IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email:editoriijcs@ipasj.org
Volume 6, Issue 4, April 2018 ISSN 2321-5992

IMPROVED ENSEMBLE SUBSPACE CLASSIFIER WITH HIGHER PRECEDENCE ACTIVE SAMPLING ON CLASS ASSOCIATION RULES

Kayal Padmanandam[1] , Kannan Subramaniam[2]

[1]Research Scholar, Research &Development Centre, Bharathiar University, Coimbatore, India.

[2]Associate Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai, India.

ABSTRACT

Ensemble classifier an assemblage of classifiers, is a prospective practice for improvising the capacity of individual classifiers. Ensemble classifier has produced a powerful accuracy related to single classifiers. The aim of this work is to explain the role of sampling done on class association rule to enhance the predictive power of the algorithm Ensemble Classification on Fuzzy Utility Mining with active sampling(ECFUM-AS). This classifier learning system uses random subspace algorithm that outfits training of increased attributes, despite many fuzzy classifications escalates attribute magnification. The sampling technique used here for accuracy improvisation can play a better prediction in the decision-making system. As soon as the research essences for better accuracy factor, we focused on subspace sampling. Subsequently, the infinite sampling space is created only with class association rules with higher order confidence precedence, which has agreed an immense progress in algorithm's accuracy. Results elevated the accuracy about 6 to 7 percent on increased rule set.

Keywords: Association Rules, Classification algorithm, Sampling Methods, Knowledge based System, Knowledge Discovery

1. INTRODUCTION

The Introduction describes the rudimentary of the algorithm ECFUM [1] which is the foundation of this work. Next it gives an intuition about sampling and its techniques used with the algorithm for improved accuracy in prediction.

1.1 ECFUM

Ensemble classification [2] has been proven as powerful prediction classifier. The Ensemble Classifier Fuzzy Utility Mining (ECFUM) algorithm uses the Subspace [3] type of ensemble for classification. It is a supervised learning algorithm used to predict the leadership capabilities of work force. The expanse of humanoid skills is ambiguous and need scalable measures for predicting the genuine rate of credentials. This has taken the implementation of attributes fuzzification on grounds of class member functions. Figure 1,2,3 explains well about the working of the algorithm ECFUM.

Algorithm 1: FAR Generation

Input: Dataset D,Weight w,Fuzzy support and confidence threshold as min_supp,min_conf.

Output: Fuzzy association rule set carrying attributes associations.

Step 1: Scan database D,find 1-item frequent set L_1.

Step 2: Using L_{k-1} generate candidate itemset c_k

Step3: For each itemset c in c_k calculate fuzzy support,if it is greater than min_support, add to L_k, else step 4.

Step 4: If it is lesser than min_support ,find min_suf of the itemset C. If it satisfies min_suf threshold add to L_k else prune the itemset.

Step 5: For each itemset l in L_k,calculate Fuzzy confidence,if is greater than min_conf,then add it to Fuzzy association rules FARs.

Step 6:continue step 2 untill L_k is greater than 1,else go to step 7.

Step 7: Prune un-interesting rules.

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email:editoriijcs@ipasj.org
Volume 6, Issue 4, April 2018 ISSN 2321-5992

Figure 1 . FAR Generation Algorithm

Algorithm 2: Calculate min_support and min_suf

Input: Itemset I,f_{ij}-Fuzzy membership of j^{th} value of ith attribute,w_{ij}-weight of j^{th} value of i^{th} attribute.

output: Itemset support and Fuzzy partial weighted SUF

Step 1: For each transaction t in D,Find the fuzzy value associated with item(i) in I,calculate the fuzzy value of I using min operator for each transactions.

Step 2: sum up all the fuzzy transaction values of each calculated in step 1 and divide by the size of the dataset which is respresented as supp(I).

Step3:after step 1 ans 2, if the supp(x) is less than min_supp for the respective transaction ,find the weight(W_i)of each item in itemset(I_i) and multiply withcorresponding fuzzy value as I_i X W_i (IW).

