

# Issues of Agent-Based Distributed Data Mining

Matthias Klusch  
Deduction and Multiagent  
Systems  
German Research Centre for  
Artificial Intelligence  
Stuhlsatzenhausweg 3  
66123 Saarbruecken,  
Germany  
klusch@dfki.de

Stefano Lodi  
Department of Electronics,  
Computer Science and  
Systems  
IEIIT-BO/CNR  
University of Bologna  
Viale Risorgimento 2  
40136 Bologna BO, Italy  
slodi@deis.unibo.it

Gianluca Moro  
Department of Electronics,  
Computer Science and  
Systems  
University of Bologna  
Via Rasi e Spinelli 176  
47023 Cesena FC, Italy  
gmoro@deis.unibo.it

## Categories and Subject Descriptors

I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—*Multiagent Systems*; I.5.3 [Pattern Recognition]: Clustering—*Algorithms*

## General Terms

Algorithms, Theory

## Keywords

Data Mining, Knowledge Discovery, Clustering, Agents

## 1. INTRODUCTION

The increasing demand to scale up to massive data sets inherently distributed over a network with limited bandwidth and computational resources available motivated the development of *distributed data mining* (DDM). DDM is expected to perform partial analysis of data at individual sites and then to send the outcome as partial result to other sites where it is sometimes required to be aggregated to the global result. Quite a number of DDM solutions are available using various techniques such as distributed association rules, distributed clustering, Bayesian learning, classification (regression), and compression, but only a few of them make use of intelligent agents at all.

The main problems any approach to DDM is challenged to cope with concern not only with scalability but also with issues of autonomy and privacy. For example, when data can be viewed at the data warehouse [4] from many different perspectives and at different levels of abstraction, it may threaten the goal of protecting individual data and guarding against invasion of privacy. These issues of privacy and autonomy become particularly important in business application scenarios where, for example, different (often competing) companies may want to collaborate for fraud detection but without sharing their individual customers' data or disclosing it to third parties. One lesson from the recent research work on DDM is that cooperation among distributed DM processes may allow effective mining even without centralised control. This in turn leads us to the question

whether there is any real added value of using concepts from agent technology [1] for the development of advanced DDM systems.

## 2. WHY AGENTS FOR DDM

Considering the most prominent and representative agent-based DDM systems to date: BODHI, PADMA, JAM, and Papyrus (details in [2]), we may identify the following arguments in favor or against the use of intelligent agents for distributed data mining.

*Autonomy of data sources.* A DM agent may be considered as a modular extension of a data management system to deliberately handle the access to the underlying data source in accordance with given constraints on the required autonomy of the system, data and model. This is in full compliance with the paradigm of cooperative information systems.

*Interactive DDM.* Pro-actively assisting agents may drastically limit the amount a human user has to supervise and interfere with the running data mining process, e.g., DM agents may anticipate the individual limits of the potentially large search space and proper intermediate results.

*Dynamic selection of sources and data gathering.* In open multi-source environments DM agents may be applied to adaptively select data sources according to given criterias such as the expected amount, type and quality of data at the considered source, actual network and DM server load.

*Scalability of DM to massive distributed data.* A set of DM agents allow for a divide-and-conquer approach by performing mining tasks locally to each of the data sites. DM agents aggregate relevant pre-selected data to their originating server for further processing and may evaluate the best strategy between working remotely or migrating on data sources. Experiments in using mobile information filtering agents in distributed data environments are encouraging [6].

*Multi-strategy DDM.* DM agents may learn in due course of their deliberative actions which combination of multiple data mining techniques to choose depending on the type of data retrieved from different sites and mining tasks to be pursued. The learning of multi-strategy selection of DM methods is similar to the adaptive selection of coordination strategies in a multi-agent system as proposed, for example, in [5].

*Security.* Any failure to implement least privilege at a

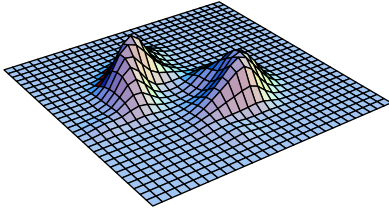


Figure 1: A density estimate

data source could give any mining agent unsolicited access to sensitive data. Agent code and data integrity is a crucial issue in secure DDM: Subverting or hijacking a DM agent places a trusted piece of (mobile) software—thus any sensitive data carried or transmitted by the agent—under the control of an intruder. If DM agents are even allowed to migrate to remote computing environments methods to ensure authentication and confidentiality of a mobile agent have to be applied. Finally, selective agent replications may help to prevent malicious hosts from simply blocking or destroying the temporarily residing DM agents.

