Fuzzy Associative Classifier for Distributed Mining

B RaghuRam
Kakatiya Institute of Technology and Scince
Warangal
Andhrapradesh

JayadevGyani
Jayamukhi Institute of Technological Sciences
Narsampat, Warangal
Andhrapradesh

B Hanmanthu
Kakatiya Institute of Technology and Science
Warangal
Andhrapradesh

ABSTRACT
Distributed data mining extracts the knowledge from distributed data sources without considering their physical location. The need for such systems arises from the fact that, in real time many data bases are distributed geographically in different locations.Often transferring data produced at local sites to centralized site for extracting knowledge results in excessive time and transmission cost and may also raise privacy issues. These reasons emphasis on need of distributed mining algorithm. In order to overcome lack of efficient associative classification techniques in field of distributed data mining this paper proposes an associative classification model on distributed databases. By considering the efficiency of fuzzy association rules in providing accuracy, intuitiveness and in overcoming the problem of crisp partitioning this model adopts fuzzy associative rules for classification. The proposed model accuracy tested on UCI data bases given encouraging results.

General Terms
Data Mining, Distributed Computing.

Keywords
Associative classification, Fuzzy association rules, Distributed data bases.

1. INTRODUCTION
In real time many modern applications fall into the category of distributed systems. For cost and time effective learning these systems need to support distributed data mining (DDM)[1]. The applications of DDM can be found with different scopes in situational aware computing, intrusion detection, weather forecasting and web mining, etc.DDM also shown huge impact on business intelligence applications this is because many of companies are spread across different regions and their data transactions are stored at different sites. As such it is 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. Imposing tradition centralized data warehouse based mining models in which total data brought to a centralized site is time consuming and results in huge communication cost. These reasonsmake developing distributed and incremental mining algorithm on distributed information sources is very important and challenging.

The classification is a significant technique in data mining with applications in industrial and scientific domains. Recently, distributed data mining, in particular, distributed classification, has attended active research. Efficiency of a classification model is evaluated by two parameters, namely the accuracy and the interpretability of the model. Even afterassociative classification methodsproved to be accurate and interpretable than other traditional classification models[2] there is lack of associative classification in distributed environment.Considering the importance of associative classification we proposed a model for performing classification based on association rules in distributed data base environment.

The efficiency and intuitiveness of associative classifier mainly depends on efficiency of association rule extraction methods. Considering this factor crisp and fuzzy association rule extraction methods are studied. Crisp association rule mining algorithms can mine only binary attributes. Assuming we have three such ranges for the attribute Income, namely up to 20K, 20K-100K, 100K and above. Income 35K and even Income 98K would fit in the second partition. Using ranges will cause uncertainty at the boundaries of ranges, leading to loss of information and small changes in the intervals may lead to very different results, which can be misleading decision maker.

To answer this problem it is better to adopt fuzzy ranges instead of crisp attribute values. The fuzzy ranges represented in the interval (0, 1), instead of having just 0 and 1. A transaction having multiple attributesfall in to a fuzzy range between 0,1. Thus many fuzzy methods based association rule extraction methods proposed [3], in which quantitative values for numerical attributes are converted in to fuzzy binary values. The pre processing method adopted by our model presented in section 3. Many of fuzzy associative classifiers are proposed in literature but there are no proposals which can perform fuzzy association rule based classification on distributed data bases.

Motivating from above factors we developed a fuzzy associative classification model which can perform classification over distributed data bases. The contribution of the paper as follows. 1. We proposed an associative classifier model which can perform associative classification on distributed data bases.2.Depend up on the proposed model a classification based fuzzy associative algorithm proposed.3. Finally we proposed a global ranking model to build a compact associative classifier.

2. RELATED WORK
The Classification Based on Associations (CBA), proposed by Liu [2] is a first association rule-based classifiers reported in the literature, generates all the classification association rules (CARs) that satisfy user-defined support and confidence thresholds. Classification based on Multiple class-Association Rules (CMAR) introduced by Li [4], applied FP-growth method. The importance of the minimum rule sets for classification isalso outlined by Yang [5]. They present their minimum set rule algorithm and investigate the relationship between error rates and minimum support/ confidence. In Hu &Li[6] suggested an optimal association classifier (OAC) that is less sensitive to the missing values in an unseen test data. The problem how to set and tune the basic mining parameters outlined by Zaiane&Antonie [7] and a new rule pruning approach is presented applying the so-called ARC-AC classifier developed by Zaiane et al [8]. An information gain based association rule classification (GARC) method was also proposed by Chen [9]. The above specified all associative classifications suffer from the following problems like huge set of rules which are hard to interpret and crisp logic based rules presented by these rules may not give intuitive look to end user.

