Sanjay Kumar Sen, Dr Sujata Dash, Subrat P Pattanayak / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp. 342-348

AGENT BASED META LEARNING IN DISTRIBUTED DATA MINING SYSTEM

[1]Sanjay Kumar Sen, [2]Dr Sujata Dash, [3]Subrat P Pattanayak,
[1]Faculty (Comp Sc & Engg.), BEC, Bhubaneswar
2 Principal, KMBB College, Bhubaneswar
[3] Faculty (Comp Application), RIMS, Rourkela

ABSTRACT

The data mining technology is used to identifying patterns and information from a huge quantity of data. In a single repository data base where data is stored in central site, then applying data mining algorithms on these data base, patterns are extracted, which is clearly implausible and untenable for many realistic problems and databases. To deal with these complex systems has revealed opportunities to improve distributed data mining systems in a number of ways. Furthermore, in certain situations, data may be inherently distributed and cannot be merged into a single database for a variety of reasons including security, fault tolerance, legal constraints, competitive reasons, etc. In such cases, it may not be possible to examine all of the data at a central processing site to compute a single global model. Here, we develop techniques that scale up to large and possibly physically distributed databases. Meta-learning (learning from learned knowledge) – a technique dealing with the problem of computing a global classifier from large and inherently distributed databases. This paper, describes meta-learning and JAM system (Java Agents for Meta-learning), which is an agent-based meta-learning system for large-scale data mining applications. Several important desiderata of data mining systems are addressed (i.e., scalability, efficiency, portability, compatibility, adaptivity, extensibility and effectiveness) and a combination of AI-based methods and distributed systems techniques are presented. We applied JAM on the real-world data mining task of modeling and detecting credit card fraud with notable success.

Keywords **: compatibility, distributed data mining, global classifier. meta-learning, portability, scalability.**

1. Introduction

Data mining systems aim to discover patterns and extract useful information from facts recorded in databases. One means of acquiring knowledge from databases is to apply various machine learning algorithms that compute descriptive representations of the data as well as patterns that may be exhibited in the data. Most of the current generation of learning algorithms, however, are computationally complex and require all data to be resident in main memory which is clearly untenable for many realistic problems and databases. Furthermore, in certain situations, data may be inherently distributed and cannot be merged into a single database for a variety of reasons including security, fault tolerance, legal constraints, competitive reasons, etc. In such cases, it may not be possible to examine all of the data at a central processing site to compute a single global model. Traditional data analysis methods that require humans to process large data sets are completely inadequate. Applying the traditional data mining tools to discover knowledge from the distributed data sources might not be possible [19]. Hence, knowledge discovery from multi-databases has became an important research field and is considered to be a more complex and difficult task than knowledge discovery from mono-databases [22]. The relatively new field of *Knowledge Discovery and Data Mining* (KDD) has emerged to compensate for these deficiencies. *Knowledge discovery* in databases denotes the complex process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [8]. *Data mining* refers to a particular step in the KDD process. According to the most recent and broad definition [8], "data mining consists of particular algorithms (methods) that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data." A common methodology for distributed machine learning and data mining is of two-stage, first performing local data analysis and then combining the local results forming the global one [20]. For example, in [21], a meta-learning process was proposed as an additional learning process for combining a set of locally learned classifiers

