Liet. matem. rink., T. 42, spec. nr., 621–629

Data mining technologies based on IBM intelligent miner

Regina KULVIETIENĖ, Jelena MAMČENKO (VGTU)
e-mail: regina_kulvietiene@gama.vtu.lt, jelena@gama.vtu.lt

Introduction

High competition level and companies desires to get much more gain from Information technology (IT) investment had induce new technology beginning. Intelligent mining technology could answer for different science, knowledge and business questions. Right choice of data analyzes and visualization tools could help analysts who are interested in to view and solve real problems. These technology systems are realizing new data analyzes form, which is supported by intellectual decisions. Using modern analyzing methods of Data Mining (DM) a lot of organizations are increasing their gain and manufacture, reducing outlays and extending clients satisfaction.

The application of these technologies without limits, it can be everywhere where data are. The main interest was taken by commercial enterprises that are making projects using Data Warehousing.

Useful information might be hidden in the data in the form of implicit patterns and connections that are not easy to discern using conventional data queries and statistical calculations.

DM is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data. All can use extracted information to form a prediction or classification model, or to identify similarities between database records. The resulting information can help you make more informed decisions.

1. Data mining process

The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various businesses, scientific, and governmental organizations around the world is daunting. According to information from research center, only scientific organizations store each day about 1 TB (!) of new information. And it is well known that academic world is by far not the leading supplier of new data. It becomes impossible for human analysts to cope with such overwhelming amounts of data.

While DM does not eliminate human participation in solving the task completely, it significantly simplifies the job and allows an analyst who is not a professional in statistic and programming to manage the process of extracting knowledge from data.

1.1. *Data mining*

Information technology has developed rapidly over the last three decades. Many organizations store increasingly large volumes of data on their computer systems. Useful information might be hidden in the data in the form of implicit patterns and connections that are not easy to discern using conventional data queries and statistical calculations.

DM is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data. All can use extracted information to form a prediction or classification model, or to identify similarities between database records. The resulting information can help you make more informed decisions.

The Intelligent Miner (IM) for supports a variety of data mining tasks. For example, a retail store might use the IM to identify groups of customers that are most likely that are most likely to respond to new products and services or to identify new opportunities for cross-selling. An insurance company might use the IM with claims data to isolate likely fraud indicators.

1.2. *Mining based on intelligent miner*

Converting Data Warehousing to information is a complex process can be successfully used to do decisions. Transforming the contents of a data warehouse into the information that can drive decision-making is a complex process that can be organized into four major steps.

- Data selection

To solve a problem we need only a part of data from information systems. At first, data must be selected to further analysing. Sometimes in this case we need to join several tables. Later given records are filtered.

- Data transformation

The first data transformation is implemented when tables are ready. Transformation is depending on methods which are using during analyse process.

- Mining the data

Using one or several mining function the transformed data are got. Frequently, selected data are still indistinct. To make more informed results data must be transformed further and functions must be controlled before new data mine.

- Interpreting the results

Transformed data processed continually by several methods when you want to get necessary knowledge or information

The Fig. 1 shows the basic data mining process.

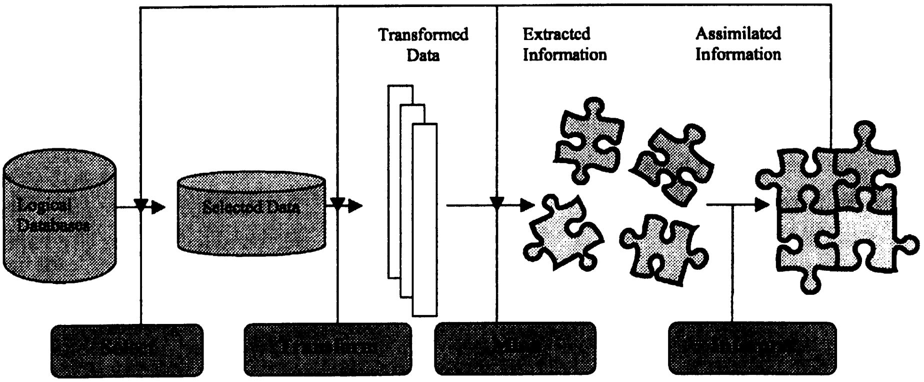

Fig. 1. The DM process.

