A Comparative study of Multi Agent Based and High-Performance Privacy Preserving Data Mining

Md Faizan Farooqui
Sr. Lecturer, Department of Computer Sc and Application, Integral University, Lucknow

Md Muqeem
Sr. Lecturer, Department of Computer Sc and Application, Integral University, Lucknow

Dr. Md Rizwan Beg
Professor, Department of Computer Sc & Engg, Integral University, Lucknow

ABSTRACT

Data mining is an extraordinarily demanding field referring to extraction of implicit knowledge and relationships, which are not explicitly stored in databases. Agent paradigm presents a new way of conception and realizing of data mining system. The purpose is to combine different algorithms of data mining to prepare elements for decision-makers, benefiting from the possibilities offered by the multi-agent systems. While the emerging field of privacy preserving data mining (PPDM) will enable many new data mining applications, it suffers from several practical difficulties. PPDM algorithms are difficult to develop and computationally intensive to execute. Developers need convenient abstractions to reduce the costs of engineering PPDM applications. The individual parties involved in the data mining process need a way to bring high-performance, parallel computers to bear on the computationally intensive parts of the PPDM tasks. This paper discusses the comparative study between multi agent based data mining and high-performance privacy preserving data mining. This paper offers a detailed analysis of the agent framework for data mining and its overall architecture and functionality are presented and also challenges in developing PPDM algorithms with existing frameworks, and motivates the design of a new infrastructure based on these challenges.

General Terms

Data Mining

Keywords

Privacy-Preserving Data Mining, Distributed Data Mining, Cluster Computing , multi-agent.

1. INTRODUCTION

The databases and data warehouses become more and more popular and imply huge amount of data which need to be efficiently analyzed. Knowledge Discovery in Databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases [1][2]. Agents' technology has proven to be particularly useful in several applications particularly when they imply managing user profiles. In data mining the sharing of data creates many potential privacy problems. Many organizations have restrictions on data sharing. Instead of dispensing entirely with cooperative data mining, research has instead focused on Privacy Preserving Data Mining (PPDM), which uses various techniques, statistical, cryptographic and others, to facilitate cooperative data mining while protecting the privacy of the organizations or individuals involved. In addition, because complex computation is often required, high performance and parallel computing technologies are necessary for efficient operation, adding yet another level of complexity to development. Furthermore, the system seamlessly integrates high performance computing technologies, to ensure an efficient data mining process. While there is much research that discusses available algorithms and techniques in PPDM, few studies focus on high performance computational architectures that support them.

2. FRAMEWORK FOR DATA MINING BASED MULTI-AGENT

A Multi-Agents System (MAS) consists of processes proceeding at the same time, therefore several agents living at the same time, sharing common resources and communicating between them [3]. The MAS must respect the standards of programming defined by the FIPA i.e., Foundation for Intelligent Physical Agents. We first discuss a generic architecture of data mining framework based multi-agent. Then we study the architecture of the Data Mining System based multi-Agent.

2.1 Generic Architecture of Data Mining Framework

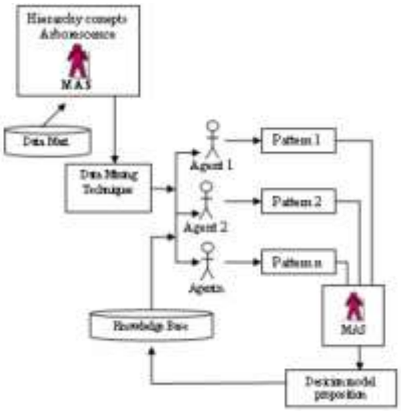


Fig. 1 Generic architecture of data mining framework based multi-agent [19]

The figure 1 describes the data mining frame work based multi agent. The MAS will be present, first, in the level of concepts hierarchy definition. After, when the resulting models from the data mining process are elaborated, the MAS is used to present the best adapted decision to the user. The proposition of decision-aid is stored in knowledge base, which could be useful in a later decision-making. Thus, it must be available among the agents which are in the entry of MAS. The identification of the agents will be done in the following section. Different methods of data mining have been proposed. Each method could have more than one algorithm. During the KDD process applied on geographic data, the user can be opposed to a problem of choosing between methods or even between algorithms. The conception of a system of data mining based multi-agent begins by the description of agents and the interaction between them [19].

