Journal of Applied Economic Sciences
Issue 4 (6)/Volume III /Winter 2008

ADVANCES IN KNOWLEDGE DISCOVERY IN DATABASES

Valentin **PUPEZESCU**, Felicia **IONESCU**
Politehnic University of Bucharest, **Romania**
Electronics, Telecommunications and Information Technology Faculty
vpupezescu@yahoo.com, fionescu@tech.pub.ro

Abstract:
The Knowledge Discovery in Databases and Data Mining field proposes the development of methods and techniques for assigning useful meanings for data stored in databases. It gathers researches from many study fields like machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, data visualization and grids. While Data Mining represents a set of specific algorithms of finding useful meanings in stored data, Knowledge Discovery in Databases represents the overall process of finding knowledge and includes the Data Mining as one step among others such as selection, pre–processing, transformation and interpretation of mined data. This paper aims to point the most important steps that were made in the Knowledge Discovery in Databases field of study and to show how the overall process of discovering can be improved in the future.

Keywords: KDD, Knowledge Discovery in Databases, Data Mining, Knowledge Management

1. Introduction

Knowledge Discovery in Databases (KDD) is the non–trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayadd, Piatesky–Shapiro, and Smyth, (1996)].

Data Mining (DM) represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data [Fayadd, Piatesky–Shapiro, and Smyth, (1996)].

The KDD process has developed due to the immense volume of data that must be handled easier in areas such as: business, medical industry, astronomy, genetics or banking field. Also, the success and the extraordinary development of hardware technologies led to the big capacity of storage on hard–disks, fact that challenged the appearance of many problems in manipulating immense volumes of data. Of course the most important aspect here is the fast growth of the Internet.

The core of the KDD process lies in applying DM methods and algorithms in order to discover and extract patterns from stored data but before this step data must be pre–processed. It is well known that simple use of DM algorithms does not produce good results. Thus, the overall process of finding useful knowledge in raw data involves the sequential adhibition of the following steps: developing an understanding of the application domain, creating a target data set based on an intelligent way of selecting data by focusing on a subset of variables or data samples, data cleaning and pre–processing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, the data mining step, interpreting mined patterns with possible return to any of the previous steps and consolidating discovered knowledge [Fayadd, Piatesky–Shapiro, and Smyth, (1996)].

Typical DM tasks are: classification – is learning a function that maps (classifies) a data item into one of several predefined classes [Weiss and Kuliakowski (1981); Hand (1981)], regression – is learning a function that maps a data item to a real–valued prediction variable [Fayadd, Piatesky–Shapiro, and Smyth, (1996)], clustering – is the partitioning of a data set into subsets (clusters), association rules – determine implication rules for a subset of record attributes, summarization – involves methods for finding a compact description for a subset of data [Fayadd, Piatesky–Shapiro, and Smyth, (1996)], dependency modelling – consists of finding a model that describes significant dependencies between variables [Fayadd, Piatesky–Shapiro, and Smyth,

(1996)], change and deviation detection – represents the search for finding the most important changes in the data from previous measured values [Fayadd, Piatesky–Shapiro, and Smyth, (1996)].

The discovery of knowledge in databases contains many study areas such as machine–learning, pattern recognition in data, databases, statistics, artificial intelligence, data acquisition for expert systems and data visualization. The most important goal here is to extract patterns from data and to bring useful knowledge into an understandable form to the human observer. It is recommended that obtained information to be facile to interpret for the easiness of use. The entire process aims to obtain high–level data from low level–data.

In terms of applying the KDD process there are a wide variety of sciences in which it can be used such as biology, medicine, genetics, astronomy, high–energy physics, banking, business and many others. DM methods and algorithms can be applied on a multitude of information from plain text to multimedia formats.

2. State Of The Art in KDD

The studies made about knowledge discovery in databases are advanced regarding DM methods and algorithms used to extract knowledge from data.

The main goals of KDD are: verification and discovery [Fayadd, Piatesky–Shapiro, and Smyth, (1996)]. With the verification goal, the system takes account only of user's hypothesis. Following the discovery goal, the system acts autonomous in finding useful data [Fayadd, Piatesky–Shapiro, and Smyth, (1996)]. Further, the discovery goal is subdivided in prediction and description [Fayadd, Piatesky–Shapiro, and Smyth, (1996)]. In the prediction goal, the system extracts patterns from data in order to predict future behavior of some entities. Description focuses on finding human–interpretable patterns describing the data. The importance of prediction and description goals for certain data mining applications can vary very much. However, in the context of KDD, description tends to be more important than prediction. This is in contrast to pattern recognition and machine learning applications (such as speech recognition) where prediction is often the primary goal of the KDD process [*Knowledge Discovery in Database*]. These goals are achieved with DM tasks.

