Mobile-Agent Based Distributed Fuzzy Associative Classification Rules Generation for OLAM

B. RaghuRam
Department of Computer Science
Pondicherry University
Pondicherry, India
raghu9b.naik@gmail.com

G.Aghila
Department of Computer Science
Pondicherry University
Pondicherry, India
aghilaa@yahoo.com

***Abstract*— The capabilities of distribute data base to store huge amount of data and providing scalability, integrity leads to many of real time databases are stored in distributed nature. In order to apply data mining in real time applications it is important to provide efficient distributed data mining techniques. As a result of the use of Online analytical mining (OLAM) technology in new fields of knowledge and the merging of data from different sources, it has become necessary for OLAM models to support distributed data mining technology. Data from different sources are not always consistent with the format and some sources may not reliable. In this paper we proposed a frame work for mobile-agent-based distributed analytical mining which can perform analytical mining on distributed and heterogeneous database systems and can manage reliable and unreliable sources differently. More over this architecture capable of performing mining on global and local data bases separately. Based on this architecture a flexible and efficient mobile-agent-based fuzzy association classification rules generation algorithm which can mine and present the global and local associative classification rules at the same time.**

Keywords-Mobile Agents; Online Analytical Mining; Classification

I. INTRODUCTION

The efficient integration of the concept of data warehousing, online analytical processing (OLAP) and data mining systems converges to OLAM results in an efficient decision support system. OLAM is first proposed by J.Han [1]. The conventional mining algorithm faces complexity for multidimensional data mining and also provides results with interpretation difficulties. Consideration of user defined meta rules is also a difficult task in traditional mining approaches. Considering the effective and easiness of OLAP [2] technology in data analysis, the proposal for On-Line Analytical Mining came which combines the features of OLAP operations with data mining technique in order to perform effectively on multidimensional databases and data cubes. The advantage of OLAM is that using OLAP operators it provides interactive data mining by allowing user to select required dimensions, abstraction levels of data and also make it possible to view results using different visualization techniques.

In real time many databases are distributed in nature. Many of companies are spread across different regions, so the data transactions are likely stored at different sites. As such it leading companies to adopt distributed database structure to store databases. At the same time the transactions in the distributed database may be changed time to time. Because of that, developing distributed and incremental mining algorithm on distributed information sources is very important and challenging. So in order to apply OLAM on real time databases it is important to explore and develop applications of OLAM on distributed databases.

In case of performing OLAM operation on distributed data bases systems, collecting data form different distributed local sources for data warehouse is major challenge because the data in distributed data bases keep on changing. The data in different sources may contain most of not related fields so selecting subject oriented data is another problem. The data in distributed local sources are in different formats so changing that format as data warehouse requirements is one more problem.

It quite obvious that out of the different data sources some of the data sources may not be company owned so that kind of data should be considered as unreliable data so showing the results distinctly with out considering unreliable source information and with considering unreliable source information is preferable one. Out of information in distributed data sources identifying unreliable data is another problem to be faced in case of using distributed data sources for OLAM.

By considering need of up dating information at frequently and considering the selection and changes that as to be carried out on data and mainly considering dynamic nature of system we conceive that fallowing mobile agent oriented architecture will be use full for performing OLAM on distributed databases. In order to realize that we proposed mobile agent oriented distributed analytical mining architecture. Mobile agents [3] are intelligent programs that can migrate on computer networks. The concept of having mobile agents carrying out tasks is creating a new paradigm for network-enabled distributed computing.

The classification is a significant technique in data mining with applications in industrial and scientific domains. The

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

efficiency of a classification model is evaluated by two parameters, namely the accuracy and the interpretability of the model. Many studies show that Associative Classifiers give better accuracy and interpretability than other traditional classification models [4].Considering the importance of associative classification we proposed an algorithm for performing associative classification of data cube and we shown how this algorithm can be adopted for our architecture.

In order to provide model which manages information distinctly form reliable sources and form unreliable sources our proposed model uses a fuzzy multi dimensional model proposed by Carlos Molina et al [5]. The model provides aggregations intuitively to end user by means of fuzzy logic. All OLAP operators of model can effectively manage data imprecision resulting from merging of data from heterogeneous sources both in fact and dimension. The model provides low fuzzy confidence value for data from unreliable source comparing to data from reliable sources. The information with confidence more or equal to user specified threshold will be considered by aggregation operators. Motivating from the efficiency of the model in performing fuzzy taxonomies and managing imprecision the fuzzy multidimensional data cube [5] has been adopted as base in our model.

