ISSN:0975-9646

A Review on Multi-Agent Data Mining Systems

Siddhant Patil[#1], Sayali Karnik [#2], Vinaya Sawant[#3]

[#1, #2]Final Year B.E. Students, [#3]Assistant Professor of Information Technology Department, Mumbai University
SVKM's D. J. Sanghvi College of Engineering, Mumbai, India

Abstract— **Data mining and intelligent agents have emerged as two fields with immense potential for research. Every intelligent agent is self-sufficient, acting independently within its boundary while collaborating with other agents to perform the assigned task efficiently. The ability of agents to learn from their experience complements the data mining process. Agent mining helps to overcome the challenges faced by data mining in a distributed heterogeneous environment. Recently, a lot of research has been conducted on the role of agents in the data mining. The paper focuses on the existing multi-agent data mining system architectures and the roles of agents in them.**

Keywords— **Agent, agent mining, multi-agent, pikater, meta learning, distributed data mining, MADM, DMMAS, decision support system, MAS, performance optimization.**

I. Introduction

We are living in the data age. With huge amounts of data being produced around the world everyday, comes the need to deal with it efficiently in order to extract useful information from it. Since the data comes from a diverse range of sources, including social networking sites, supply chains and government databases; it is usually unstructured following no particular format or layout. Hence, we need to process the incoming data to find useful information in it. Data mining is the process used wherein intelligent methods to extract interesting data patterns and knowledge from large amounts of data [1]. However, the rate at which data is produced is very high, and we need efficient methods of mining to keep abreast with it. To ensure higher performance, we use the concept of agents to support the data mining process known as agent mining. The distributed nature of agent mining brings several advantages to data mining such as autonomy, scalability, reliability, security, interactivity and high speed [2]. Agents can be used to automate the various tasks like data selection, data cleansing, and data pre-processing, to perform classification, clustering and knowledge representation. As an emerging area, a lot of research can be performed in this field. The main areas of research include agent-based data warehouse, agents for information retrieval, agents for distributed and parallel learning, information gathering agents and mobile agents for distributed data mining [3].

In this paper, section II describes the concept of an agent, while section III explains agent mining. Along with the above, section IV of this paper describes the concepts of multi-agent systems and section V describes the existing multi-agent data mining systems, each benefitting the data mining process in its own way. Finally, the section VI provides the conclusion.

II. Agent

An agent is as a software program that is used to perform a specific task on behalf of another entity i.e. an individual or another program.

Definition by Stuart J. Russel and Peter Norvig,

An intelligent agent is defined as anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators [4].

Definition by N. R. Jennings and M. Wooldridge,

It is a computer system that is capable of flexible autonomous action in order to meet its design objectives. By flexible, we mean that the system must be:

1) *Proactive:* Agents should not simply act in response to their environment; they should be able to exhibit opportunistic, goal-oriented behaviour and take the initiative when required rather than respond to it after it has occurred.
2) *Reactive:* Agents should perceive their environment and respond in a timely fashion to the changes that occur in it i.e. act in a response to a stimulus.
3) *Social:* Agents should be able to interact, when they deem appropriate, with other agents and humans in order to complete their own problem solving and to help others with their activities [5].

Apart from the properties states above, an agent can also be:

1) *Mobile:* An agent is mobile if it has the ability to move from one node to another node.
2) *Adaptable:* An agent is adaptable if it can adjust intelligibly to changes in environment.
3) *Rationality:* An agent performs its function to achieve its goal and must not act in a way that will prevent it from achieving its goal.

III. Agent Mining

Data mining is a multidisciplinary process of discovering interesting patterns in large data sets in order to assist the decision making process. Agent mining refers to the application of autonomous intelligent agents in the field of data mining in order to support and enhance the knowledge discovery and decision making process while providing high performance and scalability. Due to their autonomous, flexible, mobile, adaptable and rational nature, agents are an excellent choice for parallel, multisource, distributed mining. Fig. 1 shows agent mining as a two way process. It consists of agent driven data mining as well as data mining for agents. In agent driven data mining, for instance, agents can be used for data selection, data integration, data pre-processing, classification, clustering, association rules mining as well as knowledge presentation. In data mining for agents, data mining is used to extract knowledge from large datasets in the form of decision trees or data induces rules, which provide logic for the intelligent agents. For instance, consider an enterprise resource planning system that maintains a log of all decisions and actions taken by a company. Using data mining, the developer can identify, code and encapsulate the logic behind these decisions and actions into agents that are robust and trustworthy enough to replace the human decision making process.