Step 4:sum up all(IW) and divide by the number of occurences of the itemset(I_i) as

$$suf = \sum Ii \times Wi \div N$$

Figure2 .Min_supp,Min_suf Calculation Algorithm

Algorithm 3: Building the ECFUM Classifier
Input: Rule itemset *Ri* and dataset *D*
Output: ECFUM Classifier .

Step 1: For a transaction *t* in dataset *D*, organize the rule-itemset *Ri* using random subspace method to classify t (refer section 4.2.5)
Step 2: Traverse the data set D and follow Step 1 for all transactions,and takes count of COV .
Step 3: Prune un-interesting rules from the rule set and calculate classification error- rate .
Step 4: Iterate Step 1 to 3 until the error rate is minimized.

Figure 3. ECFUM Classifier algorithm

1.2 Sampling

Sampling plays a vital role in solving numerous databases and knowledge mining problems [4]. Sampling allows Big data scientist, analytical modelers and miscellaneous data analysts to exert with a small, controllable amount of data to build and execute analytical models more rapidly, despite the fact of producing accurate findings. Sampling is predominantly useful with data sets that are too big to analyze. But the important part of sampling is the “magnitude of the sample”. Literature equally supports the cases of immense and trivial mining of samples from dataset. Many researches have given a surprise fact that, small samples can even provide knowledge, and often bigger samples can do it with enhanced data manipulations [5].

Sampling is expected to construct demonstrative samples from huge data, trusting a classifier will outperform with such samples than being trained with ample data set. [6]. Here, the framework uses sampling on class association rules for classifier training based on resultant samples. The knowledge mining literature provides multiple sampling algorithms [7,8,9] in three discrete types as follows.

1.2.1 Sampling Categories

Realm of sampling defines the term "population" as the maximum set of possible observation, and a "population element" or a sample is a set of observation that belongs to the "population. Infinite sampling states, a population element is rested back to the population after each selection, hence population elements are unconfined for approaching echoes. Whereas Finite sampling states, a population element is not rested back to the population after selection for approaching echoes, hence population elements are confined.

1.2.2 Sampling Methods

a) **Static -** The action of sampling is achieved, deprived of any information but a database carries. Random sampling method is the most commonly used algorithm for static sampling [6].

b) **Dynamic –** It is discerned from static lonely in sample selection. On every echo, with the selected sample, a classifier is built and assessed. If the resultant classifier does not reach the accuracy of satisfaction, the algorithm is made to iterate once more with other samples.

c) **Active Sampling** stay differentiated from other types, lonely by selecting items on echoes. Such sampling aims for diminished numeral frequent items picks, for the classification learning system towards gaining the maximum classifier perception [6]. This is realized by calculating Effective Estimation Score(EES) which explicates the maximum knowledge gain. In our algorithm [1] the EES is found by selecting class association rule with higher order precedence of confidence only.

2. ENSEMBLE CLASSIFIER FUZZY UTILITY MINING WITH ACTIVE SAMPLING (ECFUM -AS)

ECFUM-AS developed from ECFUM [1] and FPWUM [10] is an ensemble classifier which uses infinite samples with the active sampling techniques to select higher order precedence class association rules. An upright decision-making is possible only when the construction of item sets is meaningful. Hence it has been observed to construct it with intense care based on min_sup, min_conf and min_suf. Also, the features weight treatment, factor of improvisation has been applied exclusively for "hidden item set". This stratagem has surely increased the performance of the algorithm [10]. The higher order precedence sampling technique is not applied at the time of item set generation. Rather they are applied after the generation of class association rules(CARs). This is because to generate more combinations of item set for association rule generation and thereby to increase class association rule generation. The CARs voting technique, ranking and pruning are well explained in [1].

2.1 Improved Ensemble Classification with Active sampling

Initially the algorithm begins with "Fuzzy Associative Classifier-(FAC)" [11] which is a classification procedure grounded on "Fuzzy Association Rules" (FARs) mining. Progressively it uses partial weighting factor for attributes using a novel measure called "SUF (Skill Utility Factor)" leading to the development of the algorithm FPWUM [10]. Well along, the algorithm has taken the direction towards Ensemble classification, since it has proven literatures on accuracy than single classification techniques, leading to the development of ECFUM. It adopts "Random Subspace Ensemble" for the classifier learning system which is an administered algorithm holding the details of features with corresponding labels. The algorithm uses thirteen features including the class. All the feature's values, excluding class (which has 4 probable labels) are fuzzified scoring on fuzzy member roles [11]. The construction of FARs and CARs takes place ensuring the occurrence of vital pruning and refinement of feature choice complications. "The item set's association are revealed in Fuzzy Association Rules and those association's class are revealed by Class Association Rules."