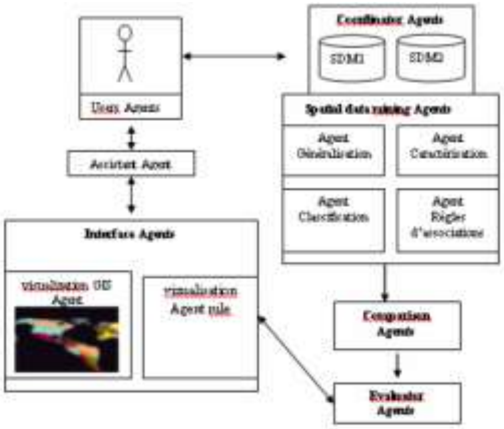


Fig. 2 General architecture of Data Mining System based multi-Agent [19].

2.2 General Architecture of Data Mining System based multi-Agent

This model makes it possible to take into account the various elements and to estimate the consequences of the various actions undertaken by the simulation or the construction of scenarios. A system based on such a model allows the evaluation of the strategic actions and consequently the anticipation of the phenomenon and the determination of the adequate strategy. The field of data mining shows characteristics favorable to a modeling multi-agent: a system of which the browsing by traditional methods leads to too large combinative. Thus the approach multi-agent is adopted. The various agents identified on this level are the following: an evaluator agent, comparison agent, a coordinator agent, one or more agents for data mining and an interface agent [19]. The general architecture of Data Mining System based multi-Agent is given by figure 2 which illustrates the interaction between the described agents.

3. ARCHITECTURE FOR PRIVATE AND HIGH PERFORMANCE INTEGRATED DATA MINING

The availability of frameworks, simple development environments are lacking, it is especially difficult to integrate the Privacy preserving data mining level of mining with the use of local high-performance computing resources (e.g. grid, clusters and specialized hardware). Architecture for Private and High-performance Integrated Data mining seeks to overcome these imitations. The design is influenced by several desired data, which have been explicitly identified in the literature or found lacking in other systems. Architecture for Private and High Performance Integrated Data Mining is explicitly built on a two-tier system of Privacy preserving data mining, which differentiates it from other systems. On the first tier, different organizations also called parties in the PPDM context communicate with each other, typically using secure, privacy preserving communications. The second tier includes grids and clusters within a particular party. Treating these tiers distinctly helps the developer to manage the complexities Inherent in each level [20].

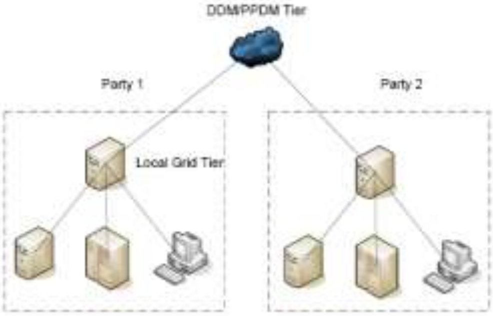


Fig. 3. A two tier PPDM architecture [20].

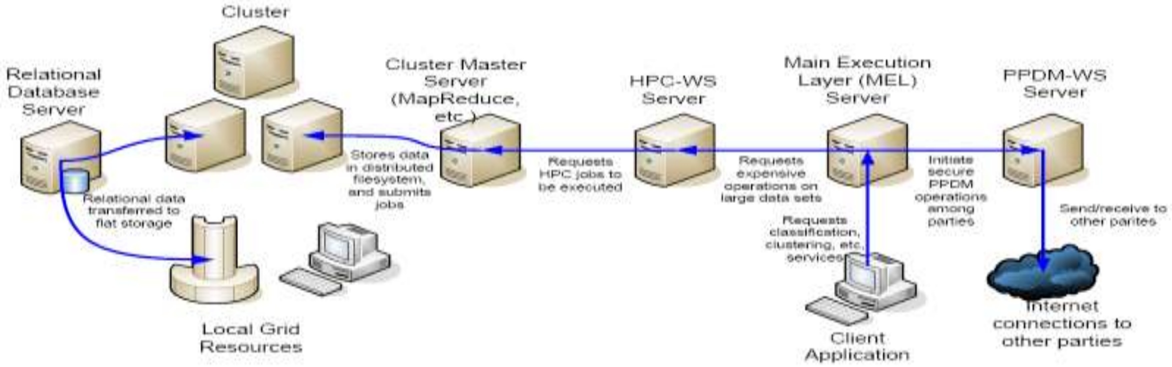

Cluster
Relational Database Server
Cluster Master Server (MapReduce, etc.)
HPC-WS Server
Main Execution Layer (MEL) Server
PPDM-WS Server
Stores data in distributed filesystem, and submits jobs
Requests HPC jobs to be executed
Requests expensive operations on large data sets
Initiate secure PPDM operations among parties
Send/receive to other parites
Relational data transferred to flat storage
Requests classification, clustering, etc. services
Local Grid Resources
Client Application
Internet connections to other parties

Fig. 4. A typical Architecture for Private and High Performance Integrated Data Mining system stack [20].