For each DM task there were developed a wide variety of data mining algorithms and methods such as: decision trees, decision rules, non–linear regression, classification methods (the neural networks play an important role here), example based methods, models based on relational learning.

Intersecting the KDD field with parallel computing or distributed computing can develop the existent DM algorithms. The most important problem in the process of finding knowledge is the optimized applying of all KDD steps. Each step from the KDD process takes an amount of time. Besides analyzing certain neural network methods of classification, in this thesis we will focus also on the possibility of minimizing the amount time consumed on some KDD steps.

When the KDD term was introduced back in 1989 by the researcher Gregory Piatesky–Shapiro, there weren't too many data mining instruments for resolving one single task. A good example is the C4.5 decision tree algorithm [Quinlan, (1986)] and SNNS neural network, or parallel–coordinate visualization [Inselberg, (1985)]. This tools were hard to use and required important data preparation [Piatesky–Saphiro, (1991)].

The second–generation data mining systems were called suites and were developed by vendors, starting from 1995. These tools took into account that the KDD process requires multiple types of data analysis, and most of the effort is spent in the data cleaning and preprocessing steps. Suites like SPSS Clementine, SGI Mineset, IBM Intelligent Miner, or SAS Enterprise Miner allowed the user to perform several discovery tasks (usually classification, clustering, and visualization) and also supported data transformation and visualization. One of the

most important advances, pioneered by Clementine, was a GUI (Graphical User Interface) that allowed users to build their knowledge discovery process visually [Piatesky–Saphiro, (1991)].

By the year 1999, there were over 200 tools available for solving different tasks but even the best of them addressed only a part from the overall KDD framework. Data still had to be cleaned and preprocessed. The development of this type of applications in areas like direct marketing, telecom, and fraud detection, led to emergence of data–mining–based "vertical solutions". The best examples of such applications are the systems HNC Falcon for credit card fraud detection, IBM Advanced Scout for sports analysis and NASD KDD Detection system [Kirkland, (1999); Piatesky–Saphiro, (1991)]

A very important issue is the way that data was stored over the time. Many years the main approach was to use a specific DM method or algorithm on a data set. In most cases the data set was stored in a centralized database. In present, because of big volumes of data the main solution is to use distributed databases systems. For mining in this data in the traditional way it is supposed that all data stored on local computers should be transferred on a central point for processing. In most cases this would be impossible because the existent connection bandwidth won't permit such big transfers. A very important matter is that when big transfers are made over the Internet can appear security issues: the intimacy of client's data must be kept. Thus, a new KDD study area appeared that was called Privacy Preserving Data Mining. This field focuses on studying the security risks that can occur in the KDD process. Because the number of steps that are included in the KDD framework is relative big client's data that are mined can be violated. Privacy Preserving Data Mining tries to create algorithms that may prevent such problems [University of Munich Institute for Computer Science Database and Information Systems].

In the last years the KDD process was approached from two perspectives: parallel and distributed computing. These directions led to the apparition of Parallel KDD and Distributed KDD. In Parallel KDD, data sets are assigned to high performance multi–computer machines for analysis. The availability of this kind of machines is increasing and all algorithms that were used on single–processor units must be scaled in order to run on parallel–computers. The Parallel KDD technology is suitable for scientific simulation, transaction data or telecom data. Distributed KDD must provide solutions for local analysis of data and global solutions for recombining local results from each computing unit without causing massive data transfer to a central server. Parallel computing and distributed KDD are both integrated in Grid technologies. One of the creators of Grid concept, I. Foster wrote the followings: *The real and specific problem that underlies the Grid concept is coordinated resource sharing and problem solving in dynamic, multi–institutional virtual organizations (VO). The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem–solving and resource–brokering strategies emerging in industry, science, and engineering. Among them, Data Mining is one of the most challenging.*

Grid computing emerged because computational power is falling behind storage possibilities. The annual doubling of data storage capacity managed to reduce the cost of a terabyte and now many researchers in physics or astronomy discuss the possibility of mining into petabyte archives. The solution to these problems lies in dramatic changes taking place in networking. All Data Mining algorithms and methods must be adapted to operate intelligent with raw data stored in a distributed way. Managing such great quantities of data over such big geographical distances brings in discussion the security problem again. Developing study fields like Privacy Preserving Data Mining is crucial for keeping the intimacy over client's data intact.