(decision trees in particular) for a global classifier. Hidden Markov Model is the statistical tools for engineer and scientists to solve credit card fraud can be detected using Hidden Markov Model during transactions[4]. Abhinav Srivastava, Amlan Kundu, Shamik Sural, Senior Member has shown the system of credit card fraud detection by using HMM model by bearing in mind a cardholder's spending habit without fraud signature and have also suggested a method for finding the spending profile of cardholders, [7]as well as application of this knowledge in deciding the value of observation symbols and initial estimate of the model parameters.. Machine-learning algorithms have been deployed in , in detecting credit card fraud[11], in steering vehicles driving autonomously on public highways at 70 miles an hour [13], and in computing customized electronic newspapers [10], to name a few applications. Adnan M. Al-Khatib in his research paper "to present a high accuracy method or prototype to detect Card-Not-Present (CNP) Fraudulent transactions in the e-payment systems" by integrating data from multiple databases (e.g., bank transactions, federal/state crime history DBs);[3] and then using suitable and effective data mining and artificial intelligence (AI) tools to find unusual access sequences. Recently data mining techniques have been successfully applied to intrusion detection in network-based systems [14]. The main focus of this paper is on the management of machine learning programs with the capacity to travel between computer sites to mine the local data. The term "management" denotes the ability to dispatch and exchange such programs across data sites, but also the potential to control, evaluate, filter, resolve compatibility problems and combine their products (that can too be intelligent agents). Data mining refers to the process of extracting automatically, or semi-automatically novel, useful, and understandable pieces of information (e.g., patterns, rules, regularities, constraints) from data in large databases. One way of acquiring knowledge from databases is to apply various machine learning algorithms that search for patterns that may be exhibited in the data and compute descriptive representations. Machine learning and classification techniques have been successfully applied in many problems in diverse areas with very good success. Although the field of machine learning has made substantial progress over the past few years, both empirically and theoretically, one of the continuing challenges is the development of inductive learning techniques that effectively scale up to large and possibly physically distributed data sets. This paper investigates data mining techniques that scale up to large and physically distributed databases. In this respect, we induct some additional features into JAM (Java agents for Meta-learning), an agent-based distributed data mining system that supports the remote dispatch and exchange of learning agents across multiple data sources and employs meta-learning techniques to combine the separately learned models into a higher level representation.

2 Meta-learning

Meta-learning[19] is loosely defined as learning from learned knowledge. Meta-learning is a recent technique that seeks to compute higher level models, called meta-classifiers, that integrate in some principled fashion the information cleaned by the separately learned classifiers to improve predictive performance. In meta learning process a number of learning programme is executed on a number of data subsets in parallel then collective result is collected in the form of classifiers.

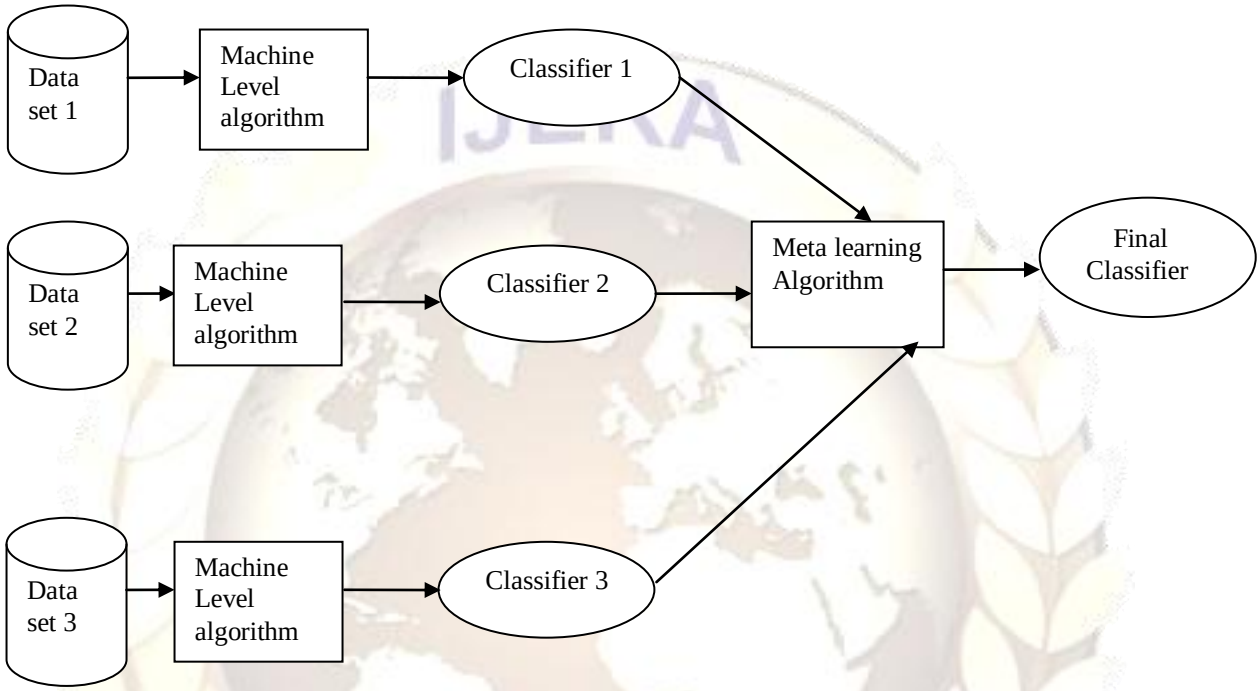

Figure – 1

From Figure 1, the different stages of a simplified meta-learning scenario are

1.There are three data sets namely, data set 1. data set 2, data set 3 which are called base-level training data sets.