In associative classification rule mining the generation of association rules are very much important and crucial step. Because associative classification is performing associativeclassification on distributed data base system it is interesting to study the models applying distributed association rule generation algorithm.Algorithm Count Distribution [10]**,** which is the adaptation of Apriori algorithm [11], has been proposed for the parallel mining environment. The PDM [12] algorithm tries to parallelism the fp-groath algorithm. FPM [13] adopts the count distribution approach and has incorporated two powerful candidate pruning techniques, distributed pruning and global pruning. By studying above approaches we adopted FDM model for generating frequent item sets on distributed data bases considering its effectiveness in generating local and global frequent item sets with less number of message interchanging which results in effective time and cost optimization.

To our knowledge, there are very few published studies that have focused on distributed associative classification. Thakur [14] proposed a parallel model for CBA system [2]. Another approach is presented by Djamil[15] to generate associative rules for classification in fp-groathapproach. Irrespective differences in association rule generation approach both models will implemented on distributed environment based up CDapproach[9]. In which training data set is partitioned among *P* processors, and in every iteration all sites calculates local counts of candidateset and broadcasts these to all other processors. After a synchronization step, association rules are generated from frequent itemsets. This strategy inherits two major limitations of CD approach that is tight synchronization and the duplication of the entire set of candidates at each site due to which these models gives huge communication cost. An agent based approach for mining fuzzy association rule based classifier presented by raghuram[16] but model is based up on centralized approach in which agentsgathers data from all distributed sources to centralized source and associative classification algorithm performed on centralized source. The major problem of this approach is huge amount of data as to transfer from all sources to centralized source.

In order to overcome such problem with associative classification algorithms we proposed a fuzzy associative classification model that can perform accurate and intuitive associative classification on distributed environment with optimized time and communication cost.

3. FUZZ PRE PROCESSING MODEL

3.1 Expert Driven Fuzzy Partitioning

To provide accurate and intuitive classifier our model adopts fuzzy association rule based classification. In order to extract fuzzy association rules from data sets the quantitative attributes of the data sets need to pre processed by dividing into fuzzy partitions. To do that our model adopts expert driven fuzzy partitioning method in which human expert will decide partitions by considering user intuitiveness. To decide attributes membership values in fuzzy partitions our model used Trapezoidal fuzzy membership function. The process explained in following age attribute example. The fuzzy method uses Young, Middle-aged and Old, partitions for age attribute and then ascertain the fuzzy membership μ (range [0, 1]) of each crisp numerical value in these fuzzy partitions as shown in Fig.1. Thus, Age = 35 may have μ = 0.6 for the fuzzy partition Middle-aged, μ = 0.3 for Young, μ = 0.1 for Old. And Age = 59 may have μ = 0.3 for Middle-aged, μ = 0.1 for Young, μ = 0.3 for Old. By using fuzzy partitions, we preserve the information encapsulated in the numerical attribute. Thus, many fuzzy sets can be defined on the domain of each quantitative attribute by experts. Once partitions are decided the original dataset transformed into an extended one with attribute values having fuzzy memberships in the interval [0, 1]. In case of binary attributes which does not have quantitative values, the μ values by default taken as 1.

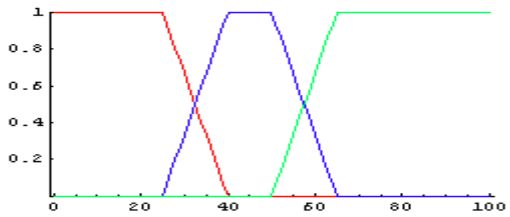

Fig 1: Fuzzy membership function for age

3.2 Support Count in Fuzzy Model

As our model adopts fuzzy associative model it is need decide how support rules with fuzzy quantities attributes are counted. Our model adopts Fuzzy Aprioriwhich is a modified version of the original Apriori algorithm [11] which can take fuzzy membership values of itemsets into consideration. FuzzyApriori counts the support of each itemset in a manner similar to the counting in Apriori, the only difference is thatit calculates sum of the membership function μ corresponding to each record where the itemset exists. The non quantitative item μ value considered as 1 on their presence or 0 for their non presence. Thus,the support for any itemset is its sum of membership functions over the whole fuzzy datasets divided by number of occurrences. The confidence of a rule will be calculated as in Apriori algorithm by measuring inter dependency factor among the items present in a corresponding itemsets.

