

Software Tool For Agent-Based Distributed Data Mining

Vladimir Gorodetsky
Head of Intelligent system Lab.
SPIIRAS, 39, 14-th Liniya
St. Petersburg, 199178, Russia
Ph. +7-812-3233570
gor@mail.iias.spb.su

Oleg Karsaeyv
Senior Researcher of Intelligent
system Lab.
SPIIRAS, 39, 14-th Liniya
St. Petersburg, 199178, Russia
Ph. +7-812-3233570
ok@mail.iias.spb.su

Vladimir Samoilov
Research Fellow of Intelligent
system Lab.
SPIIRAS, 39, 14-th Liniya
St. Petersburg, 199178, Russia
Ph. +7-812-3233570
samovl@mail.iias.spb.su

Abstract— *The paper scope is multi-agent technology and software tool for the joint engineering, implementation, deployment and, possibly, use of applied multi-agent distributed data mining and distributed decision making systems. The core problem of distributed data mining and decision making technology does not concern particular data mining techniques, because the respective library of classes can be extended when necessary. Instead of this, its core problem is development of an infrastructure and protocols supporting coherent collaborative work of distributed software components (agents) responsible for data mining and decision making. The paper is focused on architecture of multi-agent distributed data mining and decision making system, on its design technology, software tool and on the protocols of software tool agents' interaction, mainly, in distributed data mining and decision making processes. The presented software tool is implemented and validated on the basis of several case studies from data fusion scope.*

1. INTRODUCTION

Distributed Data Mining (DDM) aims at extraction useful pattern (rules, frequent patterns, association rules, etc.) from distributed data sets. Distributed decision making (in the paper—classification) aims at combining decisions produced by distributed solvers on the basis of data of particular sources. A lot of modern applications fall into the category dealing with distributed decision making. Applied tasks can be of different natures, for example, data and information fusion for situational awareness; scientific data mining in order to compose the results of diverse experiments, design a model of a phenomena, sensor data fusion, intrusion detection task, etc. One of the newly arisen tasks is learning of coordination in multi-agent systems (MAS). Web Intelligence area needs use of DDM for mining data distributed over the Internet, for example for mining marketing data.

From practical point of view, DDM problem is of great concern and ultimate urgency. Indeed, enormous number of databases accumulating experience in different areas of theory and practice constitute extremely rich, valuable and useful sources of knowledge that are still waiting to be discovered to enrich both science and industry. It is worthy to note that many data sources are either private or classified what excludes its centralized processing.

Recently, distributed data mining, in particular, for the purposes of distributed classification, which is based on using multiple classification models from distributed data sets, has become an active research area ([16], [1], [12]), [2], [3], [4], [11], [13], [14], [9], etc.). However, to date the prime attention in this area is paid to particular algorithms for distributed data mining and combining particular decisions produced on the basis of distributed data sources. Although DDM is becoming a critical technology, several important aspects of DDM, e.g. cooperation protocols of distributed software components both in distributed data mining and distributed decision making as well as new technologies like multi-agent one are about out of the scope of the current research. These aspects are in the focus of the paper. The paper primarily considers a DDM technology for the purposes of distributed classification with assumption that data sources are distributed and heterogeneous and possibly not available for centralized mining.

Distributed data used for producing decisions in such a decision making system can be of different physical nature (sensors' data, preprocessed data, experts' information, etc.) and of different accuracies. Particular data may be incomplete, uncertain and be measured in different scales. The most specific peculiarity of a DDM task is that each data source provides only a fragment of a phenomena specification or partial awareness, and DDM objective is to design distributed knowledge base and meta-level mechanism for combining of decisions produced on the basis of data fragments thus reconstructing, for example, a whole phenomena model of whole situational awareness.