2.These data sets are executed by machine level programme to produce three base classifiers.

3. These three base level classifiers are trained from the meta-level training to compute final classifier.

Meta-learning Techniques

In this section we describe previous research in meta-learning and in particular address the following specific research issues:

1. Using a variety of statistical, information-theoretic and the characterization of datasets is performed.

2. By applying a set of algorithms at the base level and combining these through a meta learner information is extracted.

3. To accelerate the rate of learning process, knowledge is extracted through a continuous learner.

In meta-learning (learning from learned knowledge) technique dealing with the problem of computing a global classifier from large and inherently distributed databases. A number of independent classifiers – "base classifiers" -are computed in parallel. The base classifiers are then collected and combined to a „meta-classifier" by another learning process. Meta-classifiers can be defined recursively as collections of classifiers structured in multi-level trees [12]. Such structures, however, can be unnecessarily complex, meaning that many classifiers may be redundant, wasting resources and reducing system throughput.

- The predictive accuracy of base classifiers is improved.
- Assuming that a system consists of several databases interconnected through an intranet or internet, the goal is to provide the means for each data site to utilize its own local data and, at the same time, benefit from the data that is available at other data sites without transferring or directly accessing that data.

4. Advantages of Meta-learning

The same base learning process is executed in parallel by meta-learning on subsets of the training data set which improves efficiency. Because the same serial programme is executed in parallel which improves time complexity. Another advantage is that learning is in small subsets of data which can easily accommodated in main memory instead of huge amount of data. Meta-learning combines different learning system each having different inductive bias, as a result predictive performance is increased. A higher level learned model is derived after combining separately learned concepts. Meta-learning constitutes a scalable machine learning method because it generalizes to hierarchical multi-level meta-learning. Also most of these algorithms generate classifiers by applying the same algorithms on different data base.

Scalability

The data mining system is highly scalable because its performance does not hamper as the data sites increases. It depends on the protocols that transfer and manage the intelligent agents to support the data sites.

Efficiency

It refers to the effective use of the available system resources. It depends on the appropriate evaluation and filtering of the available agents which minimizes redundancy.

5. JAM (Java Agent for Meta-Learning) :

Meta-learning system is implemented by JAM system(Java Agents for Meta-learning) which is a distributed agent-based data mining system. It provides a set of learning agents which are used to compute classifier agents at each site. The launching and exchanging of each classifier agents takes place at all sites of distributed data mining system by providing a set of meta-learning agents which combined the computed models those computed models at different sites. We have achieved this goal through the implementation and demonstration of a system we call JAM (Java Agents for Meta-Learning). To our knowledge, JAM is the first system to date that employs meta-learning as a means to mine distributed databases. A commercial system based upon JAM has recently appeared [15].

The JAM Architecture:

First of all local classifiers are computed on each local data site by executing learning agents. Then these local computed classifiers are exchanged between local sites combine with each local classifiers through meta-learning agents. Each local data sites is administered by local configuration file which is used to perform the learning and meta-learning task. To supervise agent exchange work and execution of meta-learning process smoothly, each data site is employed by GUI and animation facilities of JAM. After computing the base and meta-classifiers, the JAM system executes the modules for classification of desired data sets. The configuration file manager(CFM) is used as server which is responsible for keeping the state of the system up to date.

