International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Minimal subset generation on Lazy Associative Classification based on cogency and Harmonic Mean

S.P. Siddique Ibrahim[1], Dr. S.P. Syed Ibrahim[2]

[1]Assistant Professor, Department of computing science and Engineering, Kumaraguru College of Technology, Coimbatore, India
[2]Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, India

Email: [1]siddiqueibrahim.sp.cse@kct.ac.in, [2]syedibrahim.sp@vit.ac.in

Abstract--- Lazy learning associative classification method gains higher accuracy than counterpart eager method. The majority of the existing lazy associative classification algorithms generate an exponential number of subsets that increase the computation time. Moreover, lazy learning associative classification with traditional support and confidence measure leads to missed out some rare and prime subsets. The proposed method overcome this problem by focusing on the important feature of the given test instance. As a outcome the proposed system is able to generate a minimal number of high-quality rare itemset as a subset. The proposed rare lazy associative classifier produces high quality subsets and increases classifier accuracy, according to the empirical results. Experiment results show that the proposed algorithm attains better accuracy and reduce the computation time than traditional lazy learning associative classification.

Keywords: Lazy Associative Classification, Feature selection, Classification, Subset generation, Gain ratio

1. INTRODUCTION

At present day, no one can lead a life without sharing a data, to connecting people all over the world people let sharing gigabytes data in the form of digital images, videos, audio and keep all these digital data in local and cloud storage. Social media plays major part and it is a field of collecting and sharing many data and these data are stored in a common storage device. In each computer aided field requires data for development through which knowledge can be gained. The business value from these data can be extracted from data mining techniques. Data mining is an enduring method which is highly involved in the development of the previously unknown and interesting pattern from dense database. Frequent, infrequent, closed itemset, rare finding are most significant areas in data mining. Association Rule Mining (ARM) [1] is noted in the form of A->B, which means an itemset A associate with B in database. Classification is a promising field in datamining to construct the model based on previous known data commonly considered as supervised learning. This task consists of Divide and Conquer [10], PRISM [8], PART [6], Statistical [4], Tree Induction [5], , Neural network [13, 9], and Navie Bayes [4] Decision tree algorithm is well suitable for new class prediction process than all these algorithms This algorithm has simple flow chart like structure to represent decisions and its consequences.

Generally, Association Classification is promising research approach in recent year applies in many real time applications including medical field [7, 12]. This enriching technique has the similar slant features of both classification and association rule mining then then hybrid technique called Associative Classification (AC). The AC build the classifier model based on strong search of frequent class association rules pattern in which it assigns the target class for the unknown instances. Eager AC and lazy learning AC algorithms works under associative classification proposed by Plastino and Merschmann [3, 15]. Eager algorithm has generates rulesets based classifier from training set and the later part predict the class in a final stage. Whereas, the lazy method is delaying the prediction phase until user given new test instance. This method projects only the data in the training datasets, which is already in the test samples. Hence the eager method takes more computation and time consuming process. This complication is well handled by lazy method [2, 29, 24, 8] proposed by Elena etal. Ibrahim et al introduced minimal rule generation process in lazy learning method which regains better classification result with low computation cost with higher accuracy [29, 8].

In order to mine rare rules, setting the minimum support threshold value to a very low value is necessary, but it leads to the generation of too many rules and sometimes many of these rules may not help during rule generation. The real time dataset con- sists of items that are of non-uniform in nature. Some items are frequently occurring and few items rarely appear. A rare itemset consists of items that appear in few transac- tions of the datasets and are pruned because they do not meet the minimum requirement.