Figure 4 shows the stack of systems comprising a typical Architecture for Private and High Performance Integrated Data Mining installation within a single party. Organizational data to be mined is frequently stored in a relational database server. Because a relational database manager is typically insufficient for flexible data mining, and because these servers are often intimately involved in core business processes, this data is converted and transferred to a high-performance distributed file system (e.g. HDFS [9], or grid-based storage). The PPDM process begins with a request from a client, typically as part of a larger application, for the output of a specific PPDM algorithm (e.g. a classified test point, a classifier model, a set of clusters, or a set of association rules). The algorithms are available by unique services representing the algorithm, a partitioning (vertical, horizontal, or arbitrary) and a specific implementation. The PPDM services layer and the High-Performance Computing (HPC) respectively support these needs. The PPDM services layer acts as a gateway to external parties. The HPC services layer is a generic interface that interacts with a pluggable set of cluster and grid runtime systems (e.g. MapReduce) to perform the local mining of the database which will become part of the larger PPDM algorithm. It will store and access training databases, and submit compute intensive jobs through the appropriate channels. Having these broad collections of service-based functions available, meets Architecture for Private and High Performance Integrated Data Mining's requirements for flexibility [20].

3.1 Components of Architecture for Private and High Performance Integrated Data Mining

Each layer of Architecture for Private and High Performance Integrated Data Mining a development model must first be established to provide a simple yet powerful abstraction for PPDM development.

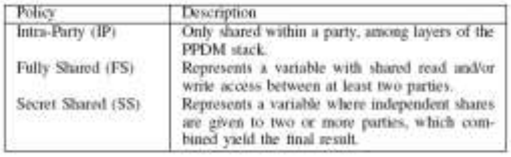
3.1.1 Program Structure

In order to bridge computations on a grid or cluster with DDM/PPDM computations, a simplified interface is needed.

3.1.2. Shared Variables

One technique APHID employs to simplify PPDM application development is the use of shared variables. These variables work similarly to those of traditional shared memory systems, with the exception that they have a particular policy attached.

Table 1. Types of variable sharing policies [20].



Policy	Description
Intra-Party (IP)	Only shared within a party, among layers of the PPDM stack.
Fully Shared (FS)	Represents a variable with shared read and/or write access between at least two parties.
Secret Shared (SS)	Represents a variable where independent shares are given to two or more parties, which combined yield the final result.

3.1.3. Main Execution Layer

The Main Execution Layer (MEL) is itself a collection of services. These are the high level services that compromise the full data mining algorithms themselves (e.g. Na¨ıve Bayes, k-NN, SVM, etc.) which are then easily integrated into higher-level applications [20].

3.1.4. High-Performance Computing Services

For interfacing with the training databases, and for resource intensive computing conducted during the PPDM process, the High-Performance Computing web services (HPCWS) provide a generic interface to this functionality. The HPC layer can be adapted by each party to interface with their specific HPC installation, which can include clusters, grids and specialized hardware [20].

3.1.5. PPDM Services

The PPDM Services (PPDM-WS) is responsible for both providing primitive DDM/PPDM operations (e.g. secure sum), but also for providing the send, receive and peer finding operations on which those operations are built. Developing this set of services for PPDM is efficient, because most popular SMC-based PPDM algorithms tend to utilize a small set of SMC operations. By providing a toolkit of frequently used operations, as suggested in [5], developers can easily implement numerous PPDM algorithms [20].

Table 2. Examples of supported operations at the PPDM level.

Operation	Reference
Secure Sum [10] Secure Scalar Product [18] Yao Circuits [4] Oblivious Transfer [17]	[8],[11] [12], [13], [14],[15], [11] [12], [16],[14] [7], [6]

4. COMPARISON OF MULTI AGENT BASED DATA MINING AND HIGH-PERFORMANCE PRIVACY PRESERVING DATA MINING

Table 3. Comparison between multi agent based data mining and high-performance privacy preserving data mining

Features	**Multi Agent Based Data Mining**	**High Performance Privacy Preserving Data Mining**
Agent Based Support	High	Low
Privacy Preserving	Low	High
High Performance Computing	Low	High
GUI Environment Support	High	Low
Development Cost	High	Low
Flexibility	Low	High
Level Of Abstraction	Low	High

In Multiagent Based data Mining several agent working at the same time, sharing several agents at the same time also share common resources and communicate them. Multiagent data mining provide the high GUI environment support but in contrast with High-Performance Privacy Preserving Data Mining that bring high performance parallel computing and also preserve the privacy of data. High-Performance Privacy Preserving Data Mining implementation cost is low and is also flexible in nature.

