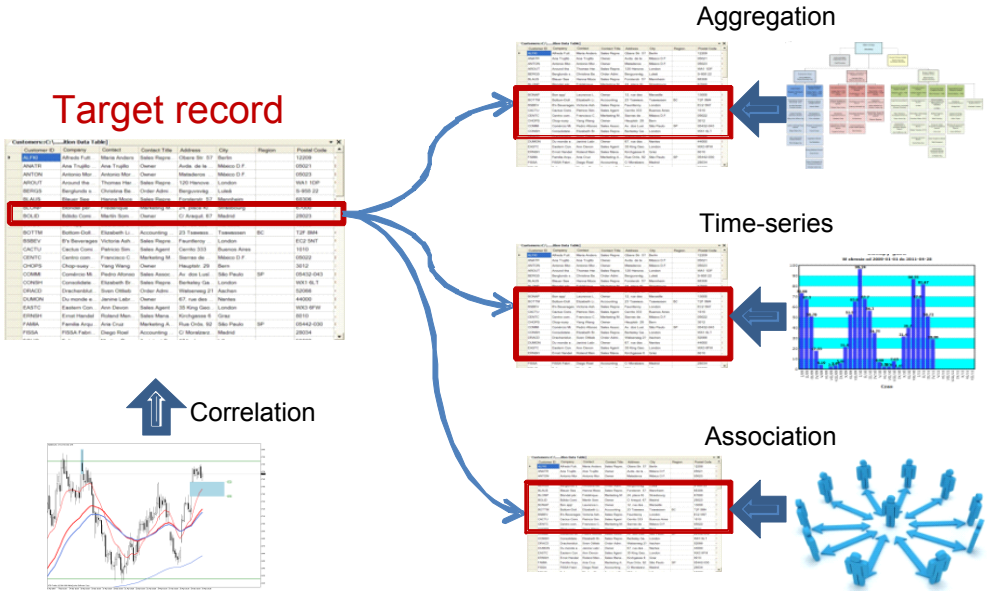

Fig. 3. Multirelational data mining schema
Rys. 3. Wielorelacyjny schemat danych w procesie eksploracji danych

Another information, often omitted in data mining processes, is staying in relation with or being associated to another entities. The difficulty comes from necessity of transformation of such data to the single-relation table, which is not supported by data mining suites. The particular form of such relation is aggregation. Basically, list of the possible relations between data objects includes, but is not limited to relations mentioned above and presented in Fig. 3.

In the schema we distinguish one table, called target relation, containing unique representation of objects (being instances of target class) to be classified. Data objects in the remaining tables are related on technical level to the target objects by means of foreign key relationship, which is a transitive relation. Semantics of such constraint is an additional knowledge, which might be utilized in data mining processes.

4. Data transformation patterns

In order to efficiently build classification models, some of the most common processing patterns were identified. The aim of such identifications is implementation of higher-level building blocks of data mining process, which can be used to preprocess multirelational data. Selected high-level multirelational data and associated preprocessing patterns are presented. On implementation level, they are further decomposed according to the representation of the data and data mining objective.

4.1. Dimensionless series of events

This pattern is applicable to information about events, which have no dimension associated with them. Only astronomical time of event is recorded – because of the nature of the event or because no more attributes can be recorded or it is economically justified.
Examples of such event are: logging to Internet application, sending SMS from mobile phone. Preprocessing of such data includes:

- relating the absolute time stamp associated with the event to the start of the time scale for the customer,
- outlier detection and removing,
- clustering series according to regularity and number of event occurrences [7].

Outcome statistics from feature extraction subprocess include average, minimum, maximum, standard deviation, event count. Such statistics are subject to clusterization, which leads to further reduction of dimensionality.

4.2. Time-series data

Time series data mining is one of the most active research areas [1, 12]. Among key problems of time-series data mining a segmentation method of data mining series is widely recognized as particularly essential [9]. Clustering time-series data might be seen as a technique of propositionalization, in which the whole sequence is represented by the cluster centroid.

The classes of time-series data can be any of the following:

- univariate time-series with irregular time points – example of such series are money withdrawal events, phone calls (numerical time-series), contacts with different customer service channels (symbolic time-series),
- univariate time-series (equal time intervals) – example of such time series is account balance or assets level,
- multivariate time-series with irregular time points – example is an investment allocation order,
- multivariate time-series (equal time intervals) – example is changing share of different assets in the portfolio

Unfortunately, most of the researches concentrate on fixed-length, uniformly sampled data. Real life records are seldom fulfilling that assumption. Usually they are characterized by:

- non uniformly sample size,
- varying length of the covered timespan.