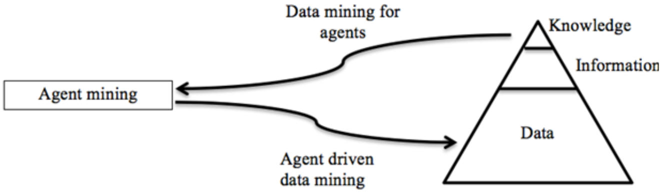

Fig. 1 Agent mining, a two way process

The agent-data mining collaboration may occur and can be analysed in a number of diverse dimensions:

- Resource dimension at data, information, and knowledge levels.
- Infrastructure dimension at infrastructure, architecture, and process levels.
- Learning dimension at learning methods, learning capabilities, and performance levels.
- Interaction dimension for coordination, cooperation, negotiation, and communication.
- Social dimension in social and organizational factors-for instance, in human roles.
- Performance dimension in the performance enhancement of one end of the coupled system.
- Interface dimension at the human-system interface, user modeling and interface design level.
- Application dimension in applications and domain problems [6].

The integration of data mining and agents provides us with benefits concerning performance and simplicity, which paves the path for the use of more intelligent and complicated agent systems in order to attain more advanced goals. Data mining requires highly trained professionals to perform the multistep process from accessing and preparing data to presenting valuable knowledge to decision makers or executives. Agent mining provides for the automation of the mining steps enabling non-experts to use the system while assisting the work of experts too. Apart from providing high performance, the data mining process supported by agents helps to increase the quality of knowledge obtained, simplify the process of identifying patterns from huge data volumes as well as help in take good decisions in real time [7].

IV. Multi-Agent Systems

The limitations of agent based technology is that there is a loose coupling between the agents achieved by introducing standardized high level agent communication languages (ACLs, e.g. FIPA-ACL) and interaction protocols. (The low level languages are used in traditional distributed computing) these high level languages try to capture the shared meaning of the messages sent. This results in interoperability among the heterogeneous systems. Thus large-scale multi-agent systems can be built for use in the real [8]. Multi-agents can solve complex issues effectively; such issues would have been too large for a single agent to solve. Agents can provide information as and when it is required and can handle the knowledge independently. They can be used in both distributed and local data mining. The multi agents thus have a high applicability since the data from different sources is different.

In a multi-agent system, there is a set of agents that work in their own sphere of influence, since the agents control different parts of the environment. If their spheres overlap, it may cause dependencies between the agents. The two principles on which the multi agents work are a high collaboration between the agents and a high degree of parallelism. The multi agents systems can be thus used when we have a complex problem to deal with, which can be broken down into sub parts, when a parallel approach will help save time, when a certain degree of redundancy is required and when the data comes from various sources and the data, controls, resources are all distributes across various system nodes [2].

V. Multi-Agent Data Mining Systems

A. Meta Learning In Multi-Agent Systems – Pikater

Pikater is a multi-agent system that makes use of meta-learning to perform a data mining task. The Pikater system enables a user to gain knowledge from a never-before-seen dataset by suggesting the best possible data mining method for mining this new dataset, obtained by meta learning over the previously obtained task results. The system stores data as well as metadata, which sometimes needs to be explicitly specified by the user with the task to be performed. This metadata consists includes the number of attributes in the datasets, number of records in the datasets, data type of the attributes and the missing values. Based on the data and metadata, the closest dataset to the dataset provided by the user is determined and the method used to mine that dataset is selected. The Pikater multi-agent system is being developed using JADE framework.

The system consists of four layers - user interface layer, computational layer, data layer and administrative layer (see Fig. 2). The task to be performed is defined by the user in a human understandable language to the UI layer. This layer is managed by the UI agents, which translates the task defined in a human understandable language into ontologies, and communicates it to the system. The computational layer consists of the data mining methods and is managed by the computing agent. The reader agents, used to read data from files and data manager service, used to access data stored in the database, make up the data layer. The agent and agent option managers control the administrative layer, which controls the entire problem solving process. The agent manager acts as a link between the interface layer and computational layer, chooses the best method from the existing mining methods as well as collects the results and provides statistical information.