The DDM area is actually challenging and puts a number of new problems resulting from distributed nature of both data to be mined and decision to be combined. Experience proved that the core problem of DDM does not concern particular data mining techniques. Instead of this, it concerns to the development of an infrastructure and protocols supporting collaborative operations of distributed software components (agents) responsible for DDM. Indeed, according to the recent understanding *distributed decision making system* is to be designed as a collection of distributed cooperative software agents, which operation has to be coordinated according to a number of protocols (distributed algorithms). On the other hand, *DDM component* destined for engineering of decision making system knowledge base and inference engine is also to be distributed and its components, training and testing agents, have also to coherently operate according to a number of protocols. Finally, according to the recent ideas *technology of MAS* and *supporting software tool* have to be capable to support distributed engineering of MAS applications and

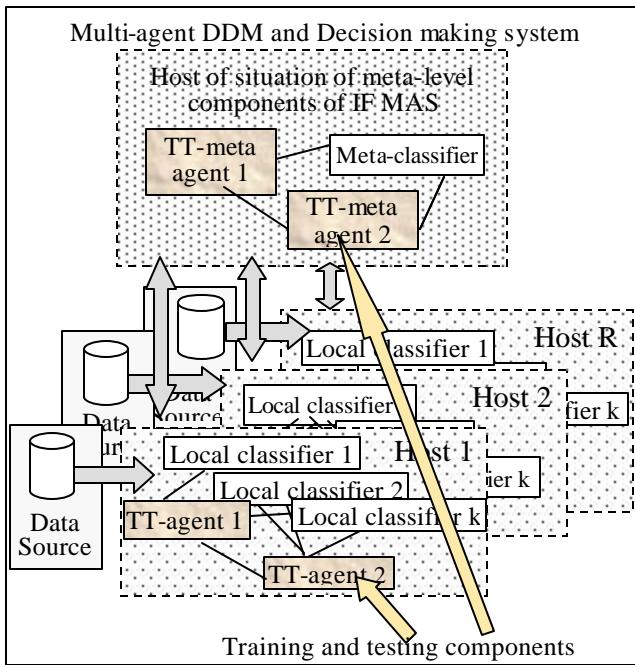


Figure 1. General view DDM and Decision making MAS architecture

therefore it is reasonable to implement the above technology as agent-mediated procedure.

The paper is focused on architecture of both multi-agent distributed data mining and decision making systems, on the agent-mediated MAS technology and software tool and on protocols of software tool agents' interaction, mainly, on distributed data mining and decision making protocols. It is organized as follows. In *section 2* we present general view of decision making MAS architecture provided with training and testing components responsible for DDM. In *section 3* the developed DDM technology is briefly presented. *Sections 4, 5* conceptually outline peculiarities of distributed ontology and meta-model of decision making design respectively. *Sections 6, 7* present the key ideas of DDM and distributed decision making and describe the respective protocols. *Section 8* outlines case studies used for the proposed technology validation. *Conclusion* summarizes the main results and future work.

2. ARCHITECTURE

In our development we consider multi-agent architecture of distributed decision making system. It comprises two types of components (Fig.1). The first type corresponds to the components operating with the source-based data and situated at the same hosts as the sources. The second type corresponds to a component operating with meta-information generated on the basis of source-based data and can be situated in any host. Each component includes classification agents (of local and meta-levels respectively) and training and testing agents (*TT-agents*) responsible for off-line supervised learning of DDM MAS.

A more detailed architecture of DDM MAS is depicted in Fig.2. The lower part depicts architecture and interactions of its source-based components while the upper one depicts architecture and interactions of its meta-level component.