The contribution of this paper is as follows. 1) The architecture of the mobile-agent-based distributed analytical mining model proposed for applying OLAM in the distributed, heterogeneous database systems. The architecture consists of a distributed analytical mining system (DAMS) and sub-DAMS. DAMS is located in the management system of the distributed database system. The sub-DAMS is located in each site of the distributed database. 2) Based on the architecture of the mobile-agent-based distributed analytical mining model an associative classification rule generation algorithm on fuzzy data cube has been proposed.

The remaining of page organized as fallows. Related work is presented in section2, the proposed architecture of mobile agent based distributed analytical mining model presented in section3, mobile agent based distributed associative classification algorithm presented in section4, evaluation presented in section 5and this paper ends with conclusion and feature work

II. Related Work

Before The model for integrating business intelligence and intelligent agent for distribute data sources are proposed by SamoBobek [6]. The proposed model mainly concentrates to overcome problem of integrating data form different sources but the model won't consider the reliable and unreliable source differently and another disadvantageous is that the model can perform mining operations only on global databases only it can not perform mining on local databases. The other model that can perform mining operations on both local and global databases proposed by Yun-Lan et.al [7] but disadvantage of model is it can not consider reliable and unreliable sources differently.

In associative classification rule mining the generation of association rules are very much important and crucial step. Because associative classification is performing associative classification on distributed data base system it is interesting to study the applying distributed association rule generation algorithm and generating associative classification rules. Algorithm Count Distribution [8], which is the adaptation of Apriori algorithm [9], has been proposed for the parallel mining environment. The PDM [10] algorithm tries to parallelism the DHP algorithm. FPM [11] adopts the count distribution approach and has incorporated two powerful candidate pruning techniques, distributed pruning and global pruning. The problems with such approaches are (1) only the global large itemsets can be mined. The information of the local large itemsets on each site cannot be provided. So it may not provided support for local business decision. (2) The information in the previous mining cannot be used to reduce the cost of current mining, since they are not incremental algorithm.

In order to overcome such problem with distributed mining algorithms and other agent based models we proposed architecture that perform analytical mining operations on global and local databases separately and that can perform mining with reliable and unreliable data bases separately.

III. Architecture Of The Distributed Anaytical Mining Based on Mobile Agent

After the architecture of the distributed analytical mining system, which is based on mobile agent, is illustrated in fig. 1. The system includes the distributed analytical mining system (DAMS) and the distributed analytical mining sub-system (sub-DAMS), the structures of DAMS and sub-DAMS are as follows:

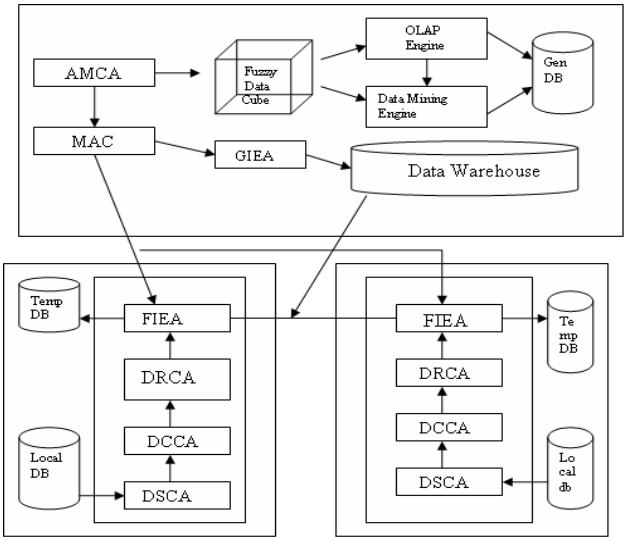

Fig 1: Architecture of Distributed analytical mining based on Mobile agent

A. The Structure of DAMS

Analytical Mining Controller Agent: This is the core agent of the distributed knowledge discovery system. The function of this agent is as follows: 1) Dispatches the mobile agent to each sub-dam and obtain the frequent item set information form local data base and storing the local frequent item sets information in to data warehouse to get the global knowledge. 2) Analyzes the data in the data warehouse by the means of OLAM.

Data Warehouse: The history data is stored in the data warehouse. The databases in the different sites of the distributed information system are often heterogeneous. The data warehouse techniques should be adopted to transfer the data to uniform format so the data can be organized and accessed efficiently. In this data warehouse data form reliable source and data from unreliable source will be distinctly store.