*Trustworthiness.* DM agents may infer sensitive information even from partial integration to a certain extent and with some probability. This problem, known as the so called inference problem, occurs especially in settings where agents may access data sources across trust boundaries which enable them to integrate implicit knowledge from different sources using commonly held rules of thumb. The inference problem is still under study as an independent thread and not any of the existing DDM systems, agent-based or not, cope with it.

### 3. DISTRIBUTED DATA CLUSTERING

*Data clustering* is the task of partitioning a multivariate data set into groups maximizing intra-group similarity and inter-group dissimilarity. In a distributed environment, it is usually required that data objects are not transmitted between sites for efficiency and security reasons. An approach to clustering exploits the local maxima of a *density estimate* (d.e.) to search for connected regions which are populated by similar data objects. In [3], a scheme for distributed clustering based on d.e. has been proposed, which we briefly recall. Every participating site computes a d.e. based on its local data only. Then, every site applies information-theoretic *regular multi-dimensional sampling* to generate a finite, discrete, and approximate representation of the d.e., consisting of its values at a finite number of equidistantly spaced locations. The samples computed by all sites are transmitted and summed (by location) outside the originating site, e.g., at a distinguished helper site. The resulting list of samples, which is an approximate representation of the true global d.e., is transmitted to each participating site. Every site executes a density-based clustering algorithm to cluster its local data with respect to the global d.e., the values of which can be computed from the samples by means of a *sampling series*. Notice that a d.e. is not a band-limited function, therefore sampling produces aliasing errors, which increase as the number of samples decreases. Figure 1 and

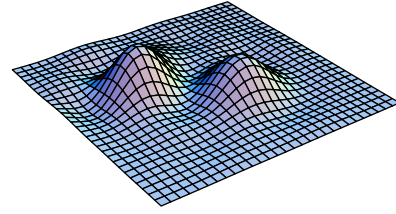


Figure 2: Reconstructed density estimate

Figure 2 represent the global d.e. of a test data set, and the corresponding d.e. obtained as a sampling series, respectively. The size of the list of samples in bytes was 25% the size of the data set in bytes. Although spurious undulations are present, the structure of the local maxima of the d.e. in the region populated by data objects did not change.

We propose to implement the approach by a society of agents. For example, in a real scenario all participating agents belong to different competing organizations, which agree to cooperate in order to achieve some common goal, without disclosing the contents of their data banks to each other. Each agent will negotiate with other agents to evaluate the advantages and risks which derive from participating to the distributed mining task. In particular, considerable security risks arise from the potential ability of the other agents to carry out inference attacks on density estimates. The resulting disclosure of sensitive information could be exploited as a competitive advantage by the organizations which own the malicious agents. Other aspects an agent has to evaluate in order to autonomously decide whether it should participate or not, include, but are not limited to, investigating a probabilistic model of trustworthiness of participating agents, the relation between trustworthiness and the topology of participating agents, and the probability of incurring coalition attacks.

### 4. REFERENCES

- [1] M. Klusch. Information agent technology for the internet: A survey. *Data and Knowledge Engineering. Elsevier Science*, 36(3):337–372, 2001.
- [2] M. Klusch, S. Lodi, and G. Moro. Agent-based distributed data mining: The kdec scheme. In *AgentLink*, volume 2586 of *LNCS*. Springer, 2003.
- [3] M. Klusch, S. Lodi, and G. Moro. Distributed clustering based on sampling local density estimates. In *Proc. of 18th IJCAI (IJCAI-03)*, Acapulco, Mexico, August 2003. To appear.
- [4] G. Moro and C. Sartori. Incremental maintenance of multi-source views. In *Proceedings of 12th ADC, Brisbane, Queensland, Australia*, pages 13–20. IEEE Computer Society, February 2001.
- [5] M. Prasad and V. Lesser. Learning situation-specific coordinating in cooperative multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1999.
- [6] W. Theilmann and K. Rothermel. Disseminating mobile agents for distributed information filtering. In *Proc. of 1st Sympos. on Mobile Agents*, pages 152–161. IEEE Press, 1999.