After these steps the algorithm ECFUM is amended by making a required imperative difference in the sampling data. The ECFUM algorithm carries the complete set of class association rules. The CAR is build up with all constraint of threshold on support, confidence and the SUF measures. Whereas the same ECFUM with active sampling (ECFUM (AS)) carries the CAR which has only higher confidence. These rules are called as higher precedence CARs. These sampled CARs have been used in training, validation and testing of the model. This difference makes lot of sense on terms of accuracy enhancement. Also, we did a comparative study on the support factor. Increasing support on the same argument has no big differences which explain that increase in confidence does the required job. Also, the CAR which

has been produced by SUF pass [10] has higher confidence. The comparative study of the algorithms accuracy is given in Table 2.

2.2 Experimental Observations

Table 1 explains the various classifier algorithms accuracy result at various threshold level. Here ECFUM shows highest performance accuracy at various support and confidence thresholds. Table 2 presents the comparative study of the algorithm ECFUM and ECFUM-AS for increased rule set. Receiver Operating Characteristics (ROC) curve envisions the space of classifiers, in two-dimensional view of true positive rate versus false positive rate. It is used frequently in analysis and diagnostics researches [12]. The ECFUM-AS ROC is shown in Figure 5. Fact of knowledge behind the curve is, the curve and the accuracy are relational. When the curve raises and permit through superior left the accuracy is superior, and when the curve falls and so is the accuracy too [13].

(1)

Table 1. Accuracy in % of ECFUM on various support and confidence Threshold

Algorithm/Threshold	0.1,0.6	0.2,0.5	0.2,0.7
ECFUM	96.4	89	83.2
Boosting	50.4	54.4	58
Bagging	93.2	82	81
KNN	92.8	85.2	80
RusBoosted trees	50.4	45.6	56.4
Svm	81.6	87	78.8

Table 2. Improvement on accuracy % on increased rules with ECFUM(AS)

ECFUM(AS)	ECFUM	Threshold 1,2,3
89%	82%	0.2,0.4
86%	80%	0.2,0.6
86%	80%	0.4,0.8

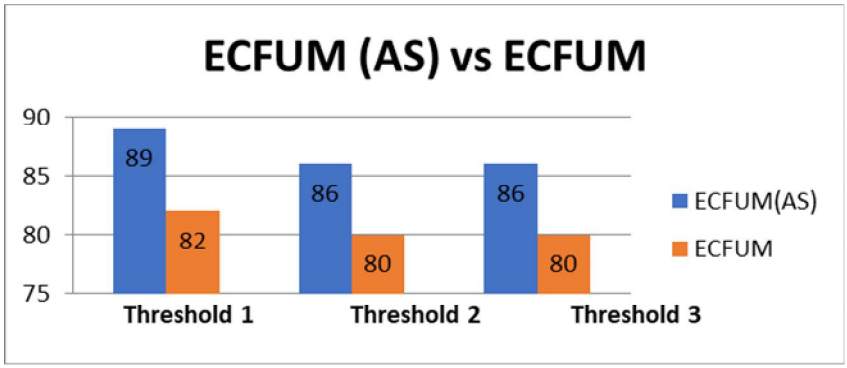

Figure 4. ECFUM(AS) vs ECFUM Accuracy chart

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email:editoriijcs@ipasj.org
Volume 6, Issue 4, April 2018 ISSN 2321-5992