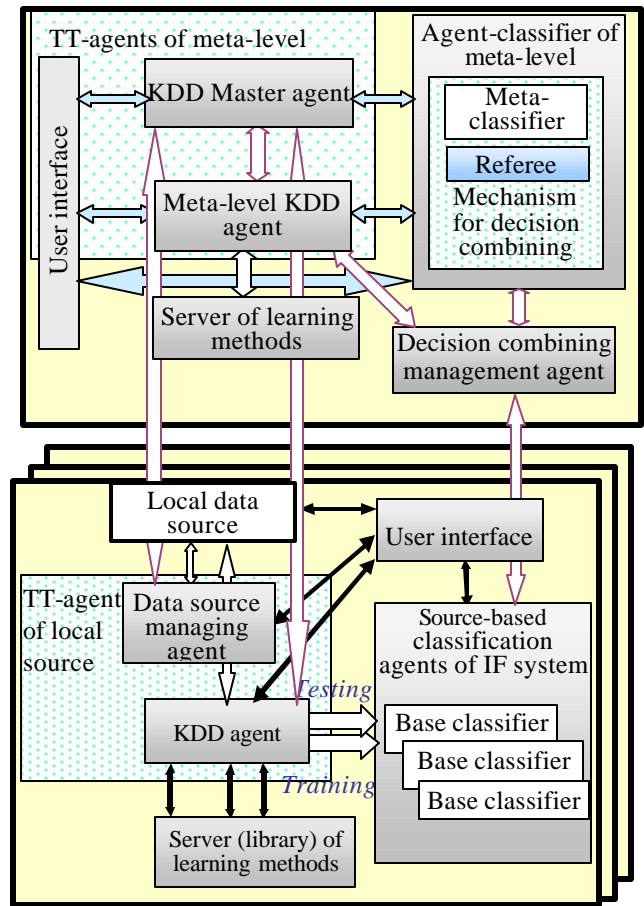


Figure 2. Architecture of meta-level (upper part) and source-based (lower part) components of DMM MAS

The source-based components of DDM MAS (Fig.2, lower part) and their functions are as described below.

Data source managing agent

- Participates in the distributed design of the shared component of the application ontology;
- Collaborates with meta-level agents in management of training and testing of particular source-based classifiers and in forming meta-data sample for meta-level training and testing;
- Supports gateway to databases through performing transformation of queries from the language used in ontology into SQL language.

KDD agent of data source

- Trains and tests of source-based classification agents and assesses the designed classifier's quality.

Classification agents of data source

- Produce decisions using source-based data. They are subjects of training performed by TT-agents.

Server of learning method (not an agent)

This component comprises a multitude of classes implementing particular KDD methods, metrics, etc. The set of classes is extendable.

The meta-level components of DDM MAS (Fig.2, upper part) and their functions are as described below.

Meta-Learning agent (“KDD Master”) – TT-agent

- Manages the distributed design of DDM MAS application ontology;
- Computes the training and testing meta-data sample;
- Manages design of meta-model of decision making.

Meta-level KDD agent

- Trains and tests of meta-level classification agent and assesses its quality.

Decision making management agent

- Coordinates operation of *Agent-classifier of meta-level* and *Meta-level KDD agent* both in training and decision combining modes of their performance.

Server (library) of KDD methods (not an agent)

This component comprises a multitude of classes implementing particular KDD methods, metrics, etc.