Next–generation grids will face many problems such as the management and exploitation of the overwhelming amount of data produced by applications but also Grid operations, and the intelligent use of Grid resources and services. The new generations of Grids should contain

knowledge discovery and knowledge management functionalities, for both applications and system management.[Cannataro, (2003)]

3. KDD and Neural Networks

The Data Mining step is at the heart of the KDD process. Many of the DM tasks are achieved with the help of neural networks. The computing model of these networks is the human brain, so neural networks are supposed to share some brain abilities to learn and adapt in response to external inputs. When exposed to a set of training or test data, neural networks can discover previously unknown relationships and learn complex non–linear patterns in the data.

One of the most important function that our brain is able to do is the ability to classify between two things. The evolutional success of many species was made possible through this ability of discerning between what is friendly and what is dangerous for a particular specie.

Although classification is very important, sometimes it can be overdone. Because of our limited storage capacity it's crucial for us to be able to group similar notions or objects together. This ability is called clustering and is one of the main tools involved in reasoning. Because of this skill we can think in terms of abstracts notions and solve problems by seeing the hole picture and neglect unimportant details.

Regression is learning function that maps a detail of data into a real variable used for prediction. This is similar with what people do: just from seeing a few examples people can learn to interpolate between the examples given in order to generalize to new problems or cases that were not encountered. In fact the ability of generalize is one of the strongest points in using the neural network technology.

Neurophysiologists proved that the learning theory of brain called Hebbian is true: people store information by associating ideas with other related memories. In our brain there is a complex network of semantically related ideas. The theory says that when two neurons are activated in the brain at the same time, then the connection between them grows stronger and that physical changes in the synapse of the neurons took place. Associative memory branch from the neural network filed deals with implementing models that describe the associative behaviour. Neural networks such as Binary Adaptive Memories and Hopfield networks shown to be limited capacity, but working, associative memories.

These DM operations are resolved with different neural models: backpropagation networks, recurrent backpropagation networks, Self–organizing Maps(SOM), Radial Basis Function networks(RBF), Adaptive Resonance Theory networks (ART), probabilistic neural networks, Fuzzy perceptrons. Each structure can make certain DM operations.

We use in our experiments the following distributed architectures only for resolving the classification task:

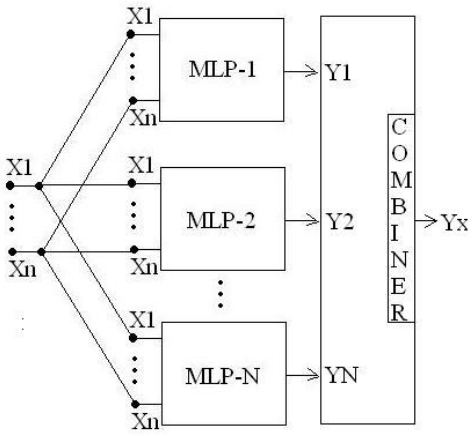

Figure 1. Multilayer Perceptron Committee Machine.

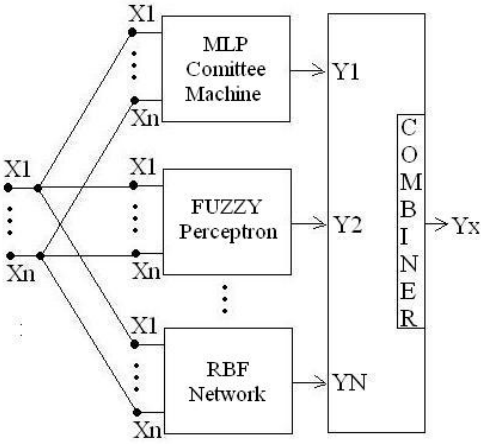

Figure 2. General Committee Machine

In the General Committee Machine (G–CM) arhitecture we have multiple neural topologies runing in the same time in a distributed way. The first block from G–CM (MLP–CM) is working in a distributed manner by itself [Mukarram, and Tahir, (2007)]. This method is prefered because of the advantages that every topology has to offer. Multilayer perceptron should be more resistant to noise than other topologies. The Fuzzy Perceptron and RBF Neural Networks are used mainly for speed. Natural organisms are equipped with multiple instruments for analyse certain problems so the proposed architectures must be fast and reliable. The entire system is autonomous and will try not to use more computational power than it's necessary. In the MLP–CM, each MLP–block will start with random weights and will work in parallel in order to reach a global error faster than normally. The result will depend on the randomly generated weights so some of the MLP–blocks will reach local minimums but some of them will reach a global minimum. Each block will have its error function. The final results will be transmited to the combiner in order to select the best result.