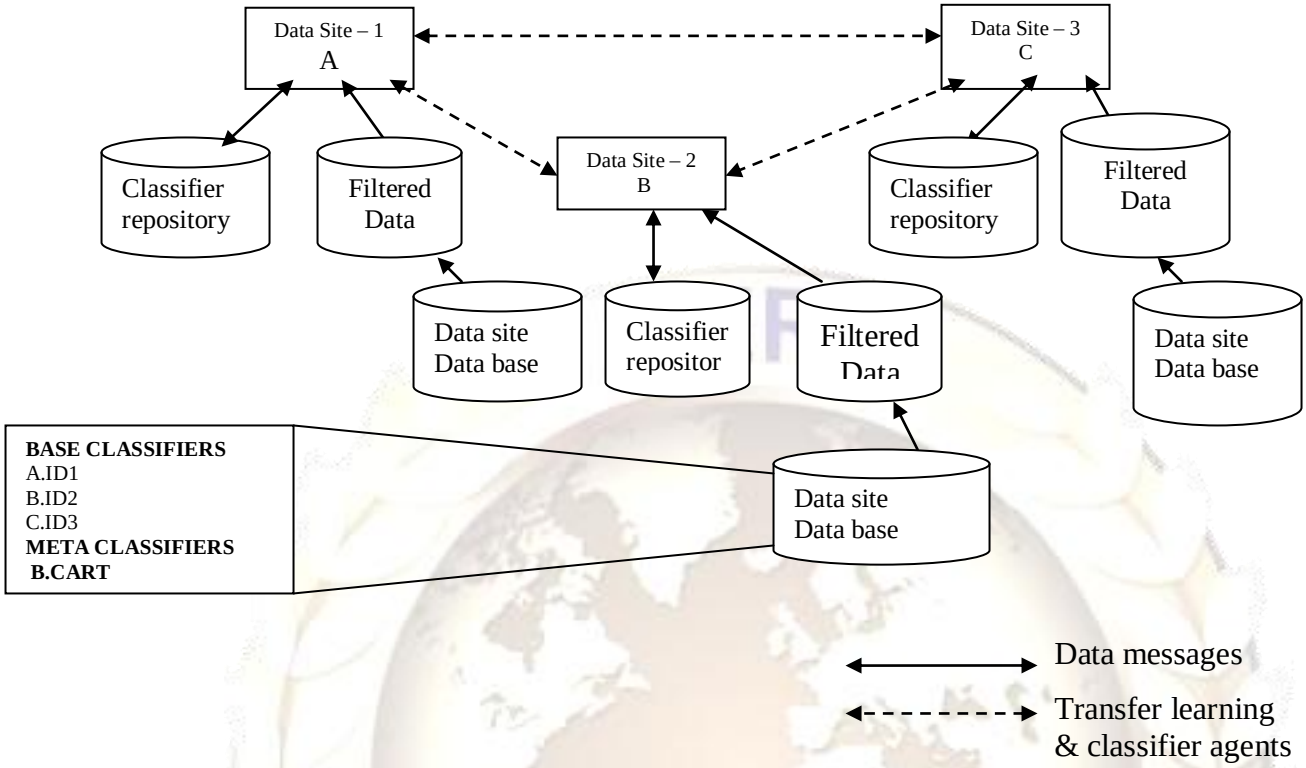

Figure 1.1. The JAM Architecture

The above figure is an example of architecture of meta learning system. There are three data sites namely A of Data site1,B of Data site2 and C of Data site3 respectively share their learning task by exchanging their local classifiers. Figure 1.1 depicts the JAM system with three data sites A, B, C while exchanging their classifier agents. Here B has imported the RIPPER classifier agent from A and ID3 classifier agent from C and combined with its own classifier agent CART to form a local meta-classifier agent BAYES. JAM is under distributed protocols in which participating data base sites execute independently and also collaborate with other data sites. Data mining system with JAM system solves the problem of how to make a learning system evolve and adjust according to its changing environment. For example in medical science the data that are related to various things for example the types of dosages, treatments and data which are in medical data base changes over time and also another example is the credit card data where new security system are introduced and also new ways to commit fraud are framed. Though the traditional data mining systems are static which can not adapt to new systems, but data mining systems with JAM are adaptive to new environment. This adaptively in JAM is achieved by employing meta-learning techniques to design learning systems capable of incorporating into their accumulating knowledge the new classifiers that capture patterns which are learner on new data sources.