5. CONCLUSIONS

Motivated by the increasing of complexity and numbers of data mining techniques, we studied and compared in this paper the decision support and data mining process to identify the main difficulties. As the field of Privacy Preserving Data Mining continues to evolve, yielding new algorithms and support for high-performance computing, Architecture for Private and High Performance Integrated Data Mining systems will continue to evolve as well.

6. ACKNOWLEDGEMENT

I am grateful to my Head of Department Dr. Rizwan Beg for providing constant support and encouragement for completing this paper, also placed immense faith in my abilities and gave me enough freedom and space to explore ideas. Without his guidance and support this work would not be completed. I thank him for everything he has done for me. I also thank my friends and colleagues for their guidance and comments to make the work better. Finally, I would like to thank my family and friends for their love and support.

7. REFERENCES

[1]. Fayyad U.M., Piatetsky-Shapiro G., Smyth P. (1996), « From Data Mining to KDD: an overview », AAAI/MIT Press, 1996.

[2]. Han J. et Kamber M. (2002), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Canada, 2002.

[3]. Ferber J. (1995), Les Systèmes multi-agents vers une intelligence collective, interEditions, France.

[4]. A. Yao, “How to generate and exchange secrets,” in Proc. 27th Annual Symposium on Foundations of Computer Science, 1986, pp. 162–167.

[5]. C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, “Tools for privacy preserving distributed data mining,” SIGKDD Explorations, vol. 4, no. 2, pp. 28–34, 2003.

[6]. B. Pinkas, “Cryptographic techniques for privacy-preserving data mining,” SIGKDD Explor. Newsl., vol. 4, no. 2, pp. 12–19, 2002.

[7]. M. Kantarcoglu and J. Vaidya, “Privacy preserving naive bayes classifier for horizontally partitioned data,” in IEEE ICDM Workshop on Privacy Preserving Data Mining, Melbourne, FL, November 2003, pp. 3–9.

[8]. H. Yu, J. Vaidya, and X. Jiang, “Privacy-preserving svm classification on vertically partitioned data,” in Pan-Asia Conference on Knowledge Discover and Data Mining (PAKDD), Singapore, 2006, pp. 647–656.

[9]. “Welcome to hadoop!” http://lucene.apache.org/hadoop/, 2007.

[10]. B. Schneier, Applied Cryptography, 2nd ed. John Wiley & Sons, 1995.

[11]. J. Secretan, M. Georgiopoulos, and J. Castro, “A privacy preserving probabilistic neural network for horizontally partitioned databases,”Aug. 2007.

[12]. G. Jagannathan, K. Pillaipakkamnatt, and R. Wright, “A new privacypreserving distributed k-clustering algorithm,” in Proceedings of the 2006 SIAM International Conference on Data Mining (SDM), 2006.

[13]. Z. Yang and R. N. Wright, “Improved privacy-preserving Bayesian network parameter learning on vertically partitioned data,” in ICDEW ’05: Proceedings of the 21st International Conference on Data Engineering Workshops. Washington, DC, USA: IEEE Computer Society, 2005, p. 1196.

[14]. G. Jagannathan and R. N. Wright, “Privacy-preserving distributed k-means clustering over arbitrarily partitioned data,” in KDD ’05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. New York, NY, USA: ACM Press, 2005, pp. 593–599.

[15]. H. Yu, X. Jiang, and J. Vaidya, “Privacy-preserving svm using nonlinear kernels on horizontally partitioned data,” in Selected Areas in Cryptography, Dijon, France, 2006.

[16]. Y. Lindell and B. Pinkas, “Privacy preserving data mining,” Journal of Cryptology, vol. 15, no. 3, 2002.

[17]. M. Naor and B. Pinkas, “Oblivious transfer and polynomial evaluation,” in STOC ’99: Proceedings of the thirty-first annual ACM symposium on Theory of computing. New York, NY, USA: ACM Press, 1999, pp. 245–254.

[18]. B. Goethals, S. Laur, H. Lipmaa, and T. Mielik¨ainen, “On private scalar product computation for privacy-preserving data mining.” In Information Security and Cryptology - ICISC 2004, 7th International Conference, Seoul, Korea, December 2-3, 2004, Revised Selected Papers, ser. Lecture Notes in Computer Science, C. Park and S. Chee, Eds., vol. 3506. Springer, 2004, pp. 104–120.

[19]. H. Baazaoui Zghal, S. Faiz, and H. Ben Ghezala, “ A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data” In World Academy of Science, Engineering and Technology 5 2005

[20]. Jimmy Secretan, Anna Koufakou, Michael Georgiopoulos, “APHID: A Practical Architecture for High-Performance, Privacy-Preserving Data Mining”