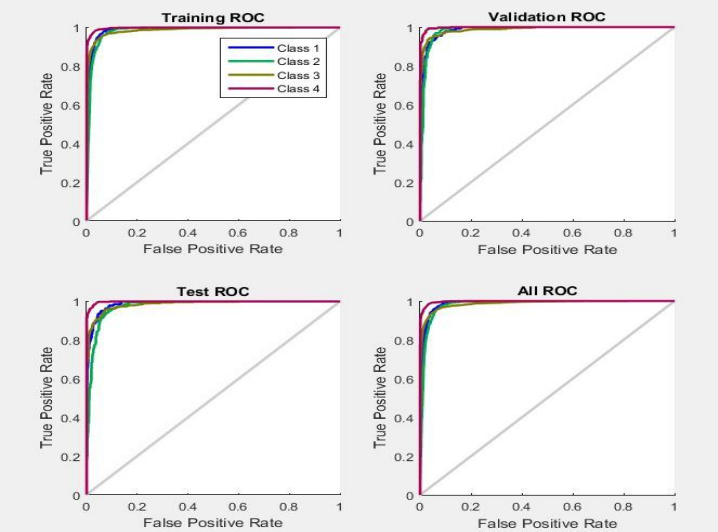

Figure 5. ROC of EPFUM(AS) at different phases

3.Conclusion

The algorithm makes use of the measure support for positive item set and applies SUF for hidden interesting item sets. Once the class association rules are generated, the main challenge of the algorithm is to increase the prediction accuracy using the sampling technique "Higher precedence Active sampling". As anticipated the results are appreciable and appropriate for the model human leadership skills prediction system. The fuzzy logic employed with ensemble subspace learners in this classification, supports the algorithm for efficient decision making.

References

1. P.Kayal and S.Kannan, "An Ensemble Classifier Adopting Random Subspace Method Based on Fuzzy Partial Mining," Indian Journal of Science &Technology ,vol 10(12),March 2017.
2. A.Rahman and S.Tasnim,"Ensemble Classifiers and Their applications:A review ,"International Journal of Computer Trends and Technology,vol.10,pp.31-35,2014.
3. Ho and T.Kam, "The Random Subspace Method for Constructing Decision Forests," IEEE Trans. Pattern Analysis and Machine Intelligence,vol.8,pp.832-844,1998.
4. T.C.Venkatesan, P.Vinayaka and Y.Sabharwal, "Analysis of sampling techniques for association rule mining," IBM India Research Lab, New Delhi.
5. R.M.Kaplan,D.A.Chambers and R.E.Glasgow, "Big data and large sample size:A cautionary note on the potential, for Bias,"Clinical and Translational Science,vol.7,pp.342-346,2014.
6. M.Aounallah, S.Quirion and W. Mineau, "Distributed Data Mining vs. Sampling Techniques: A Comparison," Canadian Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence 3060, pp. 454–460, 2004.
7. M.J.Zaki,S.Parthasarathy,W.Li and M.Ogihara "Evaluation of Sampling for Data Mining of Association Rules,"IEEE Proc. RIDE,pp.42-50,1997.
8. A.H.Milley, J.D.Seabolt and J.S.Williams, "A SAS Institute Best Practices Paper: Data Mining and the Case for Sampling," SAS Institute Inc., Cary, NC,1998.
9. D.Jensen and J.Neville, "Correlation and Sampling in Relational Data Mining,"Univeristy of Massachusetts, MA 01003,2001.
10. P. Kayal and S.Kannan, "A Partial Weighted Utility Measure for Fuzzy Association Rule Mining," Indian Journal of Science and Technology, vol.9(10)March,2016.
11. P.Kayal and S.Kannan, "Building Fuzzy Associative Classifier Using Fuzzy Values," International Journal of Science and Research,vol.3,pp.1498-1500,2014.
12. T.Fawcett, "An introduction to ROC analysis" Pattern Recognition Letters,vol.27,pp.861-874,2006.
13. H.Zweig and C.Gregory, "Receiver-operating characteristic (ROC) plots: A Fundamental evaluation tool in clinical medicine". Clinical Chemistry, vol.39, pp.561-577,1993.

Kayal Padmanandam, is a Research scholar in R&D center, Bharathiar University, Coimbatore, India. She had her Master of Technology in computer Science from Manonmaniam Sundaranar University,Tirunelveli, India in 2007 and M.Sc., from M.K University, Madurai,India in 2005. She has published various research articles in International/National journals and has gained a profound knowledge in academics and professional careers of Industry, Teaching & Research for a period of 10Years.