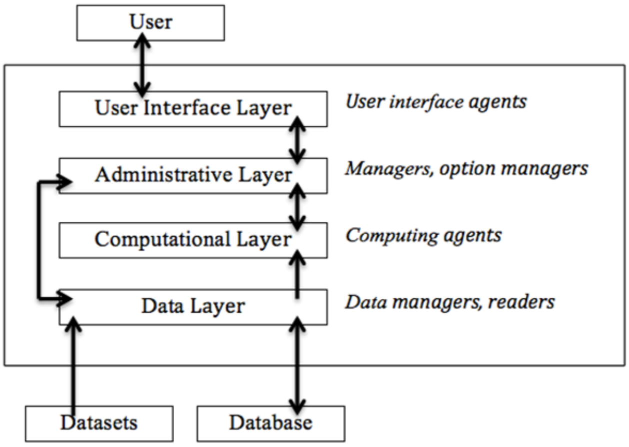

Fig. 2 Four layer abstract architecture and the agents that constitute that layer

Apart from providing a convenient, user-friendly and reusable system for data mining, the system suggests a data mining method for datasets that have not been encountered before. The use of multi-agent technology provides extensibility by allowing addition of new components to the existing system by use of structured ontology language and the international standards of agent's communication [9].

B. Using Multi-Agents Systems In Distributed Data Mining (MADM)

The Multi Agent Based Distributed Data Mining is the integration of multi-agent systems and distributed data mining wherein the concept of cooperative agents is used in data mining to overcome the challenges faced in a distributed environment like limited bandwidth, sensitivity of confidential data, limited distributed computing resources and complexity concerning multiple large systems generating huge amounts of data. The use of multi agent systems in distributed data mining allows dynamic selection of sources and data collection as well as provides scalability, security, reliability, interactivity, autonomy of the data mining process. In the MADM approach (see Fig. 3), KQML or FIPA-ALC is used as standard agent communication languages to assist interaction between the agents.

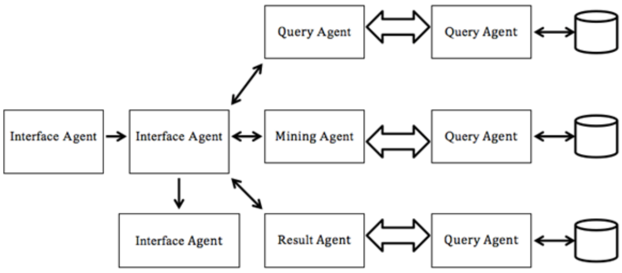

Fig. 3 The MADM approach

The commonly used agents in the MADM approach (see Fig. 3) include:

1) *Interface agent:* The interface agent is responsible for communication with the user, which includes accepting the task to be performed as input and providing the results as output. It is responsible for inter-agent communication.
2) *Agent manager:* On receiving a request from the interface agent, the agent manager forms a plan to complete the request. The agent manager is responsible for the completion of the user request, which it attains by assigning the work to different agents. The results are communicated to the interface agent. It is responsible for synchronization of the agents.
3) *Data agent: The* main function of the data agent is to supply data from multiple sources to the mining agent. The data agent maintains the metadata information of all the data sources.
4) *Mining agent:* The mining agent implements the mining algorithm. The mining agent initiates the mining technique based on the information provided by the knowledge module, such as the appropriate type of method for the problem at hand, the requirements of the method, form of input data, etc.
5) *Result agent:* The result agent receives the data mining result from the mining agent. The result agent is responsible for the presentation and visual representation of the knowledge with the help of the visualization primitives and report templates it maintains.
6) *Broker agent:* The broker agent contains the names, ontology and capabilities of all the agents registered with it. On receiving a request, the broker agent provides the corresponding names of the agents in order to fulfil the request.
7) *Query agent:* A query agent is created for each user request. The query agent uses the knowledge module schemas to generate queries in order to complete a user request.

Other agents used include an ontology agent for maintaining the ontologies, mobile agent for processing the data at each node and sending back the results to main host, pre-processing agent for performing data pre-processing tasks such as data leaning and post data mining agent to evaluate the time, accuracy, speed of the data mining agents [2].

C. Multi Agents based Data Mining for Intelligent Decision Support Systems (DMMAS)

DMMAS is a data mining multi-agent system that follows a real time agent mining approach to mine large datasets in a distributed environment. DMMAS uses JADE for the agent platform, WEKA as the data mining engine and MySQL as a database technology to store the data. The code for DMMAS is written in Net Bean IDE. In this approach (as in Fig. 4), a typical dataset is divided into N rows. The number of rows is the further divided into I segments. The number of segments can be anything between 1 to N (I = N/q where q = 1 to N). Each segment is assigned an agent, which operates on that segment only to generate its own rules.