6. Fraud and Intrusion Detection

The present data mining system must be flexible to accommodate not only data and patterns which are almost changing from time to time, but also machine learning algorithm and data mining technology which are changing over time. JAM is designed using object-oriented methods that can be implemented independently of any particular machine learning program or any meta-learning or classifier combining technique. The learning and meta-learning agents are designed as objects. JAM provides the definition of the parent agent class and every instance agent (i.e. a program that implements any of your favorite learning algorithms ID3 [16], CART [17], Bayes [18], etc.) are then defined as a subclass of this parent class. For smooth operation of electronic commerce system e.g. inter banking network etc. it required smooth operation and also it requires access only to legitimate user through verification and authentication mechanism. It also thwarts fraudulent activity attempted by frauds. To overcome fraudulent activities in financial systems from threats, we

adopted a mechanism for protection by using the models of errant transaction behaviour that consists of pattern-directed inference systems which can cautious and forewarn impending threats. By using JAM we can compute fraudulent models by analyzing the huge and distributed data base. By computing local fraud detection agents which can detect fraudulent transaction and intrusion detection activities which happened in a single financial corporation, JAM used to integrate meta-learning system that combines the collective knowledge acquired by individual local agents. JAM used to construct meta learned system by sharing the models of fraudulent transactions through exchange of classifier agents in secured agent infrastructure.. This meta classifier agent used to compute meta classifiers that used to act as sentries forewarning of possibly fraudulent transactions and threats by inspecting, classifying and labeling every transaction.

7. Detection of Fraud Label Approach

Step.1 : In total N numbers of banks, each bank Bi having own data base Di uses its trained data Ti and combines with its own learning agent to produce classifier ci.

Step 2 : Then each bank sent its local classifier agents ci to a central data repository where there is global trained data base T of all banks i.e. $T = T_1 \cup T_2 \cup \ldots\ldots T_n$

Step 3 : In central repository, with using some algorithm and by taking its trained data T and all classifiers c_i of all banks, meta classifier C is produced.

Step 4 : This meta classifier or global classifier C is sent to all banks as a data filter to mark as a fraud label.

8. Evaluation

Here we provide a data base schema of several banks which has transaction data sets. To capture fraudulent activities of banks, several information are continuously analysed by banks. The information of customer contains its credit card no, amount of transaction, details past information about transaction, transaction data and time , the age of account and card, the locational information of transactor and transaction, confidential and proprietary field, other credit card account information, the fraud label etc. The approach can be described by a suitable illustration. Let us assume that there are three banks, namely "A" and "B" and "C". Each of these banks has its own data base sites namely, DB1 and DB2 and DB3. Taking trained data T1, T2 and T3 from each of the data base DB1, DB2, DB3 respectively and combining with each of its own learning algorithm classifiers C1, C2,C3 are generated. There is a central repository data base called DB where the trained data of three data sites of three banks and three classifiers are stored. In central repository these three classifiers C1,C2 and C3 along with three data base are combined with a leaning algorithm to generate a new classifier C which is also called global classifier . This global classifier C is sent to the three data base sites of three banks i.e. DB1, DB2 and DB3 acting as a data filter to mark as a fraud label.

8. Conclusion and future work.

The main objective of the distributed data mining system is to combine information and patterns and extract from remote data bases which are distributed in multiple databases. To fulfill this object and to discover the various descriptive patterns to compute the classifiers by applying very machine learning programmes. In this paper we aim to find useful information and efficiently and more accurately from huge and distributed data bases. To achieve this objective we applied a agent based meta-learning system called JAM in this distributed data mining system. Various machine learning programmes are implemented in meta-learning system to compute the combining classifier models for smooth and effective mining of large data mining system. For JAM system in distributed data mining system., we have also discussed several issues like scalability, efficiency. Scalability is showed by employing its performance which does not hamper as the data sites increases which depends on the protocols that transfer and manage the intelligent agents to support the data sites. Efficiency is achieved by effectively using the available system resources. It has also been discussed about fraud detection approach. This paper has provided only overview of distributed data mining system but not provided detailed exposition of techniques. So it requires extensive research on all issues.