Consider the case where user specified minsupp of itemset {X, Y, Z} is 3% and the minsupp of its subset {X, Y }, {X, Z}, and {Y, Z}is 5%. Based on the count in the database {X, Y, Z} to be frequent item but none of the subsets are frequent with respect to their count. For instance, {X, Y }, {X, Z}, and {Y, Z} have support of 3% and {X, Y, Z} 4%. In this example, {X, Y, Z} is considered frequent because of its count greater than the user-defined minsupp threshold 3%. However, the traditional al- gorithms simply ignore this itemset since {X, Y }, {X, Z}, and {Y, Z} fall below the user specified threshold. In fact, certain items may appear frequently while others may only appear occasionally

This motivative proposed work is purely based on attribute selection based subset generation which decreases the hardest thing in the class rule generation. For many concerns like improvement in the field of business there is a huge demand for the many such applications. This proposed lazy attributes selection method initially compute the probability value for given test instance and further ruleset generation only to the selected attributes and does not produce huge number of class rules.

The other part is organized as follows: firstly, the previous studies in the proposed GRLLAC is presented. Next segment comprises sample computation of the GRLLAC. Experimental results and comparison is also presented. Finally, the paper present conclusion and future direction.

2. RELATED WORK

Association rule mining is as same as of used with classification called Class Based Association rule mining (CBA), whose rule format other side is a class label which is very essential and it takes more number of sweep over the database during rule generation process [16]. The CBA algorithm which picks up the most suitable rules based on minimum support and clearly ignore the uninteresting rules during classifier construction. Moreover, this algorithm puts up large efforts and takes more time in finding the unknown charactertics and produces numerous rulesets as this method adopts well known Apriori algorithm. Most of the real world applications have been implemented with AC but leads to memory overload in some dense databases [2]. CMAR [17] MMAC [14] are improved rule searching technique to utilize the tree based structure to predict the class with multiple rules to obtain the higher accuracy. Hence the frequent itemsets are clearly gained from the well known support and confidence measure from the standard mechanism of ARM. Nevertheless, many significant rules simply ignored by the algorithm as due to the high limit of support measure at the same time redundant and duplicate rules may part of the mining process and degrade the performance due to the lower limit assignment of support.

Liu et al [16] proposed the approach that bring the complete set of high quality rules for prediction that leads more efforts. Meanwhile, this method retards a high level of confidence. This algorithm set with different minimum support in every individual item based on its frequency that present in the database. To predict the medical oriental disease, Carlos et al [21] proposed a technique by using decision tree and another data mining technique was introduced by Gyu lee et al [23] proposed a computation model. A prediction system based on association rule was discovered by Carlos Ordonez found to be the best technique for heart disease diagnosis. Ibrahim et al represented enhanced weight based model for predicting heart disease in both eager and lazy methods [27,28, 29, 30]. Harleen et al presented artificial neural network based system to diagnosing the diabetic patient [20]. Palaniappan et al [22] proposed Navie bayes based automated system to predict the condition of heart disease in this model significant part depends upon algorithm to enhance the accuracy of classification from the provided datasets. According to background work, it prominently acquires definite issues that can be addressed for class rule generation [11, 18, 19]. The proposed algorithm appropriate for dataset consist of huge number of dimensions and needs high quality rules with minimal consuming period.

ALGORITHM FOR ATTRIBUTE SELECTION BASED ASSOCIATIVE CLASSIFICATION

A. Problem Description

This paper assumes that the training dataset *T* consist of set of items with h instances is represented by < AT_1, AT_2 …AT_h , C>, and |T| rows. Where AT_1, AT_2 …,AT_h are sequence of fields with C class Labels. A ruleset R has expressed in the form R -> C, where the left part of rule R may have one or more combination of itemset and in the consequent side must have class.

A rule R has to pass the min support threshold(min_supp) if for r, the supcount (r) / |T| ≥ minsupp, a rule r has to pass the minconf if supcount (r) / Appr (r) ≥ minconf.