4. FUZZY ASSOCIATIVE CLASSIFIER FOR DISTRIBUTED ENVIROMENT

4.1 Associative Classification Model in Distributed Environment

Data mining approaches on geo graphically distributed data sources can be applied on two ways know as centralized model and distributed model. In the centralized model the required data distributed over various sources gathered in to a centralized site where mining algorithm will be applied. Itprovides accurate results but impose huge communication cost and time. Where as in distributed environment, mining will be performed at local sites and results of local sites will be optimized based on feedback from other sites. Even though distributed model will gives less accuracy than centralized model it reduce communication cost, time complexity and makes algorithm easily scalable. In order to develop model with optimum time and communication cost we adopted distributed environment.

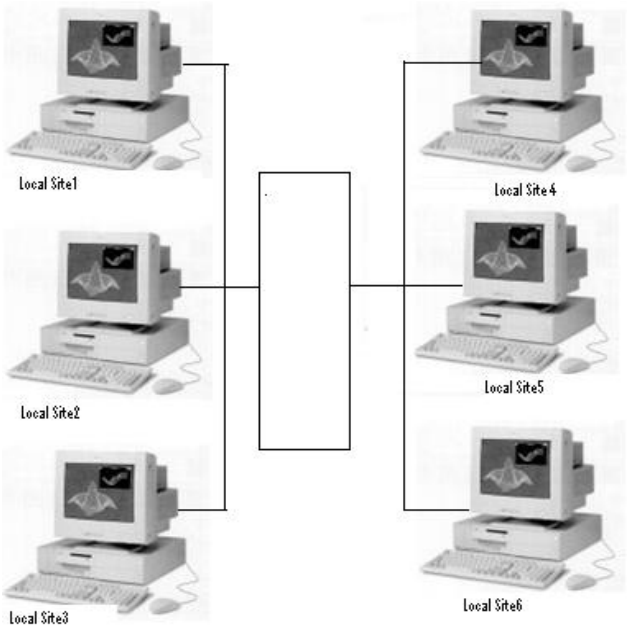

Fig 2: Distributed environment with shared nothing systems

Our model aimed to perform fuzzy associative classification on distributed data bases where data stored in various shred nothing machine connected over distributive environment as shown in Fig2. Our model assumes that data is horizontally distributed over various sources. In distributive environment we may face two criteria's out of one is the training data distributed among various sources and classification as to perform on one source. In other approach along with training data distribution the classification also need to perform on various sources. We proposed our model to handle both of these factors. The proposed distributed fuzzy association rule based classifier (DFAC) model will follow three steps in order to perform associative classification on distributed data bases.

The first step is data preprocessing step, in this step training data with known class labels present at distributed sources are converted in to fuzzy data sets as per expert view (briefed in section 3). Once data sets mapped into fuzzy data sets it is ready for training and testing associative classifier. The data at all sites segregated into training data and testing data.

In second step the model generates fuzzy association rules for classification using FDM [13] based algorithm (briefed in section 4.2) on pre processedtraining data sets at all local sites with minimum number of message transfers to provide cost effective computation.

In third steplocally large association rules broad casted to all sites or send to a particular site depend up on requirement. Where rules will be divide as per class label and ranked as per proposed ranking model (briefed in section 4.4).Finally the ranked rules of different class labels will be sorted to make use as classifier.

The classifier rules obtained in previous step will be evaluated by applying them on testing data to which class labels are known. If the result of the evaluation is up to mark the resulted associative classification rules can be used for classifying real time data at single site or at multiple sites as per requirement. The obtained classifier will be applied to classify actual data.

4.2 Classification basedFuzzy Associative Rules Extraction

In this phase the proposed algorithm for extracting globally large associative rules for classification on distributed data bases is presented. The algorithm performs following steps:

Step1 Obtains frequent candidate itemsets with specific class labels at all local sites.

Step2 Obtain globally frequent item sets by exchanging locally large item sets count among all the sites. The globally frequent items only send to next step.

Step3 generate and prune rules using the globally large item sets and forwards globally large item to step1.

The steps repeats up to all levels of itemsets are generated. Proposed algorithm shown in section 4.2. Initially it fallows Apriori algorithm [11] like method 'candidateGen' to obtain frequent one item sets at the local sites. Each item set is in form of <candset, L > where candset is a set of items and L is class label. While generating item sets the algorithm will consider same items with different class labels as different items sets in order to obtain frequent item sets specialized for class labels. The candidate item set support count is calculated as method briefed in section 3.2 will be given by 'fuzzySupcount' function in algorithm. The threshold support [0 to 1] is the support threshold established by the user. Out of item sets generated the item sets which crosses support threshold are considered as locally large item sets.