3. TECHNOLOGY FOR DDM MAS ENGINEERING

The developed technology for DDM MAS design uses

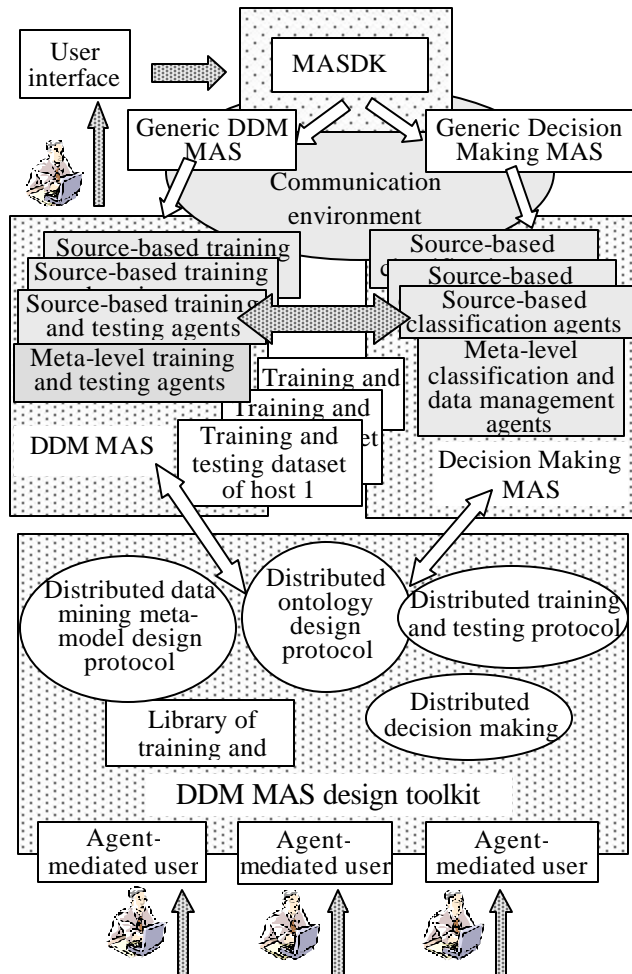


Figure 3. Explanation of the DDM MAS technology framework

software tool called *Multi-Agent System Development Kit (MASDK)* [6], which is used for the design of so-called *Generic DDM MAS*. The latter comprises agent instances supposed by the application architecture (Fig.2) and communication environment. On this phase the agents of DDM MAS are only provided with basic (reusable) functionalities supposed by MASDK platform. Deployment

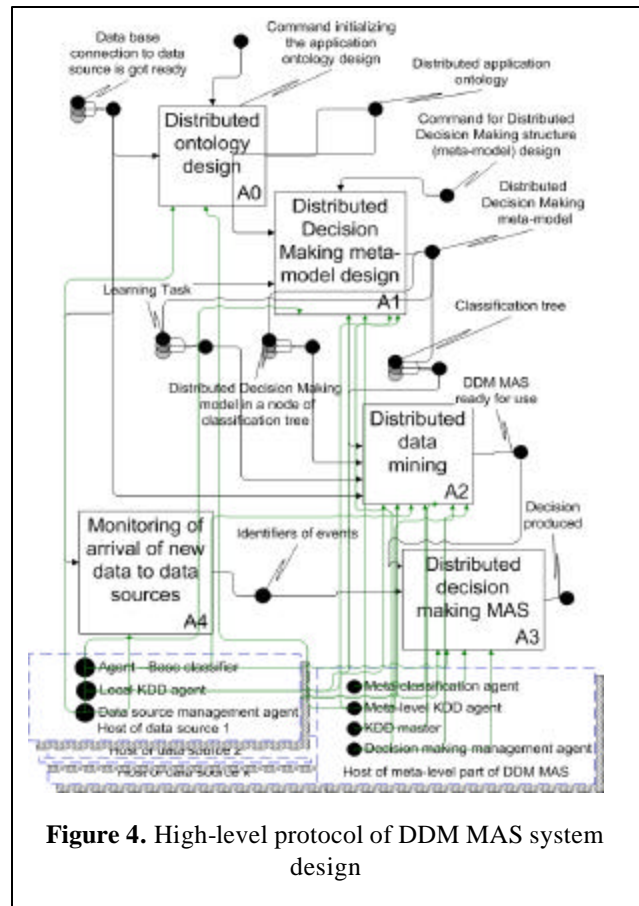


Figure 4. High-level protocol of DDM MAS system design

of the resulted DDM MAS is carried out within MASDK.

Next phase of DDM MAS design aims at specialization of the "start up" (empty) agents of *Generic DDM MAS* in order to tune them to the particular application. Fig.3 explains the hierarchy of users' activities on the specialization phase. Specialization is conducted by use of software components developed specifically for DDM MAS design. These components include a number of protocols, library of training and testing methods, and user interfaces supporting interactive and iterative distributed mode of DDM MAS specialization. The aforementioned components forms so-called DDM MAS design Toolkit. On this phase, the subjects of the design and specialization are ontology, message formats and contents, structure (meta-model) of distributed decision making (the latter also determines the structure of DDM), and training and testing functionalities selected for use in DDM.

It is supposed that DDM MAS is designed in distributed mode by several designers. Coordination of their activities is supported by a set of protocols. High-level view of the protocol of DDM MAS design, in which training of DDM MAS is a core procedure, is presented in Fig.4 in terms of standard IDEF0 diagrams. It comprises a number of processes (sub-protocols) that are as indicated below.