The outcome feature vector for value series may include:

- statistics of the time series values (average, min, max, standard deviation),
- trend parameters (depending on the timespan for the whole period or fixed numbers of non-overlapping time-frames),
- length of the last trend direction,
- last change point and total number of change points throughout the period.

4.3. Unrelated time-series

This form of time series is a distinctive pattern, as events are unrelated to the object, but may affect behavior of the customer. Examples are the stock market index values, currency exchange ratio.

The task for this kind of data is to detect time-correlation between time-series data streams or between time stream and customer-generated event stream [8]. For correlation between time-series data the following approach can be applied [11]:

- summarizing data in fixed granularity level,
- detecting change points (CUMSUM – Cumulative Sum method can be applied here),
- calculating the statistical correlation between aggregated data points and corresponding time-distance.

4.4. One-to-many associations

Although one-to-many relationship is practically the only kind of relationship between the tables in RDBS, in this context one-to-many association means relation, where target object is connected by the foreign keys with set of other objects, and both the connection and attributes of child objects are changing so slowly, it can be assumed they are constant. The linked records can represent possession of the customer (SIM Cards, accounts, houses etc.), elements of social network (linked friends), purchased products.

The most obvious transformation of such object is summarization of the objects linked, grouped by the child object type.

To reflect the difference in information gain associated with connected objects, optionally aggregated vector of TF-IDF can be calculated [10], similarly as in representation of text documents.

5. Pattern implementation in Rapid Miner environment

Identified transformation patterns can be used for construction software components used in data mining processes. Some of the patterns mentioned above have already been implemented as a process templates in Rapid Miner[1]. This open-source Java environment enables defining process templates, which can be later used as building blocks in higher level processes. The platform was selected mainly because of its openness (possibility to invoke models implemented in other open-source tools: R[2] and WEKA[3]) and its user-friendly interface appealing to business users.

The key features of the solutions, utilized in pattern implementation, are as follows:

- macro-variables and evaluation of expressions defined with macro-variables,
- automatic optimization of model parameter values (i.e. selection of number of clusters based on value of Davies-Bouldin index),
- openness of the platform with possibility to integrate with other open-source platforms (R, Weka).

The implemented components were used in data mining process aiming at classification of customers according to their probability of breaking the contract in asset management company, where practically all available knowledge of the customer is limited to behavioral, time-stamped data describing his orders and contacts.

[1] Rapid Miner homepage: http://rapid-i.com/content/view/181/190/.
[2] R Project of Statistcal Computing homepage: http://www.r-project.org/.
[3] WEKA homepage: http://www.cs.waikato.ac.nz/ml/weka/.

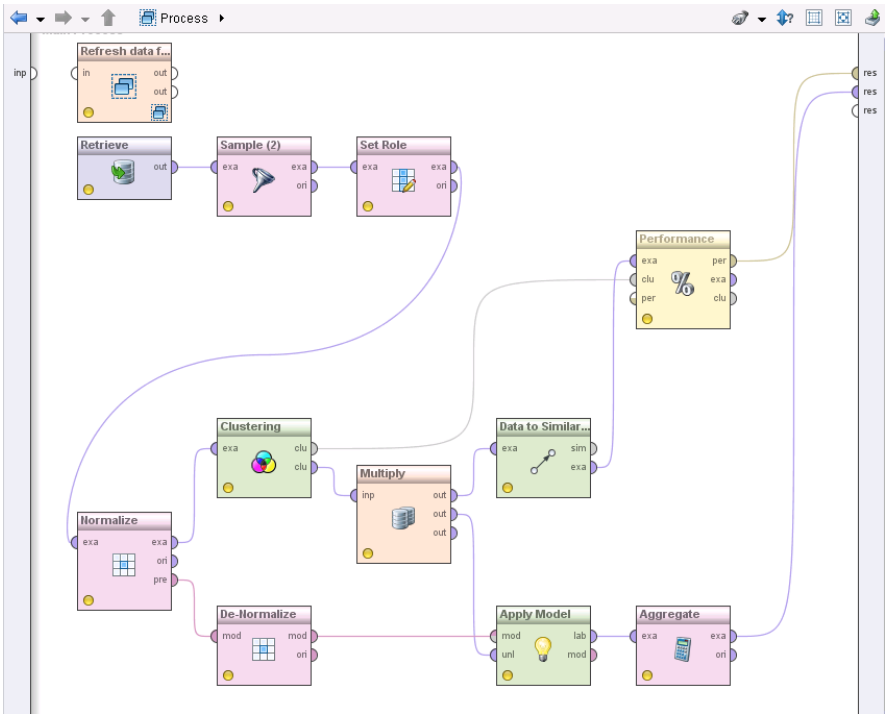

Fig. 4. Rapid Miner data mining sample template implementation
Rys. 4. Przykład implementacji procesu w środowisku Rapid Miner

6. Conclusions

Bulk of the data mining tools offers set of data transformation building blocks limited to elementary transformations corresponding to data mining algorithms. Constructing data mining data flow from such detailed transformation is time consuming and prone to errors.