After generating locally large item sets the global fuzzy support of item sets will be collected by sending messages to all other sites as in FDM[13]. By receiving these request messages local sites checks local fuzzy support count of respective candidate item sets in their data bases and inform back to corresponding sites. After gathering information from all sites now a local site get global fuzzy support count of individual item sets. The global support count of item will be sum of local supports of an item set divided by number of local sites. The item set is considered as globally large if and only if it succeeds to cross support threshold.

After generating globally large item sets using this frequent rule items the algorithm produce the associative classification rules using the 'genRule' function which works just like in Appriory [11]. In this step rules with less fuzzy confidence than user provided threshold will be pruned. These rules will be pruned again by 'pruneRule' function using pessimistic error rate method specified in C4.5 model [17]. The pessimistic error rate method is efficient pruning model which provides more concise rule items for further processing steps. This pruning method is optional one and can be replaced with any other standard pruning method.

In next iteration the algorithm generates item sets using a rule items which are proved to beglobally large. 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 associativeclassification rules are made into one set.

4.3 Classificationbased Fuzzy Association Rule Generation Algorithm

Input:

- Threshold sup: Support threshold to consider an candidate item as frequent

- Threshold conf: Confidence threshold for rules pruning.

Output:

- Set of fuzzy associative classification rules.

Algorithm:

1. C_f←pre processed local data base transactions.

2. k←0

3. C_{fk}= C_f

3. for (k=1 ; C_{fk}-1≠ ø ; k++)

a) C_k=k-candidateGen(C_{fk})

b)C_{fk}= ø

c) whileC_k ≠0

i) I = first element of C_k

ii) C_k = C_k-I

iii) for each transaction T Є C do

A) if ((ruleSubSet(I,T) and (T.class = I.class)) then

B) calculate I. fuzzySupcount

C) End

iv) if (I. fuzzySupcount>fuzzySupportThreshod)

A) get(I.globalSupport(I.fuzzySupcount))

B) if (I.globalSupport>fuzzySupportThreshold)

C) CAR_k← {CAR_k} U {genRules (I)}

D) addCfk=I

4. End

5.$PrCAR_k$= {$PrCAR_k$} U {pruneRules (CAR_k)}

6.End

4.4 Distributed Associative Classifier

After generating rules at individual sites using globally large item sets at local sites the rules can send to a particular system in distributed environment where classification has to perform or all local sites can broadcast rules to other local sites if classification as to perform at all sites.

At corresponding site in order to build a classifier using generated associative rules we are using CBA [2] based model. As per CBA model rules will be divided according to class labels and sorted in a order. In order to improve accuracy of CBA model with overlapping rules we adopted one more rule (rule 3) to three rules for rule ranking method proposed in CBA. Given two rules r_i and r_j, r_i>r_j (i.e., $r_i$precedes r_j) if one of the following holds good:

- Rule1 The confidence of r_i is greater than that of r_j.
- Rule2Their confidences are the same but support of r_i is greaterthan that of r_j
- Rule3 All the above constraints are the same but if a particular rule item r_i is present more number of sites than r_j them r_i will be given more preference.
- Rule4 All constraints are the same but r_i is generated before r_j

The set of sorted rules are considered as associative classifier. The classifier present at an individual system or at a particular system can be used per classification.

5. PERFORMANCE ANALYSIS

5.1 Experimental Setup

In order to testify the performance of our DFAC model, our experiments utilized three P4 2.40GHz PCs with 512Mb main memory and windows XP OS. The three PCs are located in 100Mb LAN. We use the pima-indians-diabetes_data(PIMA) and Heart disease data sets sets obtained from UCI Machine Learning Repository.

At first step of experiment the data sets are processed by mapping quantitative items into fuzzy sets and divided data sets horizontally among e different system connected in distributed environment. Then after data at local sites divided into training and testing data sets in 50:50 ratio using 10X10 validation method. In second set proposed fuzzy associative classification algorithm implemented on training fuzzy data setsdistributed over different systems and global fuzzy associative classification rules are generated.In next step the obtained rules at local sites broad casted to all other sites where classifier is obtained by dividing the rules according to class labelsand sorted using the proposed rule ranking model. The classifier generated send to all site because our experiment intended to classify data present at all local sites. Finally classifier applied on test data sets without class labels.The classifier efficiency calculated by comparing generated class labels against actual class labels. The fuzzy associative classification rules generated for PIMA data set shown in Fig 3.

