

# Multi-agent Technology for Distributed Data Mining and Classification

**Vladimir Gorodetsky**  
SPIIRAS, St. Petersburg,  
Russia  
gor@mail.ias.spb.su

**Oleg Karsaev**  
SPIIRAS, St. Petersburg,  
Russia  
ok@mail.ias.spb.su

**Vladimir Samoilov**  
SPIIRAS, St. Petersburg,  
Russia  
samovl@mail.ias.spb.su

## Abstract

*The core problem of multi-agent distributed data mining technology not concern particular data mining techniques although the latter is now paid the most attention. Its core problem concerns collaborative work of distributed software in design of multi-agent system destined for distributed data mining and classification. The paper presents the developed and implemented distributed data mining technology, architecture of the multi-agent software tool supporting this technology and demonstrates the key protocols used by agents in collaborative design of an applied multi-agent distributed data mining system.*

## 1. Introduction

Distributed Data Mining (DDM) aims at extraction useful pattern from distributed heterogeneous data bases in order, for example, to compose them within a distributed knowledge base and use for the purposes of decision making. A lot of modern applications fall into the category of systems that need DDM supporting distributed decision making. Applications can be of different natures and from different scopes, for example, data and information fusion for situational awareness; scientific data mining in order to compose the results of diverse experiments and design a model of a phenomena, intrusion detection, analysis, prognosis and handling of natural and man-caused disaster to prevent their catastrophic development, Web mining, etc. From practical point of view, DDM is of great concern and ultimate urgency.

Recently, distributed data mining, in particular, for the purposes of distributed classification, has attended active research ([8], [1], [7], [9], [10], [5], [6], etc.). However, in this research the prime attention is to date paid to the algorithmic aspects of distributed data mining and combining decisions. At that, important issue concerning cooperation protocols of distributed software components

both in DDM and distributed classification as well as use of new technologies like multi-agent one is paid smaller attention. These aspects are in the focus of the paper. It primarily considers agent-based DDM technology and protocols of agents' collaboration in DDM and classification for the case if data sources are distributed, heterogeneous and can be not available for centralized mining, but, DDM techniques are out of the paper scope.

In the rest of the paper *Section 2* presents general view of architectures of multi-agent DDM and associated classification systems to be designed. *Section 3* outlines briefly the developed DDM technology. *Sections 4* outlines protocols for the design of meta-model of DDM and presents the key ideas and the protocols for DDM and classification. *Conclusion* summarizes the main results.

## 2. Architecture of distributed data mining

We use multi-agent system (MAS) architectures for both DDM and Distributed Classification (DC) systems. Both of them comprise two types of components (Fig.1). The first of them unites the components handling the source-based data and tasks; they operate in the same hosts as respective data sources. The second one is composed of components handling meta-data. The latter can operate in any host. In Fig.1 the DDM MAS components are given in the left hand part whereas DC MAS components are given in the right hand one.

Let us consider the source-based components of DDM and DC MASs (Fig.1, lower part) and their functions.

### *Data source managing agent*

- Participates in the distributed design of the consistent shared by DDM and DC MAS components of the application ontology;
- Collaborates with meta-level agents in training and testing procedures of source-based classifiers and in forming meta-data for meta-level training and testing;
- Provides gateway to databases through performing transformation of queries from the language used in ontology into SQL language.