- A0. Distributed ontology design.
- A1. Decision making and DDM meta-model design.
- A2. Distributed data mining.
- A4. Monitoring of arrival of new data to data sources.
- A5. Distributed decision making.

This diagram specifies interaction of agents, intermediate and final results and activity ordering. The core of the technology is constituted by A0, A1 and A2 protocols, which are briefly explained below.

4. DESIGN OF DISTRIBUTED ONTOLOGY

The key DDM MAS peculiarities come out of the fact that data sources are *distributed* and can be *heterogeneous*. These features put problems strongly influencing on many issues of DDM MAS design. Use of *ontology* is now considered as the only approach to cope with the data distribution and heterogeneity problem.

In the design technology of applied MAS, the ontology design is considered as the first step. The developed protocol, A0, is intended to solve several specific problems [5]. The *first* of them is development of a *shared thesaurus* providing for *monosemantic understanding of the terminology* used in formal specification of domain entities. If data of sources are private then components of application ontology associated with the particular sources are designed independently. Therefore, the experts can denote different domain entities by the same name, and vice versa, they can denote the same entity differently what can lead to agents' misunderstanding. Next important problem comes out of the fact that the same entities can be represented in different sources in different data structures but in DDM procedures all of them have to be used equally. That is why it is necessary to provide distributed data *for consistent representation*. The next but not the last problem is so-called "*entity instance identification problem*" [5]. The data specifying an object is represented in several data sources and therefore each information source only partially specifies it. Object complete specification is made up of data fragments distributed over sources and to form a complete object specification, a mechanism to identify such fragments is needed.

The protocol of distributed design of distributed ontology is implemented as a component of DDM MAS Design Toolkit (see Fig. 3). The user-guided process of ontology design is also supported by use of a particular editor for ontology specification. The standard editors of such a kind are being developed in Semantic Web community [18]. As a rule, the ontology is specified in a standard language like XML, RDF, DAML+OIL. In our software tool the XML language is so far used.

5. DESIGN OF DECISION MAKING META-MODEL

Distributed classification supposes that within a problem many interacting classifiers participate in producing of decisions. In the developed architecture, decision making is organized as two-level procedure. The first level is responsible for producing classifications on the basis of particular data sources. On the second level, source-based classifications are combined according to an algorithm. The structure of distributed classifiers and their cooperation in

combining decision is called here *meta-model of distributed decision making*.

A reasonable choice of meta-model depends on many factors and as a rule this choice is made by users on the basis of personal experience and analysis of input data from many viewpoint, e.g. the sizes and dimensionalities of data of different sources, diversity of the data representation structures both within a source and in different sources, number of classes, mining algorithms at hand, etc. Actually, selection a meta-model of decision making is domain expert's responsibility, while software tool under consideration provides this expert with an instrument for such a meta-model design. The protocol A1 (Fig.4) supports this design.

It should be noted that meta-model of DDM entirely determines the structure and interaction of training tasks.

The protocol supporting meta-model design is developed in detail and implemented but we omit its specification due to the lack of the paper space.