OLAM: OLAM is the combine of OLAP and data mining. Data warehouse provides the platform for the multi-dimension data cube which can be analyzed and mined. In OLAM the mining operation can effectively make use OLAP operators for mine data in across dimensions and multi levels. The operating of slice, scan, drill and rotate can be processed.

Fuzzy Data Cube: Fuzzy datacube extracts the information from data warehouse as directed by AMCAgent. The fuzzy data cube used in this process provides different degrees of reliability to data from reliable sources and data from unreliable sources. Depends up on the cut off value for membership reliability the data cube decides whether the value should be considered for further operations or not.

Global Knowledge Base: The knowledge in the global knowledge base includes the knowledge mined using OALM engine from the data warehouse and the knowledge which is mined from all the sub-DAMS by the mobile agents and integrated by the analytical mining controller agent.

Mobile Agent Controller: MAC is the framework within which the mobile agent activities in the distributed data mining system take place. MAC is responsible for generating, activating and assembling the agents required for the data mining process. The different agent types and their tasks are briefly discussed below.

1) Global Information Extraction Agent (GIEA): This agent can be dispatched to data warehouse to extract the global data which is require by fuzzy data cube.

2) Data Subjectivity Check Agent (DSCA): This agent can be dispatched to sub-DAMS to select subject oriented data with respect to data warehouse out of the various data fields available in local databases.

3) Data Consistency Check Agent (DCCA): This agent can be dispatched to sub-DAMS to checks whether data retrieved from local data bases are in compatible format with data warehouse format or not and pre-process the data in the local database to the format of data warehouse.

4) Data Reliability Check Agent (DRCA): This agent can be dispatched to sub-DAMS to checks whether local database is reliable source (company owned) or from unreliable source (external data source) and intimate the information to data warehouse, so that unreliable data in data warehouse can be stored distinctly.

5) Frequent Item Extraction Agent (FIEA): This agent can be dispatched to sub-KDS to mine the local large itemsets and to dispatch entire information to Data warehouse.

B. The Structure of sub- DAMS

Local database: The local real time information is stored in the local database, which can be queried by the local supervisor and at the same time can be mined by the mobile agent dispatched by the supervisor.

Local knowledge database: the knowledge discovered by the mobile agent is return to the KDMS, at the same time it is saved in the local knowledge base, which can provide references to the local supervisor.

Mobile agent execution environment (MAEE): The KDMS dispatches the mobile agent to the sub-KDS, so the mobile agent execution environment has to be installed in
the sub-KDS. In this paper Aglet is used as the mobile agent platform.

IV. MOBILE-AGENT-BASED FUZZY ASSOCIATIVE CLASSIFICATION RULES GENERATION ALGORITHM

The proposed model mobile agent based fuzzy associative classification rule extraction model can work on individual local data bases and in combined data bases.

A. *Associative Classification Rule Extraction from Local Databases*

In case of performing analytical mining individual local data bases at first DSCA sends to locate the subjective data from local database. Then the DCCA agent will be send to check whether the data is in compatible format of data cube or not if not the data pre processing will be done and stored in data cube. Then FIEA agent send to the local data base to collects the frequent item sets. Using the information stored in fuzzy data cube the analytical mining can be performed across dimensions and multi levels fuzzy associative classification rules can be using algorithm1

B. Associative classification Rule Extraction from Multiple Local Databases

In case of performing analytical mining form multiple local data base DSCA sends to locate the subjective data from different databases that should be considered for data extraction. Then the DCCA agent will be send to check whether the data is in compatible format of data warehouse or not if not the data pre processing will be done and stored in data cube. The DRCA agent checks whether data is from reliable sources (company owned) or from unreliable sources (External sources). If the data is from unreliable sources it intimate the data warehouse so that it can provide membership values in less compare to data from reliable sources. Then FIEA agent send to the local data bases to collects the frequent item sets. Out of the data stored in data warehouse GIEAgent extract the required data that is needed by data cube. Finally

using the information stored in fuzzy data cube the analytical mining can be performed across dimensions and multi levels fuzzy associative classification rules can be using algorithm1.

C. Associative Classification Rule Generation Algorithm

Once the data stored to data cube by agent than OLAM model will start associative classification rule generation process. In this phase the proposed algorithm extracts associative classification rules at user selected levels of dimensions by making multiple passes to data cubes. The model shown in figure.2.The algorithm performs following steps.

1) Obtain frequent candidate rule items with and without specific class labels.
2) Generate rules using the item sets found in the previous step.
3) Prune the rules obtained in the previous step.