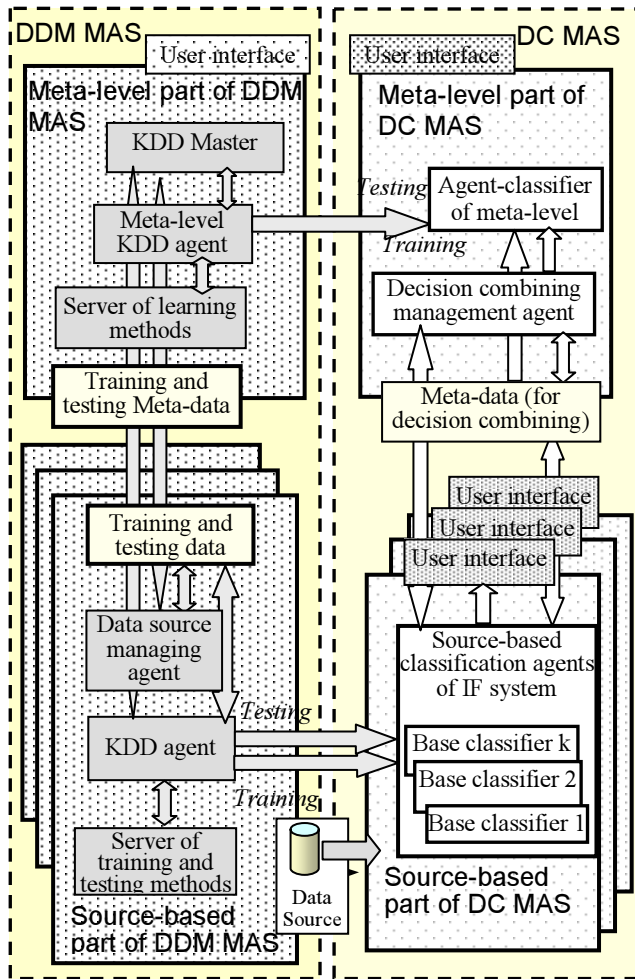


Fig.1. Architecture of DDM (left) and DC MAS (right)

*KDD agent of data source*

- Trains and tests of source-based classification agents;

*Classification agents of data source (part of DC MAS)*

Produce decisions using source-based information. They are subjects of training and testing.

*Server of training and testing methods (not an agent)*

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

The meta-level components of DDM MAS (Fig.1, upper part) and their functions are as follows:

*Meta-Learning agent ("KDD Master")*

- Manages the distributed design of common DDM and DC 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.

*Decision combining management agent*

- Coordinates operation of *Agent-classifier of meta-level* and *Meta-level KDD agent*.

*Agent-classifier of meta-level*

- It is subject of training and testing performed by *Meta-level KDD agent* of DDM MAS.

### 3. Technology of DDM and DC MASs design

The developed technology uses platform called *Multi-Agent System Development Kit*, MASDK ([6]), which is used for the design of so-called *Generic DDM MAS* and *Generic DC MAS*. Both of them comprise agents supposed by the application architecture (Fig.1) situated within a communication environment. These agents are only provided with basic (reusable) components supposed by MASDK platform deployed within a computer network.

Next phase of DDM and DC MASs design aims at specialization of the "start up" ("empty") agents of the deployed *Generic DDM* and *DC MASs* in order to tune them to the particular application. Fig.2 explains activities on this specialization phase. Specialization is conducted by use of so-called *Distributed Data Mining Design Toolkit* that comprises software components developed specifically for DDM and DC MASs design. These components include a number of protocols, library of training and testing methods, and users' interfaces supporting interactive and iterative mode of DDM and DC MASs specialization. On this phase, the subjects of the

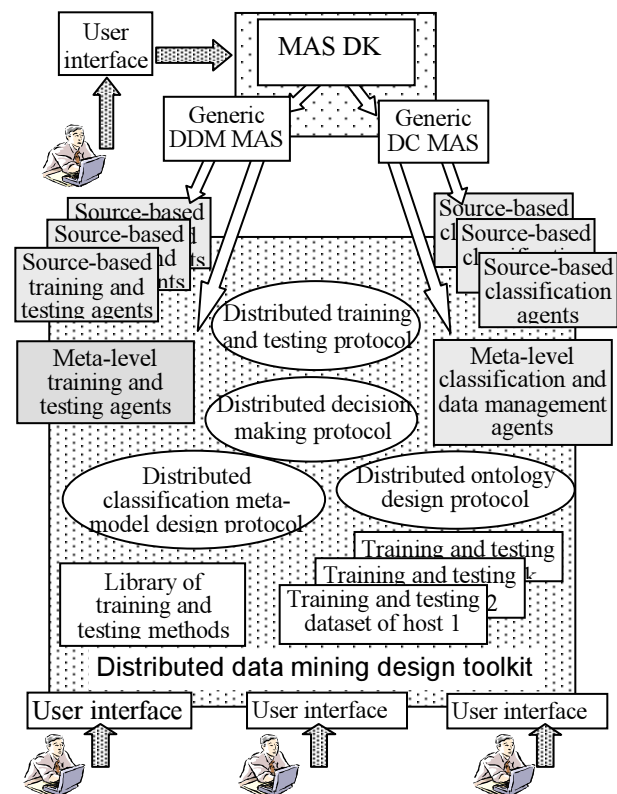


Fig.2. Explanation of the technology and tool kits used for the design of DDM and DC MAS