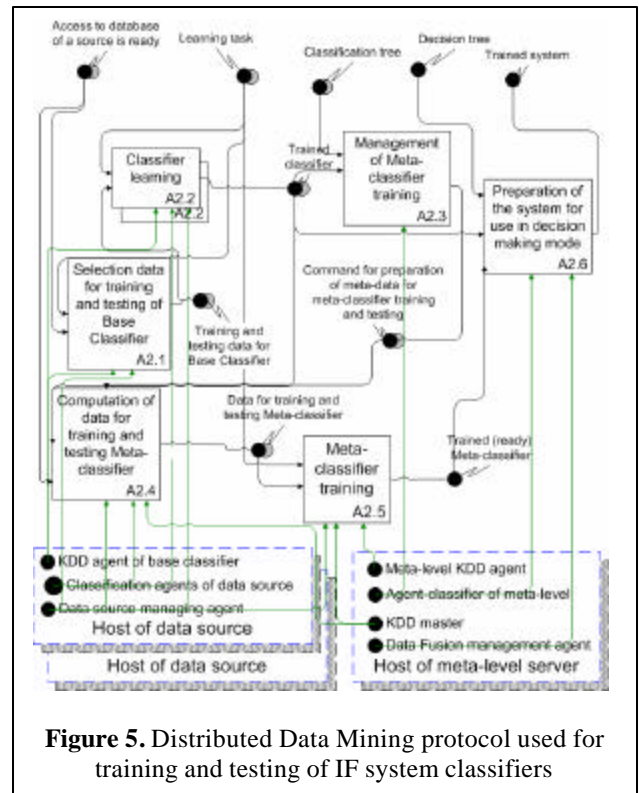


Figure 5. Distributed Data Mining protocol used for training and testing of IF system classifiers

6. DISTRIBUTED DATA MINING

Distributed data mining protocol, A3, that supports agents' collaboration in training and testing of particular classifiers and also manages decision combining is the core of DDM MAS technology. IDEF0 diagram of this protocol is presented in Fig.5.

It involves in interaction all agents of DDM MAS (Fig. 2). The basic processes constituting this protocol within the developed DDM training technology are as described below.

1. Selection of data sets for training and testing of the base classifiers (A2.1).
2. Training and testing of base classifiers (A2.2).

3. Meta-classifier training management (A2.3).
4. Computation of data for training and testing of meta-level classifier (A2.4).
5. Training and testing of meta-level classifier (A2.5).
6. Preparation of the DDM MAS system for use in decision making mode (A2.6).

The sub-protocols of the DDM constituting A2 protocol are specified in several levels of details up to the sub-processes that don't suppose distributed execution. This protocol is now implemented.

The set of techniques used for learning of base classifiers and meta-level classifier so far included into the library of training and testing methods comprises *VAM* algorithm (*Visual Analytical Mining*, [7]) for mining numerical data, *GK2* algorithm [8] for extraction rules from discrete data, and also *Frequent Pattern grows* algorithm [9] for mining association rules.

7. DISTRIBUTED CLASSIFICATION

This protocol that is one of the basic ones, determines interactions of the components of the applied decision making agents of the target MAS in producing decisions on the basis of data incoming to data sources.

The decision making process involves into interaction Data source managing agents and Agents-classifiers of data sources and also Decision making management agent and Agent-classifier of meta-level. IDEF0 diagram of this protocol is depicted in Fig.6.

In this protocol, the high level management is provided by Decision making management agent. It is responsible for monitoring of data incoming to sources, for forming

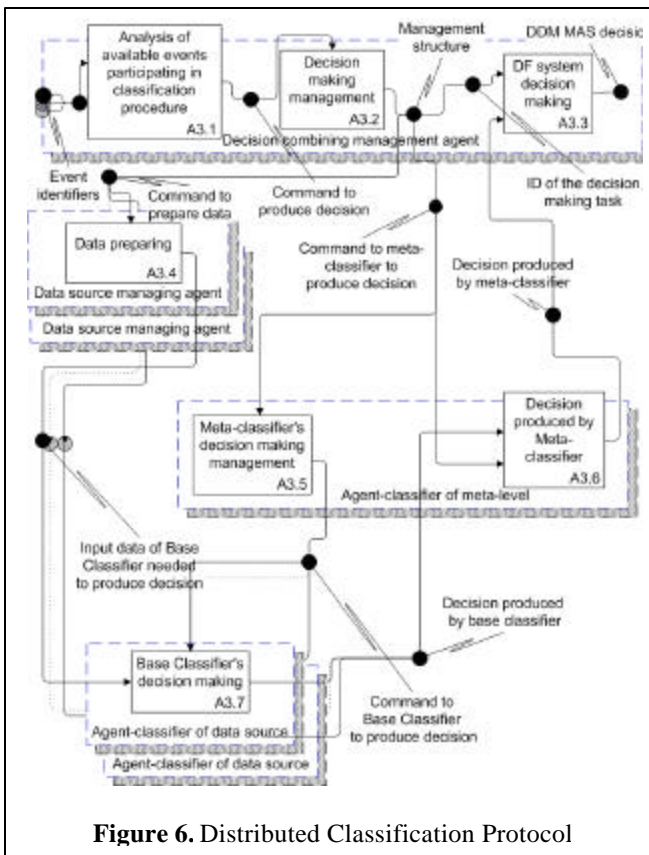


Figure 6. Distributed Classification Protocol

commands initiating decision making, for management of these processes and coordination of decision combining procedures. It implements three high-level processes:

1. Analysis of new data to be classified (A3.1);
2. Decision making management (A3.2), and
3. DDM MAS decision making (A3.3).