2. Data mining operations

Four operations are associated with discovery-driven data mining

- Creation of prediction and classification models

This is the most commonly used operation primarily because of the proliferation of automatic model-development techniques. The goal of this operation is to use the contents of the database, which reflect historical data, i.e., data about the past, to automatically generate a model that can predict a future behavior. For example, a financial analyst may be interested in predicting the return of investment of a particular asset so that he can determine whether to include it in a portfolio he is creating. A marketing executive may be interested to predict whether a particular consumer will switch brands of a product of interest. Model creation has been traditionally pursued using statistical techniques. The value added by data mining techniques in this operation is in their ability to generate models that are comprehensible, and explainable, since many data mining modeling techniques express models as sets of if . . . then . . . rules.

- Link analysis

Whereas the goal of the modeling operation is to create a generalized description that characterizes the contents of database, the goal of link analysis is to establish relations between the records in database. For example, a merchandising executive is usually interested in determining what items sell together, i.e., men's shirts sell together with ties and men's fragrances, so that he can decide what items to buy for the store, i.e., ties and fragrances, as well as how to lay these items out, i.e., ties and fragrances must be displayed nearby the men's shirts section of the store. Link analysis is a relatively new operation, whose large-scale application and automation have only become possible through recently developed data mining techniques.

- Database segmentation

As database grow and are populated with diverse types of data it is often necessary to partition them into collections of related records either as a means of obtaining a summary of each database, or before performing a data mining operation such as model creation, or link analysis. For example, assume a department store maintains a database in particular visit to the store. The database can then be segmented based on the records that describe sales during the "back to school" period, records that describe sales during the "after Christmas sale" period, etc. Link analysis can then be performed on the records in the "back to school" segment to identify what items are being bought together.

- Deviation detection

This operation is the exact opposite of database segmentation. In particular, its goal is to identify outlying points in a particular data set, and explain whether they are due to noise or other impurities being present in the data, or due to casual reasons. It is usually applied in conjunction with database segmentation. It is usually the source of true discovery since outliers express deviation from some previously known expectation and form. Deviation detection is also a new operation, whose importance is now being recognized and the first algorithms automating it are being to appear.

3. Data mining techniques

While there are only four basic data mining operations, there exist numerous data mining techniques supporting these operations. Predictive model creation is supported by supervised induction techniques; link analysis is supported by association discovery and sequence discovery techniques, database segmentation is supported by clustering techniques, and deviation detection is supported by statistical techniques. To these techniques one has to add various forms of visualization, which even though does not automatically extract information, it facilitates the user in identifying patterns hidden in data, as well as in better comprehending the information extracted by other techniques.

3.1. *Supervised induction*

Supervised induction refers to the process of automatically creating a classification model from a set of records (examples), called the training set. The training set may either be a sample of the database or warehouse being mined, the entire database, or data warehouse. The records in the training set must belong to a small set of classes that have been predefined by analyst. The induced model consists of patterns, essentially generalizations over the records that are useful for distinguishing the classes. Once a model is induced it can be used to automatically predict the class of other unclassified records. Supervised induction methods can be either neural or symbolic. Neural methods, such as back propagation, represent the model as architecture of nodes and weighted links.

For example, credit card analysis is an application for which a supervised induction is well suited. A credit card issuing company may have records about its customers, each record containing a number of descriptors, or attributes. For those customers for which their credit history is known, the customer record may be labeled with a good, medium or poor labels, meaning that the customer has been placed in the corresponding class of good (medium or poor) credit risk. A supervised induction technique producing symbolic classification models may generate the rule stating. If the customer's income is over 25.000, and the age bracket is between 45 and 55, and the customer lives in XYZ neighborhood then the customer is good. A supervised induction technique is particularly suitable for data mining if it has three characteristics:

1. It can produce high quality models even when data in the training set is noisy and incomplete.
2. The resulting models are comprehensible and explainable so that the user can understand how decision is made by the system.
3. It can accept domain knowledge. Such knowledge can expedite the induction task while simultaneously improving the quality of the induced model.

Supervised induction techniques offer several advantages over statistical model-creation methods. In particular, the induced patterns can be based upon local phenomena while many statistical measures check only for conditions that hold across an entire population with well-understood distribution. For example, an analyst might want to know if one attribute is useful for predicting another in a population of 10.000 records.

If, in general, the attribute is not predictive, but for a certain range of 100 values it is very predictive, a statistical correlation test will almost certainly indicate that the attributes are completely independent because the subset of the data that is predictive is such a small percentage of the entire population.