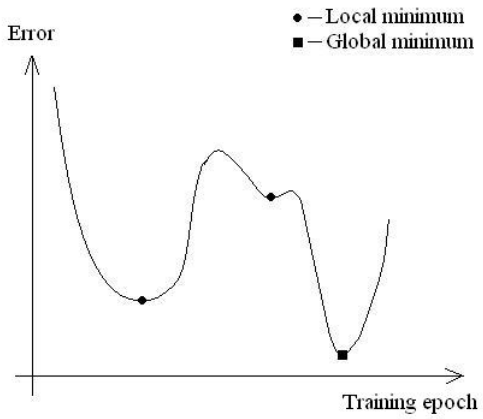

Figure 3. Example of error function

The architecture proposed in Fig.2 is the most suited for performing the clasification task in the buisness field because of the resistance to noise data.

As we can see, the ultimate goal of the neural networks is to imitate the behaviour of natural organisms.

At the base of neurology lies the studies of neurons. Neural Network field constantly benefits and depends on this field of research. Recent studies show that many of the old views should be revised because the structure and the behaviour of the natural neurons. First of all, it seems that electric synapses are more common than previously thought [Connors, and Long, (2004)]. So, instead of having an individual functionality, in some portions of the brain we have distributed computing of the information – many groups of neurons are activated simultaneous. Another observation is that dendrites and axons have the so called "voltage–gated ion channels" which can generate electric information carrying potentials from and towards the soma. This behaviour rise doubts about many actual theories that stipulate that dendrites are passive information receivers and the axons sole transmitters. These observations lead to the conclusion that neurons are more complex in structure and operations than previously thought. An important fact is also that role played by glia cells. Neurons and glias are the main components of the central nervous system. The number of the glia cells is ten times bigger than that of neurons. New studies about glia cells show that these cells are vital for the processing of information [Witcher, Kirov, and Harris, (2007)].

The McCulloch–Pitts neuron is a reduced model of the real neuron. Because of the new discoveries that are made in neurology research a new neuron model will appear eventually. These is a very feasible prediction and it can be sustained by the increase of computing power, parallel and distributed computing technologies.

4. Future Directions of Study

Although there are many future directions of study into this field of research, we can summarize the most important of them:

- The continuous development and optimisation of the Data Mining algorithms. From those presented we can predict that the neural network field can still be improved by bringing artificial neuron models as close as possible to the functionality of the biologic neuron. As we come close to the real neuron we should obtain similar performances and because of parallel and distributed technologies the functionality of large numbers of neuron units acting in a simultaneous way will be achieved.
- One of the most important study is the one about the implementation of the KDD steps on GRID platforms. Other areas of research are Parallel KDD and Distributed KDD. Because of the fact that Knowledge Discovery in Databases is an intense computing process the problem of testing specific algorithms in distributed systems is still opened.
- The continuous improvement of additional steps from the KDD process. A very important matter here is that in the real world analysed data is not pre–processed. Most of the time it is incomplete or incorrect. So the data must be pre–processed in an intelligent and optimised manner. It is well known that DM algorithms applied on wrong data it's worthless time consumption. Certain tests must be made to see exactly how the well known neural topologies behave when they have also noise besides good data at the input. This problem must be determined because there are groups of researchers that are saying that some neural topologies are more resistant to noise and other groups that are saying that all topologies are affected equally.
- The development of expert acquisition systems can resolve many of the aforementioned problems. A big problem is that many DM systems loose so much time with eliminating and correcting the rough data. It would be correct to move this process to the moment in which the data are gathered. This is one of the main reasons that slows down the development of KDD.

We propose to follow the second scheme as much as possible when data are collected for Data Mining:

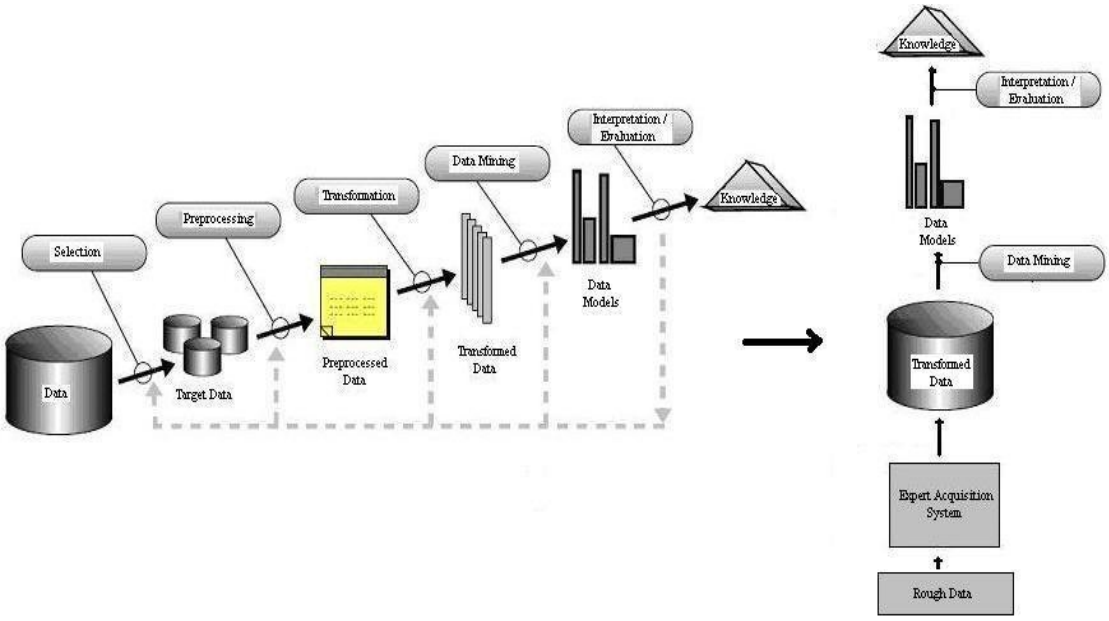

Figure 4: The classic KDD process and the new proposed model

The schemes are mainly the same but the difference is that the Expert Acquisition System (EAS) makes all the verifications at the moment when data are gathered. A big part from the selection and preprocessing steps is already done by the EAS. The best situation for KDD systems is to have good data from the start. It is well known that selection, preprocessing and transformation steps take more time than the Data Mining step.

▪ One of the most important future directions of study in KDD is also to resolve the security issues that might appear. Working in distributed systems is unavoidable because DM is an intense computing process so we can have important data exposed to other parties. This is something that must be avoided.

5. Conclusions

Knowledge Discovery in Databases process still poses many problems to the researchers.

In our future research we aim to study the performances of applying certain neural network algorithms on stored data and the possibility to improve the results by optimising the way that data is stored. We will treat also the performances of specific neural algorithms applied on data stored into centralized and distributed databases and observe exactly what effects has the noise on analysed data and how well certain topologies are more or less sensitive to it. Tests will be made to see how the dimensionality of databases affects the performances of the neural network architectures. The overall process of discovering useful information in data will be analysed and we hope to improve some steps from the KDD process.

6. References:

[1] Cannataro, M., (2003), *Knowledge Discovery and Ontology–based services on the Grid*

[2] Connors, B, Long M., (2004), *Electrical synapses in the mammalian brain*. Annu Rev Neurosci 27: 393–418.

[3] Domenico, T., *Grid–based Distributed Data Mining Systems, Algorithms and Services.*

[4] Fayadd, U., Piatesky–Shapiro, G., and Smyth, P. (1996), *From Data Mining To Knowledge Discovery in Databases*, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262–56097–6 Fayap.

[5] Foster, I., Kesselman, C., and Tuecke, S. (2001). *The anatomy of the grid: Enabling scalable virtual organizations*. Intl. J. Supercomputer Applications, 3(15).

[6] http://www2.cs.uregina.ca/~hamilton/courses/831/index.html, *Knowledge Discovery in Databases.* Department of Computer Science, University of Regina.

[7] Mukarram, A.Tahir. (2007), *Java Implementation of Neural Network*, ISBN 1–4196–6535–9.

[8] Piatesky–Saphiro, G., (1991), *Knowledge Discovery in Databases: 10 Years After*

[9] Piatesky–Saphiro, G., Knowledge Stream Partners. *The Data–Mining Industry Coming of Age.*

[10] Piatesky–Saphiro, G., *Machine Learning and Data Mining. Course Note.*

[11] The Handbook of Data Mining, (2004), *Edited by Nong Ye,* Publisher: Lawrence Erlbaum Associates, Inc.

[12] University of Munich Institute for Computer Science Database and Information Systems. *Knowledge Discovery in Databases*

[13] Witcher, M, Kirov, S., Harris, K., (2007), *Plasticity of perisynaptic astroglia during synaptogenesis in the mature rat hippocampus.* Glia 55 (1): 13–23.