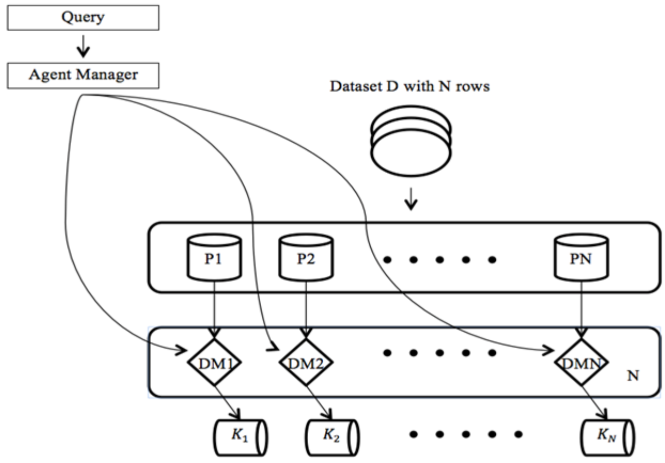

Fig. 4 The DMMAS Approach

This approach was tested using the US Census Adult Dataset, a well-known benchmark data from the University of California, Irvine machine learning repository. The dataset was partitioned into two parts as a training set and testing set. In Batch Mining Approach, the training set contained 66.6% of the dataset while testing set contained the remaining 33.3%. In the DMMAS approach, the training set was further partitioned and each partition was assigned to an agent for learning and producing a decision tree classifier. The learning algorithm used was Quinlan's C4.5 extension 8. The obtained rules depict the agent's knowledge. Similarly, the testing set is partitioned and each partition is assigned to an agent wherein it tests the classifier using the same classifier algorithm that was used to obtain the rules.

Observations were made based on the time, performance and accuracy showed that as the number of agents increased the testing and training time decreased. The accuracy improved up to a certain point after which there was no significant increment by addition of more agents. The processing speed was higher as compared to batch mining, however there was a marginal loss of accuracy. It was concluded that for large data sets, the advantage of improved efficiency overshadows any negative effect as a result of loss of accuracy, with the accuracy improving, as more and more data is available [10].

D. Performance Optimization of Data Mining Applications Using a Multi-layered Multi-agent Data Mining Architecture

In this system, a feedback from the user is used to optimise the performance of data mining. The general architecture of this system describes a five-layered approach (see Fig. 5) to achieve this goal. The five layers are:

- Problem solving ontology layer
- Optimization layer
- Intelligent technology layer
- Reactive agent layer
- Distribution and coordination layer

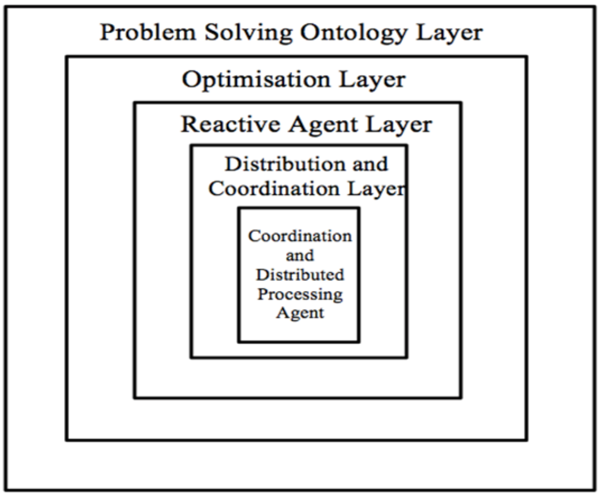
Problem Solving Ontology Layer
Optimisation Layer
Reactive Agent Layer
Distribution and Coordination Layer
Coordination and Distributed Processing Agent

Fig. 5 Multi-layer data mining architecture

The first layer i.e. the problem solving ontology layer conceptualises any particular problem to be solved during any of the phases like pre-processing, decomposition, control decision, post-processing. The optimization layer uses the feedback from the user to improve its performance. The intelligent technology layer uses agents for knowledge extraction from given data. The reactive layer provides response to the stimuli of the agent environment. The agents in this layer need not have learning and its use is mainly for real time response requiring environment. The distribution and coordination layer speeds up the data mining process using parallel or distributed data processing. It may also use the coordinate activities of agents in the data mining system.