Data source management agents are responsible for carrying out of the process (4) "Data preparing" (A3.4), that produce input data for Agent-classifiers of the respective data sources. The latter also realize the processes (tasks) "Base Classifiers' decision making" (A3.7). The results of base-level classifications form the data needed for Agent-classifier of meta-level to execute its functions (processes) "Meta-classifier's decision making management" (A3.5) and "Making decision by Meta-classifier" (A3.6). The protocols A3.1–A3.7 are specified in more details and implemented as a component of DDM MAS Design Toolkit.

8. VALIDATION OF THE SOFTWARE TOOL

The preliminary important notice to this section is that case studies were used for validation of the technology and software tool, but not for validation of the properties of the resulting performance of the designed DDM and Decision making MAS. That is why the analysis of the accuracy of the performance of the resulted MAS is not the case here.

The basic ideas of DDM MAS technology were validated by its successful use in the engineering, implementation and deployment of two case studies. They are outlined below.

KDDCup-99 –based Case study of DDM MAS system

Application corresponding to KDDCup99 dataset [16] deals with Intrusion Detection Learning Task. Inherently this data are not distributed but it was split artificially to model multiple sources and to use the result as a case study. The KDDCup99 dataset is specified by 36 attributes of various types (numerical–28, categorical–4, Boolean–4), and the total size of data records *used in case study* (but not in dataset) is equal to 33460, at that $TT=7100$ of them were used for training and testing of base classifiers and meta-classifier and the rest, $FT=26360$, were used for evaluation of the accuracy of the developed decision making MAS. The data set TT was artificially split into two data sources, DS1 and DS2 (they share 1 Boolean attribute). In turn, DS1 and DS2 data sets were also split into 3 and 4 subsets respectively. The last splitting destined for forming of training and testing data for particular base classifiers and

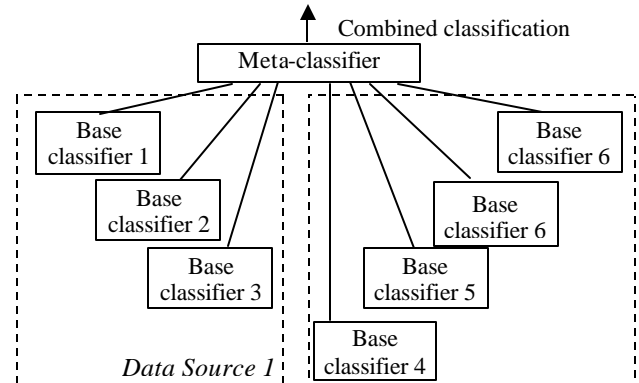


Figure 7. Classification structure in KDDCup99 case study

meta-classifier. The total number of base classifiers was chosen equal to 7 (3 of them were used in DS1 and 4-in DS2). The scheme of interaction of base classifiers and meta classifier is given in Fig.7. The base classifiers differ in attribute sets and also in training and testing data sets. The base classifiers also differ in training algorithms used. Two basic training algorithms were used in the case study: *Visual analytical Mining* ([8]), dealing with numerical data, and *GK2* ([9]) dealing with discrete data. Both algorithms and their software implementation were developed by authors.

This case study demonstrated correctness and feasibility of the main ideas of the technology and software tool.

Multi-Spectral Image classification

The second case study concerning multi-spectral image classification of Landsat Scanner image dataset ([17]) was also used for validation of the developed software tool.