Presented approach, offering high-level transformation subprocess templates reduces effort in building data mining processes. Thanks to ready-to-use components analyst can concentrate on major tasks and on matching the best transformation pattern to business problem.

Taking into account current limitation of data mining tools and growing demand for knowledge discovery in organizations in competitive markets, overcoming this limitations seems to be natural direction of evolution of Business Intelligence systems.

Identification of the patterns seems to be especially important in the area of multirelational data mining tasks. Because of the domination of relational databases in the data warehouses technology, multirelational data mining problems are the lion share of the business concerns, whereas the available tools provide hardly any support for solving it.

BIBLIOGRAPHY

1. Last M., Klein Y., Kandel A., Abraham K.: Knowledge Discovery in Time Series Databases. IEEE Transactions on Systems. Man and Cybernetics, Vol. 31, 2001, p. 160÷169.
2. Dzeroski S.: Multirelational data mining. An introduction. ACM SIGKDD Explorations Newsletter, Vol. 5, 2003, p. 1÷16.
3. Dzeroski S., Lavrac N.: Relational Data Mining. Springer, Berlin 2001.
4. Knobbe A.: Multi-Relational Data Mining. IOS Press, Amsterdam 2006.
5. Knobbe A., De Haas M., Siebes A.: Propositionalization and Aggregates. Proceeding of the 5th PKDD, 2001, p. 277÷288.
6. Krogel M. A., Wrobel S.: Transformation-Based Learning Using Multirelational Aggregation. LNAI, 2001, p. 142÷155.
7. Liao T. W.: Clustering of time series data – a survey. Pattern Recognition, Vol. 38, 2005, p. 1857÷1874.
8. Minaei-Bidgoli B., Lajevardi S. B.: Correlation Mining between Time Series Stream and Event Stream. Networked Computing and Advanced Information Management, 2008, p. 333÷338.
9. Morchen F.: Unsupervised pattern mining from symbolic temporal data. SIGKDD, 2006.
10. Rayner A.: Discovering Knowledge from Multi-relational Data Based on Information Retrieval Theory. LNAI, 2009, p. 409÷416.
11. Sayal M.: Detecting Time Correlations In Time Series Data Streams. HP Laboratories, 2004.
12. Tak-Chung F.: A review on time series data mining. Engineering Application of Artificial Intelligence, 2011, p. 164÷181.

Wpłynęło do Redakcji 13 stycznia 2012 r.

Omówienie

Reprezentacja danych o obiekcie w postaci rekordu pojedynczej tabeli powoduje pominięcie w procesie eksploracji danych dodatkowych informacji przechowywanych w innych tabelach relacyjnej hurtowni danych. Wykorzystanie tych dodatkowych danych jako predykatorów wymaga złożonego przetwarzania danych, które obecnie w niewielkim stopniu jest

wspomagane przez narzędzia eksploracji danych. Aby zwiększyć efektywność konstruowania przez analityków procesów eksploracji danych z wykorzystaniem danych wielorelacyjnych, środowisko powinno dostarczać gotowe, reużywalne bloki transformacji danych.

Pierwszym krokiem w tym kierunku jest identyfikacja wzorców wstępnego przetwarzania danych. Opierając się na danych zgromadzonych w hurtowni danych oraz celu procesu eksploracji danych, można wyróżnić wzorce przetwarzania danych związane z przetwarzaniem szeregów czasowych, asocjacji, agregacji oraz strumienia zdarzeń. Przetwarzanie tych danych prowadzi do propozycjonalizacji, czyli zastąpienia zbiorów danych o zmiennej liczbie rekordów rekordem o ustalonym zbiorze atrybutów. Wyznaczony zbiór atrybutów rozszerza wektor cech obiektu. Głównymi technikami propozycjonalizacji wykorzystywanymi we wzorcach przetwarzania danych są agregacja oraz klasteryzacja.

Dla wybranych wzorców zostały zaimplementowane podprocesy pełniące rolę bloków budowlanych w środowisku eksploracji danych RapidMiner. Wzorce procesów parametryzowane są za pomocą makrozmiennych, możliwość wielokrotnego używania dla różnych modeli danych wejściowych została natomiast osiągnięta za pomocą mechanizmu ewaluacji wyrażeń definiowanych za pomocą makrozmiennych.

Address

Marcin MAZUREK: Wojskowa Akademia Techniczna, Instytut Systemów Informatycznych, ul. Kaliskiego 2, 00-908 Warszawa, Polska, marcin.mazurek@wat.edu.pl.