As shown in Fig. 6, the performance optimising agent optimizes the parameters of the belief base. It is then fed to the prediction agent. In the initial stages, the user provides the feedback and the negative or positive feedbacks are communicated to the performance agent. As the learning builds up based on actual use feedback, the performance agent will be able to act like human agent and can take over the tasks for performance measurement. The performance measurement agent should perform certain tasks to achieve the said goal. It needs to identify and quantify user feedback. The feedback can be in linguistic non-linguistic form. The feedback should also be modelled using a historical functional relation between system predictions and user acceptance or rejection. A constraint for this could be the quality of the feedback since human feedbacks are subject to experience, environment etc. Another constraint is the search space. If the population of the algorithms is more then there is a rapid convergence for the process of selection and mutation. Also more population would require more computing resources. The post condition of the performance agent is that the process converges to a global minimum overtime. The performance agent can be triggered by the decision phase agent, user invoked coordination agent or by the coordination agent independently if optimization is the only task to be performed for the data mining system. Neural networks are used to predict the actions based on the previous data. It encompasses both the machine prediction and user feedback to learn the human behaviour. The prediction process would soon overtime be comparable to the human user.

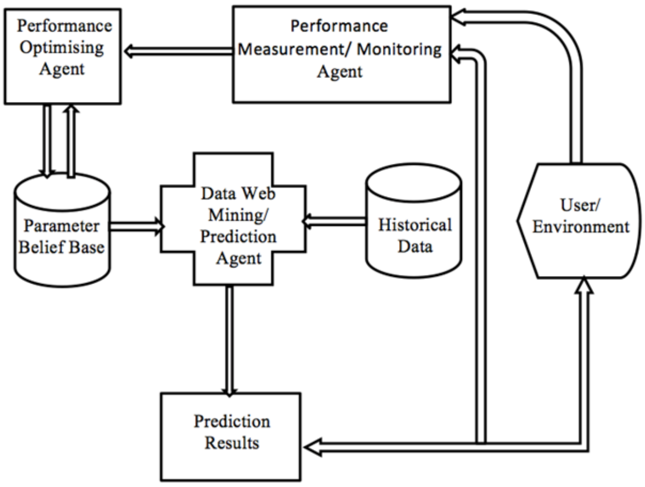
Performance Optimising Agent
Performance Measurement/ Monitoring Agent
Parameter Belief Base
Data Web Mining/ Prediction Agent
Historical Data
User/ Environment
Prediction Results

Fig. 6 Optimization of data mining system in prediction

The user is a part of the system here since the optimization is based on his feedback. The soft computing technologies like fuzzy logic, neural networks etc. provide approximate solutions to the problems. To optimise this, users need to be a part of the system, which is achieved by this research. This system can be used for credit card approval systems [11].

E. A Multi-Agent Method for Parallel Mining Based on Rough Sets

Rough sets are used to reason imprecise or incomplete data and to find relationships between them. Rough sets do not need preliminary or additional information like statistics, probability, etc. Multi agent systems are very useful for distributed systems and concurrent engineering. The basic properties of the MAS are:

1) *Decomposition:* Dividing the problems into smaller tasks so that they can be addressed in isolation.
2) *Abstraction:* Simplifying a model so that it will emphasize some properties while hiding some others.
3) *Organization:* Defining and managing problem-solving components. We can thus define the agents and the relationships between them using algebra in Rough rule mining system.

The information and the multi agent systems can be divided into many sub systems, so a subsystem can be represented using a sub agent. The sub-agent encapsulates the decision table, mining algorithm and interface of communication with others. The multi agent parallel mining based on rough rules for various mining tasks uses agent properties. Each sub-information system is an intelligent agent and every agent uses thorough set mining method. We can use the different agent interrelationships and get the final decision rules.

The data sources for data mining uses the enterprise information system S=(U, A, V, F) The strategies used mainly are that the information systems can be divided into smaller agents, that different data sources at different locations are considered to be different agents and that data sources which are isomerous can be divided into different agents. These strategies help the MAS to decrease the computing complexity of the algorithm and strengthen its abilities for dealing with the isomerous data sources. It improves the parallel data mining degree; the system performance time is decreased.

A dispatching agent, performing agent and synthesizing agent are the main components of MAS architecture.

The agent model is Ag=(IS, C, R), dispatching agent sends tasks to the performing agent based on the performing condition C. The dispatching agent gives the performance tasks and works on the formula:

<Ag_task>::=<condition_ID><Perform_Ag_ID>{<mining_ap
pr_space>}

Where, condition_ID- unique mining condition (matches with performing agent), Perform_Ag_ID-unique flag of performing agent, mining_appr_space-divided decision space and mining algorithm (Rough set theory).

