Neural networks in data mining

Neuronové sítě v data mining

A.VESELÝ

Czech University of Agriculture, Prague, Czech Republic

Abstract: To posses relevant information is an inevitable condition for successful enterprising in modern business. Information could be parted to data and knowledge. How to gather, store and retrieve data is studied in database theory. In the knowledge engineering, there is in the centre of interest the knowledge and methods of its formalization and gaining are studied. Knowledge could be gained from experts, specialists in the area of interest, or it can be gained by induction from sets of data. Automatic induction of knowledge from data sets, usually stored in large databases, is called data mining. Classical methods of gaining knowledge from data sets are statistical methods. In data mining, new methods besides statistical are used. These new methods have their origin in artificial intelligence. They look for unknown and unexpected relations, which can be uncovered by exploring of data in database. In the article, a utilization of modern methods of data mining is described and especially the methods based on neural networks theory are pursued. The advantages and drawbacks of applications of multiplayer feed forward neural networks and Kohonen's self-organizing maps are discussed. Kohonen's self-organizing map is the most promising neural data-mining algorithm regarding its capability to visualize high-dimensional data.

Key words: association rules, data mining, decision trees, genetic algorithm, Kohonen's self-organizing maps, multilayered feedforward neural networks

Abstrakt: Nezbytným předpokladem pro úspěšné podnikání v moderní ekonomice je dostatek relevantních informací. Informace lze rozdělit na dvě základní kategorie, a to na data a znalosti. Problematika sběru dat a vytváření databázových systémů je klasickou oblastí informatiky. Naproti tomu znalostní inženýrství je disciplína mladší. Její intenzivní rozvoj začal až v osmdesátých a devadesátých letech. Znalostní inženýrství se zabývá metodami formalizace, získávání, uchovávání a udržování (aktualizací) znalostí. Znalosti jsou buď získávány od odborníků – expertů, nebo jsou induktivně odvozovány ze souborů dat, obvykle uložených v databázích. Automatické odvozování znalostí ze souborů dat se v anglické literatuře nazývá data mining (vytěžování dat, dobývání znalostí). Klasickými metodami pro získávání znalostí ze souborů dat jsou statistické metody. Pro vytěžování dat se kromě statistických metod používají nové metody, které mají svůj původ v oblasti umělé inteligence. Tyto metody vyhledávají neznámé a neočekávané vztahy, které mezi daty v databázi platí. V článku je podán přehled metod, které se v současné době pro vytěžování znalostí používají. Důraz je kladen na metody založené na neuronových sítích. Jsou diskutovány výhody a nevýhody použití vícevrstvých neuronových sítí s dopředným šířením a Kohonenových samoorganizujících se map. Kohonenova samoorganizující se mapa je v oblasti vytěžování znalostí nejvíce používaným neuronovým algoritmem pro její schopnost vizualizovat mnohorozměrná data.

Klíčová slova: asociační pravidla, genetický algoritmus, Kohonenovy samoorganizující se mapy, rozhodovací stromy, vícevrstvé neuronové sítě s dopředným šířením, vytěžování dat (data mining)

INTRODUCTION

For successful enterprising in modern agricultural business, it is necessary to posses the sufficient quantity of relevant information. When responsible managers estimate risks and opportunities on the base of inexact and obsolete information, their enterprises cannot successfully compete in the modern complex agricultural market. How to gain and use information, it is studied in knowledge engineering.

In knowledge engineering, information is parted to data and knowledge and questions of formalization, gaining, storing, refreshing and retrieval of knowledge are followed. Data are considered as elementary verifiable facts. Knowledge is considered as a set of instructions, which describe how these facts can be interpreted and used. Data describe the actual state of the world. Knowledge describes the structure of the world and consists of principles and laws. From this point of view, the validity of facts is only temporary. The validity of knowledge is more permanent.

How to gather, store and retrieve data is studied in database theory, which is a standard part of informatics. In the knowledge engineering, in the centre of interest there is knowledge and methods of its formalization and gaining are studied. Knowledge could be gained from ex-

perts, specialists in the area of interest, or it can be gained by induction from sets of data. Automatic induction of knowledge from data sets, usually stored in large databases, is called data mining. Data mining is today the main part of knowledge engineering.