3.2. *Association discovery*

Given a collection of items and a set of records, each of which contain some number of items from the given collection, an association discovery function is an operation against this set of records which return affinities that exist among the collection of items (Fig. 2). These affinities can be expressed by rules such as "72% of all the records that contain

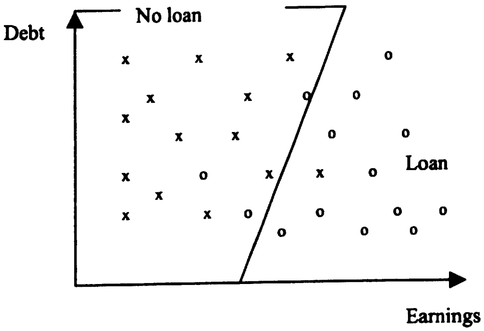

Fig. 2. Loan expansion.

items A, B and C also contain items D and E". The specific percentage of occurrences (in this case 72) is called the confidence factor of the association. Also, in this association, A, B, and C are said to be on an opposite side of the association to D and E. association discovery can involve any number of items on either side of the association. A typical application that can be built using association discovery is Market Basket Analysis. In this application, a retailer will run an association discovery function over the point of sales transaction log. The transaction log contains, among other information, transaction identifiers and product identifiers. The collection of the items mentioned above is, in this example, the set of all product descriptors. Typically, this set is of the order of 100.000 or more items. The set of products identifiers listed under the same transaction identifier constitutes a record, as defined above. The output of the association discovery function is, in this case, a list of product affinities. Thus, through association discovery the market basket analysis application can determine affinities such as "20% of the times that a specific brand toaster is sold, customers also buy a set of kitchen gloves and matching cover sets."

3.3. *Sequence discovery*

In the transaction log discussed above, the identity of the customer that did the purchase is not generally known. If this information exists, an analysis can be made of the collection of related records of

the same structure as above (i.e., consisting of a number of items drawn from a given collection of items). The records are related by the identity of the customer that did the repeated purchases.

Such situation is typical of Direct Mail application. In this case, a catalog merchant has the information, for each customer, of the sets of products that the customer buys in every purchase order. A sequence discovery function will analyze such collection of related records and will detect frequently function could also have been used in one of the examples in the previous section to discover the set of purchases that frequently precede the purchase of a microwave oven. Another example of the use of this function could be in the discovery of a rule that states that 68% of the time Stock X increased its value by at most 10% over a 5-day trading period and Stock Y increased its value between 10% and 20% during the same period, then the value of Stock Z also increased in a subsequent week.

Sequence discovery can be used to detect the set of customers associated with frequent buying patterns. Use of sequence discovery on the set of insurance claims can lead to the identification of frequently occurring medical procedures performed on patients, which in turn can be used to detect cases of medical fraud.

3.4. *Conceptual clustering*

Clustering is used to segment a database into subsets, the clusters, with the members of each cluster sharing a number of interesting properties (Fig. 3). The results of a clustering

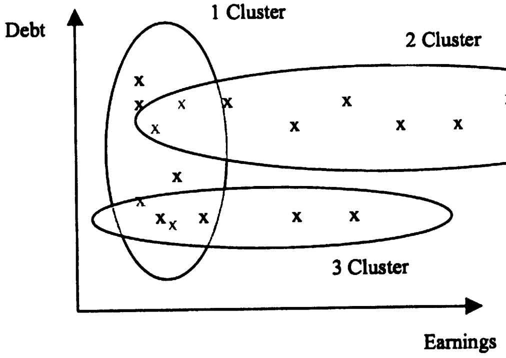

Fig. 3. Conceptual clustering example.

operation are used in one of two ways. First, for summarizing the contents of the target database by considering the characteristics of each created cluster rather than those of each record in the database. Second, as an input to other methods, e.g., supervised induction. A cluster is a smaller and more manageable data set to the supervised inductive learning component.

Clusters can be created either statistically, or using neural and symbolic unsupervised induction methods. The various neural and symbolic methods are distinguishing by (1) the type of attribute values they allow the records in the target database to take, e.g., numeric, nominal, structured objects, (2) the way represent each cluster, and (3) the way they organize the set of clusters, i.e., hierarchically or into flat lists. Once the database has been clustered, the analyst can examine the created clusters to establish the ones that are useful or interesting using a visualization component.

4. Overview of the IM components

This section provides a high-level overview of the Intelligent Miner architecture.

The IM communicates between mining and processing functions on the server, as well as between the administrative and visualization tools on the client. The client component includes a user interface from which you can invoke functions on an IM server. The results are returned to the client where you can visualize and analyze them. Fig. 4 shows the client and server components of the IM.