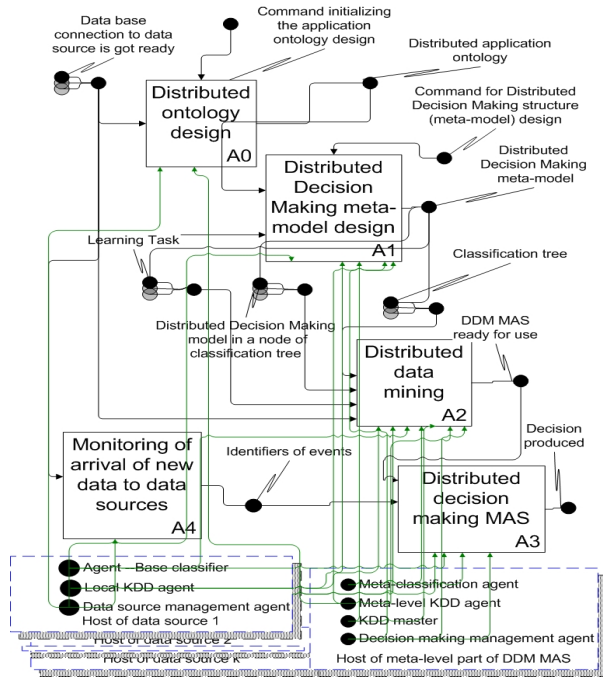


Fig. 3. High-level protocol of DDM and DC MAS design

design and specialization are (1) shared ontology of multi-agent systems, (2) DDM and DC MAS communication components specifying message formats and contents, (3) structure (meta-model) of DDM and DC procedures; (4) procedures associated with distributed training and testing of DC system performed by DDM MAS.

It is supposed that DDM and DC MAS are designed in agent-mediated distributed mode by team of designers. Coordination of their activities has to be supported by a set of agent interaction protocols. High-level view of the protocol of DDM and DC MASs specialization, in which training of DDM MAS is a core procedure, is given in Fig.4 in terms of standard IDEF0 diagrams. The core of the technology is constituted by A0, A1, A2 and A3 protocols. Some of them are briefly explained below.

#### 4. Basic DDM and DC protocols

Distributed classification supposes that several interacting classifiers participate in producing decision. In the developed architecture, classification 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.

In the developed protocol, the strategy based on use of meta-agent responsible for combining decision is applied. It supposes agent-mediated negotiations of DDM and DC MASs designers destined for development of two

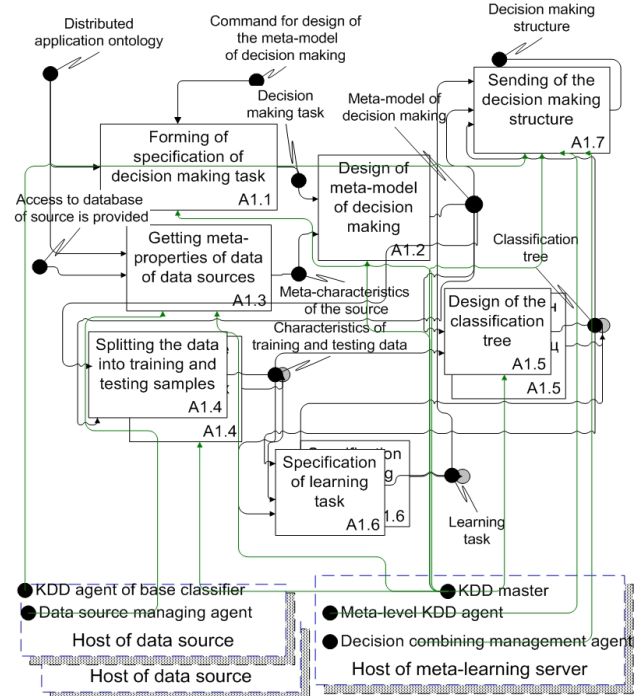


Fig. 4. Protocol for the design of data mining and decision making meta-model

structures. The first is so-called "*Classification tree*" specifying multi-class classifications task (if the number of classes is more than 2) in terms of binary ("pair-wise") classification. The second structure designed through agent-mediated negotiations of designers according to a protocol is one specifying how the local decisions are combined in meta level to produce the final one. It is called hereinafter "*Decision making structure*". An important role of "*Classification tree*" and "*Decision making structure*" is that they make it possible on this design step to abstract from concrete data mining used for training and also from particular formal specifications of node contents of both aforementioned trees and to implement the protocol in question as invariant component of Distributed Data Mining Design Toolkit (see Fig.2).

A reasonable choice of meta-model in question 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 sources, data representation heterogeneity, mining algorithms at hand, etc. Selection a meta-model of decision making is a users' responsibility, whereas software tool under consideration has to provide users with an agent-mediated instrument for the design of such meta-models. The protocol A1 (see Fig.3 and also Fig.4). supports this design.