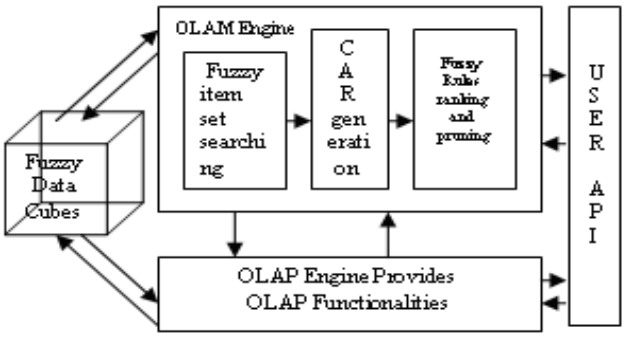

Fig 2. Architecture of GBAC model for classifying Fuzzy Data Cubes

The rules, which may not obtain in base level, may be possible at generalized higher abstraction level. The rules at higher abstraction will be less in number and more interpretive to user. In order to over come that our model provides user to select require levels for dimensions. Our models adopts support calculation model proposed by Nicolas Marin et al [12] to overcome the disadvantageous due to using same support for all levels of abstractions. At first step the approach calculates support for item sets at base level. If the item would not prove frequent at this step it generalizes items to its higher abstraction and calculates support again. In order to overcome using same confidence level for all abstraction the model also provides interesting mechanism for calculating threshold support at higher level of abstractions.

The Algorithm1 shown in figure 3 describes the proposed model for mobile agent based distributed fuzzy generalization based associative classification rule generation. The algorithm will function in two phases namely data collection phase and rule generation phase. At firth phase of algorithm (Line 1 to 13 in algorithm) the data collection activities will performed by agents as described in section 4.1 and 4.2 depending up on whether mining has to perform on local data base or global database. Once the agent's stores data in data cube the rule generation phase (Line 14 to 38 in algorithm) will be start.

Algorithm1 : Mobile- Agent-Based fuzzy Associative classification Rules generation

1. MAC create MEE agents according to the requirements
2. MAC dispatches the mobile agents to required local database
3. IF (MAC dispatch agent to single local database)
4. The DSCA agent gather the data in local disk with respect to requirements of data cube
5. The DCCA agent perform the consistency check of local database
6. The FIEA agent extracts the itemsets and stores in datawarehouse
7. End if
8. IF (MAC dispatch agent to multiple local database)
9. The DSCA agents gather the data in local disk with respect to requirements of data cube
10. The DCCA agents perform the consistency check of local database
11. The DRCA agents performs the data reliability checks
12. The FIEA agents extracts the itemsets and stores in data warehouse
13. End
14. C_1 ? φ
15. k ? 0
16. c= elements at the base le dimensions
17. for (k=2 ; C_{k-1} ? φ ;k++)
18. C_k=candidate Gen(Cfk-1)
19. while C_k ? 0
20. I = first element of C_k
21. C_k = C_k-I
22. for each transaction T $\in$ C do
23. if (rule Sub Set(I,T)) then
24. I. candidate Supcount++
25. if (T.class = I.class)then
26. I rule Supcount++
27. End For
28. if (I rule Supcount ? $threshold_{sup}$ + (1- $threshold_{sup}$) * A(I)) then
29. C_k ? C_k U {I}
30. else
31. C_k ? C_k U Generalize (I)
32. End while
33. CAR_k = genRules (C_k)
34. $PrCAR_k$ = pruneRules (CAR_k)
35. End For
36. CAR_c = U CAR_k
37. $PrCAR_c$ = U $PrCAR_k$
38. AMC Agent presents the $PrCAR_c$ to End user.

Fig 3. Algorithm 1

A rule item is frequent if its rule support is equal to or greater than threshold value. For the base level, the algorithm utilizes the same value given by user. But elements at higher levels may group several values at the base level. The problem of considering same confidence value in all hierarchical levels will reduce efficiency of data mining task by including huge set of rule sets. This problem is known as rule over fitting. So in order to obtain an intuitive and accurate classifying model, support threshold for higher abstraction levels should be higher than support at lower abstraction level. In order to avoid this problem, in this work the automated threshold generation of higher abstraction model proposed by Nicolas Marin [12] is used. The model calculates threshold using equation (1). For an aggregated item set the support threshold is defined as:

$$Threshold_I = threshold_{sup} + (1 - threshold_{sup}) * A(I). \quad (1)$$

Where $threshold_{sup}$ is the support threshold established by the user for the basic levels and A(I) is abstraction value of item I which ranges from 0 to 1. Abstraction value for an item is provided by data cube depending up on abstraction level of item. In a data cube elements at higher abstraction will present a higher abstraction value (A (I) value) than elements at lower abstraction. The abstraction of the rule item set is average of all individual item set abstraction values.