4.1. *User interface*

A program that allows you to define data mining functions in a graphical environment. You can define preferences for the user interface, which are stored on client.

4.2. *Environment layer Application Programming Interface (API)*

A set of API function that control the execution of mining runs and result. Sequences of functions and mining operations can be defined and executed using the user interface

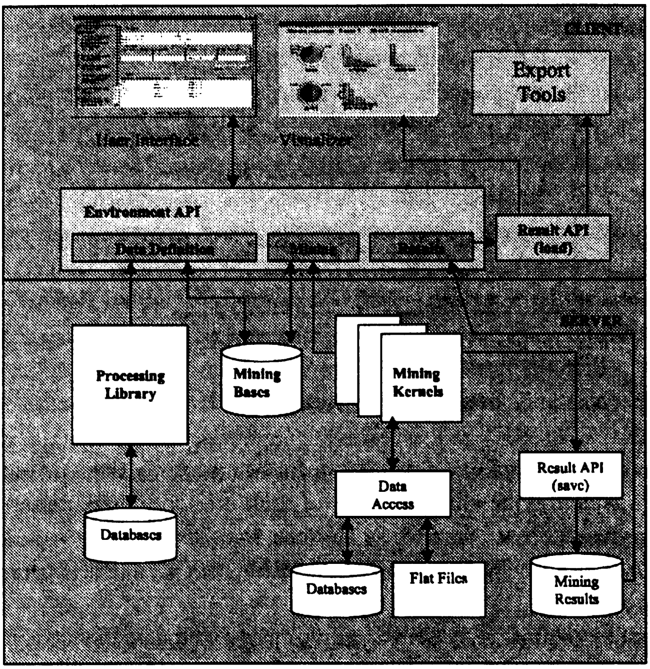

Fig. 4. The Intelligent Miner architecture.

through the environment layer API. The environment layer API is available on all server operating systems.

4.3. *Visualizer*

A tool that displays the results produced by a mining or statistical function. The Intelligent Miner provides a rich set of visualization tools. You can also use other visualization tools.

4.4. *Data access*

Data access to flat files, database tables, and database views.

4.5. *Database tables and flat files*

The data types that you can process/ the intelligent Miner components work directly with data stored in a relational database or in flat files. The data need not to be copied to a special format. You define input and output data objects that are logical description of the physical data. This logical description allows you to change the physical location of the data without affecting objects that use the data. Only the logical description must be change. The change might be as simple as changing the name of a database table.

4.6. *Processing library*

A library that provides access to database functions.

4.7. *Mining bases*

A collection of data mining objects used for a mining objective or business problem. Mining bases are stored on the server, which allows access from different clients.

4.8. *Mining kernels*

The algorithms brought into operation when you run a data mining or statistical function.

4.9. *Mining results, result API, and export tools*

The data extracted by running a mining or statistics function. These components allow you to visualize results at the client. Results can be exported for further processing or use with visualization tools.

Conclusions

Data Mining is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data. Extracted information can be used to form a prediction or classification model, or to identify similarities between database records. The resulting information can help you make more informed decisions.

The IM supports a variety of data mining tasks that is why it can be used even in medicine. The sphere of using DM is everywhere if you have data. Today primarily DM methods are using in commerce. With big success these methods are used in banking, telecommunication, insurance company, genetic engineering and etc.

References

[1] R.S. Michalski, K.A. Kaufman, Data mining and knowledge discovery: a review of issues and a multistrategy approach. textitMachine Learning and Data Mining. Methods and Applications, West Sussex, England (1998), 71–105.

[2] В. Дюк, А. Самойлснко, *Data Mining*: *учебный курс*, Санкт-Петербург.

[3] *Mining Your Own Business in Health Care*, IBM Redbooks (2001).

[4] *Mining Your Own Business in Banking*, IBM Redbooks (2001).

[5] *Mining Your Own Business in Telecom*, IBM Redbooks (2001).

[6] *Data Management Solution*, IBM Data Mining Technology, White Paper (1996).

[7] *IBM Intelligent Miner for Data*, Version 6.1.

Duomenų gavybos technologijos IBM Intelligent Miner pavyzdžiu

R. Kulvietienė, J. Mamčenko

Duomenų gavybos technologijų sistemos realizuoja naują duomenų analizės formą, paremtą intelektualiais sprendimais, leidžia gauti iš duomenų bazės žymiai gilesnes žinias, negu sudėtingiausios užklausos ir iš jų suformuotos ataskaitos.