Distributed data mining protocol, A3, that supports agents' collaboration in training and testing of particular classifiers and also manages decision combining is the

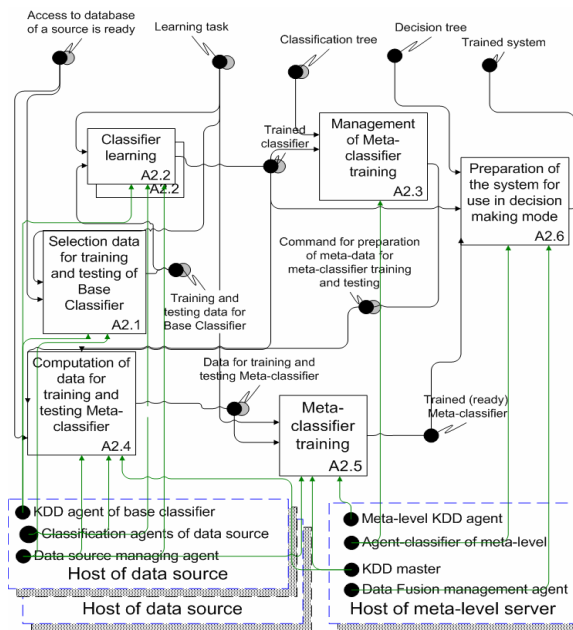


Fig.5. Distributed data mining protocol used for training and testing of DC MAS classifiers

core of DDM MAS technology. IDEFO diagram of this protocol is presented in Fig.5.

It involves in interaction all agents of DDM MAS (Fig. 1). The basic processes realizing this protocol and corresponding to the developed technology of DDM MAS training are as follows:

1. Selection of data sets for training and testing of 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 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 protocol A2 are specified on several levels of details up to the sub-processes that don't suppose distributed execution.

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

## 5. Conclusion

The paper presents the developed *multi-agent technology for DDM and DC*. Design of both DDM and DC systems puts several new non-specific tasks and challenges. Some key problems are in the paper focus. The

key problems come out of the fact that data sources are distributed, heterogeneous and, as a rule, of large scale. Other important issue is that design technology of DDM MAS supposes *collaborative activities of agent-mediated distributed users*. The paper proposes solutions concerning the above issues. *First*, it proposes *architecture* of DDM and DC MASs and *technology* for their design. *Second*, it proposes a number of well-developed *protocols* supporting agent-mediated design of applied DDM and DC MASs. The developed technology was validated on the basis several case studies, for example, on the basis of KDDCup99 data set ([11]). It is very important to note that the presented technology constitutes the conceptual basis of the software tool implemented which is implemented and practically used for rapid prototyping of the applied DDM and DC MAS.

Future research will be focused on accumulation of the experience of use this technology and software tool on the basis of design and implementation of particular DDM and DC applications.

## Acknowledgement

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

## References

- [1] T.Dietterich. Machine Learning Research: Four Current Directions. *AI magazine*. 18(4), 1997, 97-136.
- [2] V.Gorodetski, O.Karsaev, I.Kotenko. Software Development Kit for Multi-agent Systems Design and Implementation. In *"From Theory to Practice in Multi-agent Systems"*. LNAI, vol. 2296, Springer Verlag, 2002, 121-130.
- [3] V.Gorodetski, V.Skormin, L.Popyack. Data Mining for Failure Prognostics of Avionics, *IEEE Transactions on Aerospace and Electronic Systems*. 38(2), 2000, 388-403.
- [4] V.Gorodetski and O.Karsayev. Algorithm of Rule Extraction from Learning Data. In *Proceedings of International Conference on Expert Systems & Artificial Intelligence*, Paris, 1996, 133-138.
- [5] J.Han, M.Kamber. Data Mining. Concept and Techniques. Morgan Kaufman Publishers, 2000.
- [6] J.Ortega, M.Coppel, and S.Argamon. Arbitrating Among Competing Classifiers Using Learned Referees. *Knowledge and Information Systems*, 4, 2001, 470-490.
- [7] A. Prodromidis, P. Chan, and S. Stolfo. Meta-learning in distributed data mining systems: Issues and approaches. *Advances in Distributed Data Mining*, AAAI Press, 1999.
- [8] F.Provost and D.Hennessy. Scaling up: Distributed machine learning with cooperation. *Working Notes of IMLM-96*, 1996, 107-112.
- [9] K.Ting and I.Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10, 1999, 271-289.
- [10] L.Todorovski and S.Dzeroski. Combining classifiers with meta decision trees. *Proceedings of 4<sup>th</sup> European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-00)*, Springer Verlag, 2000, 54-64.
- [11] <http://kdd.ics.uci.edu/databases/kddcup99/>.