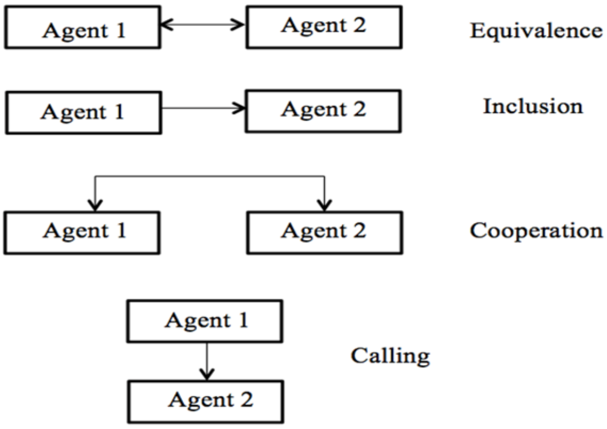

Fig. 7 Basic interrelationships among agents

Equivalence, inclusion, cooperation and calling relationships are the performing agent inter-relationships. Synthesizing agent synthesizes the results of performing agents (see Fig. 7). Using the similar formulae, final decision rules can be extracted. If the rules can't satisfy the decision maker requirements, he can justify the parameters and performing condition, then convey it back to the dispatching agent until the decision rules are satisfied [12].

VI. CONCLUSIONS

This paper thus, gives a brief idea about the concept of agents and the multi-agent mining systems. Data mining in a distributed heterogeneous environment becomes flexible, adaptable, robust and easier with the use of data mining agents. Using the multi-agent systems for the data mining process we can tackle a large amount of information and also increase the speed of dealing with that information.

The multi-agent systems can also be optimized for a better performance using some of the techniques described in this paper, thus further increasing their efficiency. Thus in conclusion we state that the concept of agent mining and the multi-agent systems have gained a huge momentum in the recent years and have a capability of delivering far more. More research in this respect can develop the data mining systems to a greater extent making them more efficient and increasing the accuracy of the mined data.

Acknowledgment

The authors of this paper wish to acknowledge Dr. Abhijit R. Joshi, Professor, Vice-Principal (Academic) & Head of Department of Information Technology along with SVKM's D. J. Sanghvi College of Engineering and Mumbai University for their kind support and the necessary facilities for the preparation of the paper.

References

[1] J. Han, M. Kamber and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed., Waltham, USA: Morgan Kaufmann, 2012.

[2] A. Fariz, J. Abouchabaka, N. Rafalia, "Using Multi-Agents Systems in Distributed Data Mining: A Survey," *Journal of Theoretical and Applied Information Technology,* vol. 73 No. 3, pp. 427-440, March 2015.

[3] L. Cao, G. Weiss, and P. S. Yu, "A brief introduction to agent mining," *Springer*, May 2012.

[4] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed., Upper Saddle River, New Jersey, USA: Pearson Education Inc., 2010.

[5] N. R. Jennings and M. Wooldridge, "Applications of Intelligent Agents," *Springer-Verlag*: New York, USA, 1998.

[6] L. Cao, V. Gorodetsky, P. A. Mitkas, "Agent Mining: The Synergy of Agents and Data Mining," *IEEE Intelligent Systems,* pp. 64-72, May/June 2009.

[7] A. M. Al-Barky and J. Ali, "Intelligent mining agent," in *8th International Conference on Computing Technology and Information Management (ICCM),* 2012, vol. 1, p. 23.

[8] E. Serrano, M. Rovatsos and J. A. Botía, "Data mining agent conversations: A qualitative approach to multiagent systems analysis," *Journal of Theoretical and Applied Information Technology,* vol. 73 No. 3, pp. 132-146, March 2015.

[9] O. Kazik, K. Pešková, M. Pilát and R. Neruda, "Meta learning in multi-agent systems for data mining," pp. 433-434, Elsevier, Jan 2013.

[10] D. Sharma and F. Shadabi, "Multi-Agents Based Data Mining for Intelligent Decision Support Systems," in *2nd International Conference on Systems and Informatics (ICSAI),* 2014, pp. 241-245.

[11] Q. Li and R. Khosla, "Performance Optimization of Data Mining Applications Using a Multi-layered Multi-agent Data Mining Architecture," in *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA),* 2005, pp. 227-231.

[12] Z. Geng and Q. Zhu, "A Multi-Agent Method for Parallel Mining Based on Rough Sets," in *Proceedings of the 6th World Congress on Intelligent Control and Automation),* June 2006. pp. 5977-5980.