All the oneitemsets are processed in the above manner and the algorithm collects the rule items whose support count is more than threshold values. Using this frequent rule items the algorithm produce the associative classification rules using

the ‘genRule’ function which performs pruning operations using confidence value. The confidence value for that rule calculated using the formula shown equation (2). In genRule function at first pruning will be performed on ruleitems that have the same condition set but referring different class labels. In such case the highest confidence set is chosen as the possible ruleitem.

Confidence=(rulesupCount/condsupCount)*100% (2)

For rule pruning any of standard models can be used in our application we used standard pessimistic error rate method specified in C4.5 model [13]. The ‘pruneRule’ function using pessimistic error rate method. The pessimistic error rate method is an efficient pruning model which provides more concise rule items for further processing steps. This pruning method is not mandatory and can be replaced with any other standard pruning method.

In the next iteration, the algorithm generates item sets using a rule item which has passed the threshold of minimum support count. The above specified steps will be repeated on all item set generated. The process will continue until all the possible item sets have been processed. Finally all the obtained pruned associative classification rules are made into one set.

The final set of rules depending up on input provided to cube (either local data or general data) can be used as classification rules. The AMC Agent will collect classification rule and present concern user.

V. PERFORMANCE EVALUATION

To assess the performance of the algorithm IDMA, we performing experiments on a cluster of computers of P-4 2.0 GHz and 526 MB of memory. The simulation program was coded in java. Our method used UCI heart databases as local databases. The notation of the database is in the form Dx, where x is the number of transactions. And the minimum support threshold is noted as minsup**.** We are assuming 2% of transaction as minimum support. The result conducted for association rule generation algorithm alone (line 14 to 38) for local data base is encouraging. The experiments with agent integration to be carried out. The time for the transmission of mobile agent on the network and the time for the mobile agent would affect the performance of the online analytical mining. The effect of transmission to mobile agent has to be calculated. After complete system is going to constructed we are planning to compare its performance with parallel data mining algorithms.

REFERENCES

[1] J.Han. Towards on-line analytical mining in large databases , ACM Special Interest Group on Management of Data, SIGMOD vol. 27, no. 1, pp. 97–107, 1998.

[2] C.E.F.Codd, S.B.Codd and C.T.Salley Providing OLAP to User-Analysts, E.F.Codd Associate,1993.

[3] Joseph Kiniry, Daniel Zimmerman, A Hands-on Look at Java Mobile Agents, IEEE Internet Computer, Vol 1(4): 21-30, 1997

[4] B.Liu, W..Hsu & Y.Ma. Integrating classification and association rule mining. Knowledge discovery and data mining ,pp. 80–86,(1998)

[5] Molina, D. Sanchez, M. A. Vila, and L. Rodríguez-Ariza, “A new fuzzy multidimensional model,” *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 6, pp. 897–912, (2006).

[6] Samo Bobek1 and Igor Perko, Intelligent agent based Business intelligent, Formatex ,2006

[7] Yun-Lan Wang, Zeng-Zhi li, Hai-Ping Zhu, “Mobile agent based distribute and incremental technique for data mining”, Cybermetrix,Xin2-5, IEEE Proceeding, November-2003

[8] R. Agrawal and J.C. Shafer, Parallel Mining of Association Rules: Design, Implementation and Experience. IEEE Transactions on Knowledge and Data Eng., 8(6): 962-969, December 1996.

[9] R.Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, in Proc. 20th Int, Conf, VLDB, pp. 478-499, 1994.

[10] T. Shintani and M. Kitsuregawa, Hash Based Parallel Algorithms for Mining Association Rules, Proc.4th Int’l Conf. Parallel and Distributed information Systems, IEEE Computer Soc. Press, Los, Californ, 2006.

[11] David W. Cheung and Yongqiao Xiao. Effect of Data Skewness in Parallel Mining of Association Rules, Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, Vol.1394, New York: SpringerVerlag, 1998

[12] Nicolas Marin, Carlos Molina, José M. Serrano, and M. Amparo Vila. A Complexity Guided Algorithm for Association Rule Extraction on Fuzzy DataCubes , IEEE Transactions on Fuzzy Systems, VOL. 16, NO. 3, pp 693-714, june (2008).

[13] J.Quinlan, C4.5: programs for machine learning, Morgan Kufman, (1993).