METHODS OF DATA MINING

Classical methods of gaining knowledge from data sets are statistical methods. In data mining, new methods besides statistical are used. These methods have their origin in artificial intelligence. They differ from classical statistical methods in the following:

1. They look for unknown and unexpected relations, which can be uncovered by exploring data in database. They try to find such regularities in data sets, which would bring the user to the new view on his field of interest and allow him to formulate new hypotheses. Statistical methods are used in a rather different way. They verify or reject hypotheses stated a priori.
2. Data mining methods can be used also in such cases, in which utilization of classical statistical methods is not appropriate. For example, when a large volume of multivariate data is concerned or when it is not possible to suppose, that data have some standard probabilistic distribution.

In data mining, the following methods for gaining knowledge are studied and used.

– Statistical methods (prediction of time series, cluster analysis etc.)
– Production rules IF . . . THEN
– Decision trees
– Genetic algorithms
– Neural networks

Production rules IF . . . THEN

Production rules form the knowledge base of expert systems with production system architecture. In expert system design, the formulation of production rules are usually result of discussion between the knowledge engineer and a team of experts. In data mining, the methods of automatic formulation of production rules are studied. Such methods are mainly elaborated for production rules called association rules.

The purpose of association rules is to discover the associations among data in large databases, i.e., to find items that imply the presence of other items in the same transaction. Association rules were first introduced by Agraval et al. (1993).

Suppose $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items, such as $T \subset I$.

An association rule is an implication of the form $X \rightarrow Y$, where

$$X \subset I,\ Y \subset I,\ X \cap Y = \emptyset$$

Association rule $X \rightarrow Y$ holds in the database D with confidence c, if c% of transactions in D, that contain X, also contains Y.

The association rule $X \rightarrow Y$ has in database $\boldsymbol{D}$ support s, if s% transactions in D contain $X \cup Y$.

Mining association rules mean to find out all association rules that have support and confidence greater than or equal to the user specified *minimum support* (*minsup*) and *minimum confidence* (*minconf*).

Problems of automatic mining of association rules were intensively pursued during the last decade and effective algorithms were designed. See for example Agraval (1994) and Holt (2001).

Decision trees

Decision tree is a possible representation of a decision function. It is used when the complete knowledge of data is not necessary for appropriate decision and when the process of gaining data is expensive. Decision tree determines which data and in which order one should collect to achieve the effective decision with minimal average cost. Decision tree thus represents knowledge and can be used for effective decision-making. There exist algorithms for automatic construction of decision trees. Automatic construction of decision trees is the traditional part of artificial intelligence.

Genetic algorithms

Genetic algorithm is a universal method for creating objects with desired properties (Holland 1975)). The method was inspired by Darwin's evolution theory. Objects are described by a sequence of symbols, which form the analogy of genom of living organisms. Capability of an object to fulfil some function is measured by fitness function. Genetic algorithm creates generations of objects. New generation arises by mutation, crossing or by their combination from those objects of actual generation, which have the greatest value of the fitness function. Evolution modelled by genetic algorithm is aimed at the generation of objects with great value of fitness function or in other words, to the objects with the desired properties.

NEURAL NETWORKS

Algorithms based on neural networks have a lot of applications in knowledge engineering. Particularly in expert systems (see for example Quah 1996) or in modelling of human decision-making (Tan 1996 or Towell 1994).

In data mining, the following neural network architectures are usually used:

– multilayered feedforward neural networks
– Kohonen's self-organizing maps.

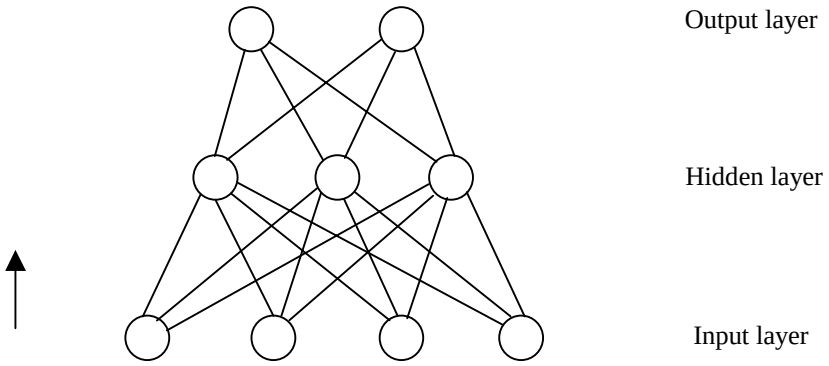

Figure 1. Multilayered feed-forward neural network (ANN)