CONCLUSION

The paper presents the developed multi-agent technology and software tool for distributed data mining and distributed decision making purposes. Design of both distributed data mining and decision making components puts several non-specific tasks and challenges. Some of them are in the focus of the paper. The key challenges come out of the fact that data sources are distributed, possibly heterogeneous and, as a rule, of large scale. Other important issue is that design technology destined for DDM and decision making supposes collaborative activities of distributed designers. Both these issues are challenging and to date are not explored in depth. The paper proposes solutions concerning the above issues. *First*, it proposes architecture of DDM MAS and respective software tool. *Second*, it proposes a number of protocols supporting both collaborative activity of designers and collaboration of agents of target applications during operation. Finally, what seems very important, it proposes the technology for distributed data mining dealing with distributed heterogeneous datasets. The main results of the paper concern analysis of these issues, proposals of respective solutions and their validation via design and implementation of two case studies.

Future research will be focused on its use for design of particular applications to accumulate experience and to make the DDM MAS software tool industry-oriented. The second direction of future research will be associated further development of the distributed agent-mediated software engineering of multi-agent systems.

ACKNOWLEDGEMENT

The work is supported by AFRL/IF and by Russian Foundation of Basic Research (grant #01-01-00109).

REFERENCES

[1] P.Chan, and S.Stolfo. Toward parallel and distributed learning by meta-learning. *Working Notes AAAI, KDD, AAAI*, 227-240, 1993.

[2] P.Chan and S.Stolfo. Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. *Working Notes of Distributed Data Mining Workshop in the 4 International Conference on Knowledge Discovery and Data Mining*, 1998.

[3] T.Dietterich. Machine Learning Research: Four Current Directions. *AI magazine*. 18(4), 97-136, 1997.

[4] J.Gama and P.Brazdil. Cascade generalization. *Machine Learning*, 41(3), 315-342, 2000.

[5] I.Goodman, R.Mahler, and H.Nguen. Mathematics of Data Fusion. Kluwer Academic Publishers, 1997.

[6] V.Gorodetski, O.Karsaev, I.Kotenko. Software Development Kit for Multi-agent Systems Design and Implementation. In B.Dunin -Keplicz and E.Nawareski (Eds.) "*From Theory to Practice in Multi-agent Systems*". *LNAI*, vol. 2296, Springer Verlag, 121-130, 2002.

[7] V.Gorodetski, V.Skormin, L.Popyack. Data Mining Technology for Failure Prognostics of Avionics, *IEEE Transactions on Aerospace and Electronic Systems*. Volume 38(2), 388-403, 2000

[8] V.Gorodetski and O.Karsayev. Algorithm of Rule Extraction from Learning Data. In *Proceedings of the 8th International Conference Expert Systems & Artificial Intelligence (EXPERTSYS-96)*, 133-138, 1996.

[9] J.Han, M.Kamber. Data Mining. Concept and Techniques. Morgan Kaufman Publishers, 2000.

[10] J.Ortega, M.Coppel, and S.Argamon. Arbitrating Among Competing Classifiers Using Learned Referees. *Knowledge and Information Systems*, 4, 470-490, 2001.

[11] A. Prodomidis, P. Chan, and S. Stolfo. Meta-learning in distributed data mining systems: Issues and approaches. In P.Chan and H.Kargupta (Eds.) *Advances in Distributed Data Mining*, AAAI Press, 1999. Available at <http://www.cs.columbia.edu/~sal/hpapers/DDMBOOK.ps.gz>

[12] F.Provost and D.Hennessy. Scaling up: Distributed machine learning with cooperation. *Working Notes of IMLM-96, AAAI-96 Workshop on Integrating Multiple Learned Models*, 107-112, 1996.

[13] K.Ting and I.Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10, 271-289, 1999.

[14] L.Todorovski and S.Dzeroski. Combining classifiers with meta decision trees. D.A.Zighen, J.Komarowski and J.Zitkov (Eds.) *Proceedings of 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-00)*, France, Springer Verlag, 54-64, 2000.

[15] D.Wolpert. "Stacked generalization". *Neural Network*, 5(2), 41-260, 1992.

[16] <http://kdd.ics.uci.edu/databases/kddcup99/>.

[17] <http://ics.uci.edu:pub/machine-learning-databases>.

[18] <http://www.semanticweb.org>.