5.2 Evaluation

In order to evaluate our distributed fuzzy association rules based classifier we compared obtained accuracy against standard associative classifier models. We adopted fuzzy support threshold as 0.2 and confidence threshold as 0.8 according to standards. From this experiment we observed that our proposed ranking model efficiently deal with rules consisting similar support and confidence values. It generated less number easily interpretable and generalized rules. This proposed distributed fuzzy association rules based classification (DFAC) model shown accuracy of 77.2% on test data bases of UCI PIMA dataset and 81.7 on test data base of UCI Heart disease data sets. The Table 1 shows accuracy of basic associative classifiers CBA [2], CMAR [4], and OAC [6] models on UCI heart database (collected form literature) and accuracy of our proposed model. On comparing accuracy of our model to other associative classification model accuracy on same data sets proves that our proposed our model provides good efficiency even though it is applied on distributed data sets as shown in Table 1.

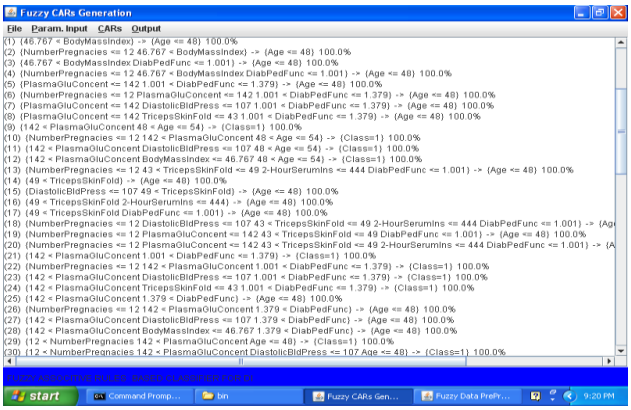

Fig 3: DFAC model results on PIMA data sets

Table 1.Classification accuracy of algorithms

Model Name / Accuracy	CBA	CMAR	OAC	DFAC
PIMA	72.4	75.1	78.1	77.2
Heart disease	81.5	82.2	81.1	81.7

6. CONCLUSION

Many of classification algorithms like decision trees are proposed on distributed data bases, but there is lack of associative classification on distributed data sources. In this paper, we proposed a distributed model which allows the associative classifier in a shared-nothing architecture. In order to provide intuitive and accurate classifier our model makes use of fuzzy association rules. The experiment conducted on UCI data sets proved that our DFAC model provide good accuracy against other conventional associative classifiers.

7. REFERENCES

[1]. M Zaki. 2000. Parallel and Distributed Data Mining: An Introduction. In Large-Scale Parallel Data Mining, pages 1–23.

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

[3]. Chen G., Yan P., Kerre E.E.2004. Computationally Efficient Mining for Fuzzy Implication-Based Association Rules in Quantitative Databases. International Journal of General Systems, 163-182.

[4]. W.Li, J. Han & J.Pei.2001. "CMAR: Accurate and efficient classification based on multiple class-association rules ",ICDM, pp. 369–376.

[5]. J.Yang.2003.Classification by association rules: The importance of minimal rule sets. The twentieth international conference on machine learning

[6]. H.Hu,& J.Li.2005.Using association rules to make rule-based classifiers robust". Proceedings of the sixteenth Australasian database conference pp. 47–54.

[7]. O.Zaiane, R. & M.-L. Antonie.2005. On pruning and tuning rules for associative classifiers. Ninth international conference on knowledgebased intelligence information and engineering systems, KES'05, pp. 966–973.

[8]. O.Zaiane,M.L.Antonie, A.Coman.2002.Mammography based classification by an association rule-based classifier". ACM, SIGKDD, pp. 62–69.

[9]. G.Chen,H.Liu,Zhang,2006. A new approach to classification based on association rule mining ", Decision Support Systems, Vol. 42(2), pp 674–689.

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

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

[12]. 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.

[13]. David W. Cheung,YongqiaoXiao.1998 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: Springer Verlag.

[14]. G.Thakur, C.J. Ramesh.2008A Framework For Fast Classification Algorithms, International Journal Information Theories & Applications V.15, pp. 363-369.

[15]. D.Mokeddem, H Belbachir.2010. Distributed classification using class association rules mining algorithm. IEEE International conference on Machine and web intelligence, Algeria.

[16]. B.RaghuRam and G.Aghila.2009. Mobile agent based distributed fuzzy associative classification rules generation for OLAM. IEEE conference on IAMA, Chennai, India.

[17]. J.Quilnlan.1993.C4.5: programs for machine learning, Morgan Kufman.