Multilayered feedforward neural networks

Multilayered feedforward neural networks (ANNs) are in essence non-parametric regression methods, which approximate the underlying functionality in data by minimizing the loss function. The common loss function used for training and ANN is a quadratic error function. ANN is used for adaptation supervised learning. Database form a training set. During training, specified items of data records are put on the input of the neural network and its weights are changed in such a way, so that its output would approximate the values in the data set. After finishing learning process, the learned knowledge is represented by the values of neural network weights. For training the algorithm of back propagation of error is often used. Back propagation of error algorithm was first introduced by Rumelhart (1988).

In Figure 1, there is an example of an ANN with 3 layers : input layer, output layer and hidden layer. It was proved, that one hidden layer is sufficient for approximation of an arbitrary continues function.

ANNs could be used in many decision-making applications. Their advantages in these applications are:

– Capability of learning from examples.
– Capability of abstraction. It means that ANNs are able to efficiently decide also in situations which did not occur in the training set.

For applications in knowledge engineering, ANNs have a great drawback. It is their inability to provide understandable reasons for their decisions. This drawback follows from the fact, that ANNs do not create and maintain any internal representation of the external world. They do not represent knowledge explicitly, their knowledge is holistic, it is distributed in values of their weights.

Kohonen's self-organizing maps

Kohonen's self-organizing maps (SOMs) have become a promising technique in cluster analysis (Kohonen 1982, 1990). They are adapted by unsupervised learning.

The unsupervised learning process in SOM can be briefly described as follows (Figure 2). The connection weights are assigned with small random numbers at the beginning. The incoming input vectors presented by the sample data are received by the input neurons. The input vector is transmitted to the output neurons via the connections. In a "winner-take-all" competition, the output neurons with the weights most similar to the input vector became active. In the learning stage, the weights are updated following Kohonen's learning rule. The weight update only occurs for the active output neurons and their topological neighbours. The neighbourhood starts large and slowly decreases in size over time. Be-

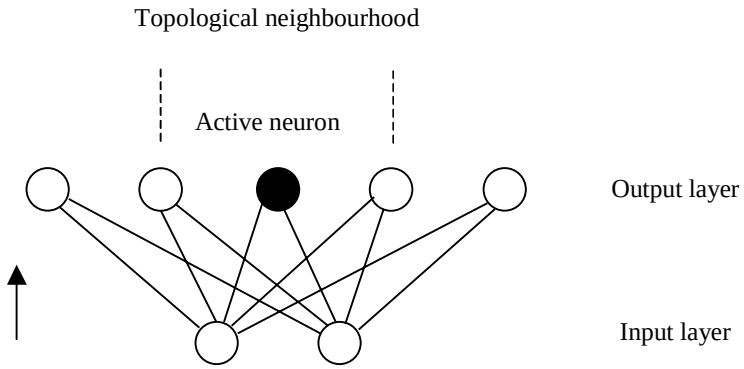

Figure 2. Kohonen's self-organizing map

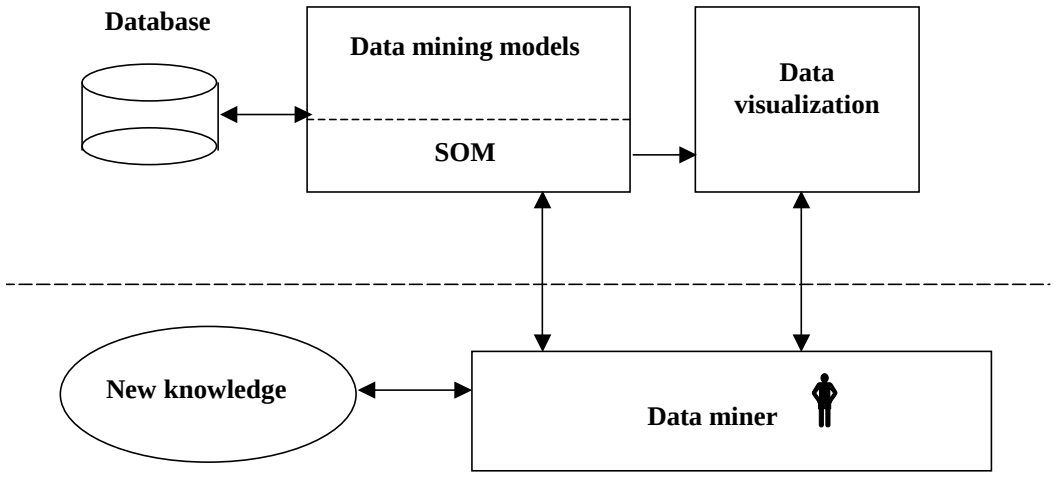

Figure 3. Apllication of SOM data mining method

cause the learning rate is reduced to zero, the learning process eventually converges.