References

[1] Adnan M. Al-Khatib and Ezz Hattab; "Mining Fraudulent Transactions in e-payment Systems"; the 9th international conference (ii WAS 2007); 2007; P.P. 179 – 189.

[2] Adnan M. Al-Khatib; "Mining Fraudulent Behavior in epayment Systems"; Ph.D. Dissertation; 2007.

[3] Adnan M. Al-Khatib, "Detect CNP Fraudulent Transactions" *World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No.8, 326- 332, 2011*

[4] SHAILESH S. DHOK "Credit Card Fraud Detection Using Hidden Markov Model International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[5] Distributed learning with bagging-like performance, *Pattern Recognition Letters* **24**, 455–471.(Datta et al., 2006) Datta, S., Bhaduri, K., Giannella, C., Wolff, R. and Kargupta, H., 2006. Distributed data mining in peer-to-peer networks, *IEEE Internet Computing* **10**(4), 18–26.

[6] (Ferri et al., 2004) Ferri, C., Flach, P. and Hernández-Orallo, J., 2004. Delegating classifiers, *ICML '04: Proceedings of the 21st International Conference on Machine learning*, ACM, New York, NY, USA, pp. 289–296.

[7] Abhinav Srivastava, Amlan Kundu, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar, Senior Member, IEEE "Credit Card Fraud Detection Using Hidden Markov Model" IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 5, NO. 1, JANUARY-MARCH 2008

[8] R. Grossman, S. Baily, S. Kasif, D. Mon, and A. Ramu. The preliminary design **of** papyrus: A system for high performance. In P. Chan H. Kargupta, editor, *Work. Notes KDD-98 Workshop on Distributed Data Mining*, pages 37–43. AAAI Press, 1998.

[9] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press, Menlo Park, California/Cambridge, Massachusetts / London, England, 1996.

[10] K.Lang. News weeder: Learning to filter net news. In A.Prieditis and S.Russel editors, *Proc.12th Intl. Conf. Machine Learning*, pages 331–339. Morgan Kaufmann, 1995.

[11] S. Stolfo,W. Fan,W. Lee, A. Prodromidis, and P. Chan. Credit card fraud detection using meta-learning: Issues and initial results. In *Working notes of AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*, 1997.

[12] P. Chan and S. Stolfo. Sharing learned models among remote database partitions by local meta-learning. In *Proc. Second Intl. Conf. Knowledge Discovery and Data Mining*, pages 2–7, 1996.

[13] D. Pomerleau. *Neural network perception for mobile robot guidance*. PhD thesis, School of Computer Sc., Carnegie Mellon Univ., Pittsburgh, PA, 1992. (Tech. Rep. CMU-CS-92-115).

[14] K.Mok W. Lee, S. Stolfo. Mining audit data to build intrusion models. In G. Piatetsky-Shapiro R Agrawal, P. Stolorz, editor, *Proc. Fourth Intl. Conf. Knowledge Discovery and Data Mining*, pages 66–72. AAAI Press, 1998.

[15] P. Chan and S. Stolfo. Meta-learning for multi strategy and parallel learning. In *Proc. SecondIntl. Work. Multistrategy Learning*, pages 150–165, 1993.

[16] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[18] R. Duda and P. Hart. *Pattern classification and scene analysis*. Wiley, New York, NY, 1973.

[19] Kargupta H., Park B., Hershberger D., Johnson E., *Collective Data Mining: A New Perspective Toward Distributed Data Analysis*. Accepted in the Advances in Distributed Data Mining, H. Kargupta and P. Chan (eds.), AAAI/MIT Press, (1999).

[20] Zhang X., Lam C., Cheung W.K., *Mining Local Data Sources For Learning Global Cluster Model Via Local Model Exchange*. IEEE Intelligence Informatics Bulletin, (4)2(2004).

[21] Prodromidis A., Chan P.K., Stolfo S.J., *Meta-learning in Distributed Data Mining Systems: Issues and Approaches*. In: Advances in Distributed and Parallel Knowledge Discovery, Kargupta H., Chan P.(ed.), AAAI/MIT Press, Chapter 3, (2000).

[22] Ahang S., Wu X., Zhang C., *Multi-Database Mining*. IEEE Computational Intelligence Bulletin, (2)1 (2003).