After learning process, similar sets of items activate the same neuron. SOM divides the input set into subsets of similar records. Therefore, SOM is a method of cluster analysis and is often used for vector humanization.

In data mining, Kohonen's self-organizing maps based cluster techniques have the following advantages over standard statistical methods.

– Data mining typically deals with high-dimensional data. A record in database typically consists of a large number of items. The data do not have regular multivariate distribution and thus the traditional statistical methods have their limitations and they are not effective. SOMs work with high-dimensional data efficiently.
– Kohonen's self-organizing maps provide means for visualisation of multivariate data, because two clusters of similar members activate output neurons with small distance in the output layer. In other words, neurons that share a topological resemblance will be sensitive to inputs that are similar. This property has no other algorithm of cluster analysis.

Data mining is not merely automatic collecting of knowledge. Human-computer collaboration knowledge discovery is the interactive process between the data miner and computer (Figure 3). Data mining is human centered and is implemented through knowledge discovery loops coupled with human-computer interaction and visual representations. The aim is to extract novel, plausible, relevant and interesting knowledge from the database.

SOM is a dynamic system, which learns abstract structures in high-dimensional input space using low-dimensional space for representation. Properly designed SOM can be used to organize the high-dimensional clusters in a low-dimensional map. These low-dimensional cluster maps can be used to assist the human in discovering knowledge because they could be easily visualized. An example of data mining technique based on SOM visualization technique is described in Wang (2002). Data mining was applied in the real estate market analysis.

CONCLUSION

Knowledge discovery in database is the nontrivial process. Many different methods of data mining are used at present. From methods based on neural networks, the Kohonen's self-organizing maps are the most promising. The main reason is that Kohonen's self-organizing maps are able to visualize high-dimensional data.

Data mining methods are important in the management of complex systems. Therefore, they have a large field of application also in economics and especially in management. Software realizations of data mining algorithms are on market, but they are very expensive. On some university websites, some of them could be gained as freeware. A large list of these websites is published in Lacko (2001).

REFERENCES

Agrawal R., Srikant R. (1994): Fast algorithms for mining association rules. Proceedings of the 20th VLDB Conference: 487–499.

Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996): The KDD process for extracting useful knowledge from volumes of data. Communication of the ACM, *39* (11): 27–34.

Quah T.S., Tan CH.L., Raman K.S., Srinivasan B. (1996): Towards integrating rule-based expert systems and neural networks. Decision Support Systems, *17* (2): 99–118.

Holland J. (1975): Adaptation in Natural and Artificial Systems. MIT Press.

Holt J., Chung M. (2001): Multipass Algorithms for Mining Association Rules in Text Databases. Knowledge and Information Systems, *3* (2): 168–183 (May).

Kohonen T. (1982): Self-organized formation of topologically correct feature maps. Biological Cybernetics, *43*: 59–69.

Kohonen T. (1990): The self-organizing map. Proceedings of the IEEE, *78* (9): 1464–1480.

Lacko J., Hurt O., Kica J. (2001): Průřez nabídkou systémů pro Data Mining. Systémová integrace.

Tan Ch.L., Quah T.S., Teh H.H. (1996): An artificial neural network that models human decision making. IEEE Computer, March: 64–70.

Towell G., Shawlik J.W. (1994): Knowledge-based artificial neural networks. Artificial Intelligence, *70* (1–2): 119–165 (October).

Wang S., Wang H. (2002): Knowledge Discovery Through Self-Organizing Maps: Data Visualization and Query Processing. Knowledge and Information Systems, *4* (1), January.

Arrived on 29th August 2002

Contact address:

Ing. Arnošt Veselý, CSc., Česká zemědělská univerzita v Praze, Kamýcká 129, Praha 6-Suchdol, Česká republika
e-mail: vesely@pef.czu.cz