Table 3.1 Sample Training Data

T. ID	**AT_1**	**AT_2**	**Class**
1	A_1	A_3	C_1
2	A_1	A_4	C_1
3	A_2	A_5	C_2
4	A_1	A_5	C_3
5	A_2	A_3	C_3

B. Attribute selection based Associative Classification

i) DIMENSIONALITY REDUCTION

Dimensionality refers to the number of input variables in the available dataset when the dataset consists of vast relative number of observations, and then some algorithms may struggle to train the effective models. To enhance the effectiveness to extract knowledge from high-dimensional datasets, dimensionality reduction methods are used. It also helps to reduce classification computation time and increase accuracy. Generally, the dimensionality reduction is divided into two categories. They are:

- Feature Extraction
- Feature Selection

Feature extraction aims to reduce the dataset’s dimensionality by building derived values from an initial set of measured data and then discarding the original features. These reduced features still accurately and completely describer the original dataset. Image processing, machine learning, pattern recognition, and natural language process- ing highly benefit from this technique.

The method of selecting the most significant variable in a dataset is known as feature selection. It influences most of the prediction feature through which it constructs the final model. It is a crucial step for effective algorithms in pattern recognition, classifi- cation, and regression.

Gain ratio is the dimensionality reduction technique by ignore all the irrelevant attributes from the mining process and retaining only the useful and important attributes that will improve the associative classification task and reduces computation. Cogency is the probability based controlling measure that bring important and rare rules from the dataset [23, 25, 29].

ii) SUBSET GENERATION

This phase of the proposed work generates a minimal number of high-quality rulesets by applying gain ratio in the subset generation. The test instance is considered as input and all possible subsets are generated. To reduce the computation and generate a high quality rules, gain ratio attribute selection method is used in this proposed algorithm and the detailed computation is presented below.

$$Gain\ Ratio\ (A) = Gain\ (A) / Split\ Info\ (A) \quad (1)$$

$$Split\ Info_A(D) = -\sum_{m=1}^{n} \frac{|D_A|}{|D|} * log_2\left(\frac{|D_A|}{|D|}\right) \quad (2)$$

$$Gain\ (A) = I\left(S_1, S_{2......}\ S_m\right) - E(A) \quad (3)$$

C. Subset Evaluation

Once the algorithm identifies significant attributes based on gain ratio, the subsets will be generated only to the selected attributes. Then, frequent class rules are identified by cogency computation. Lastly, the class label for the test instance is based on harmonic mean value.

HM (r) = 2* confidence (r) * cogency (r)/ confidence (r) + cogency (r)
(4)

3. 1 Proposed Algorithm (GRLLAC)

Input: M = Training data (m x n)
Input T = Test Instance
Output: Predicted class label for given test query
STEP1: The algorithm starts with testing dataset T
STEP2: Calculate the gain ratio for the given test instance
STEP3: Generate the subset from training dataset M only for the gain ratio attribute
STEP4: Calculate the cogency combination for selected attribute
STEP5: Set minimum cogency1 and cogency2 threshold.
STEP6: Calculate the Harmonic mean for all the generated rules
STEP7: Assign the class label based on the higher mean value attribute.

3. 1 Sample Computation

Let consider the medical dataset presented in Table 3.2, contains 12 training data and Table 3.3 contains test instance and the main task is to predict the class label for given test dataset. Among all the four fields in the training dataset ‘age’ attribute has been chosen for further subset generation since it has maximum gain ratio value.

The proposed algorithm then calculates the cogency for each subset of ‘Age’ attribute. The class label is assigned for the given test instance based on highest mean value.

Table 3.2 Training Dataset

Age	**Exc. Protocol**	**Smoking**	**Chol**	**Heart Disease**
Senior	Bruce	Yes	No	No
Senior	Bruce	Yes	Yes	No
Junior	Bruce	Yes	No	Yes
Youth	KOTTUS	Yes	No	Yes
Youth	Balke	No	No	Yes
Youth	Balke	No	Yes	No

Junior	Balke	No	Yes	Yes
Youth	KOTTUS	Yes	Yes	No
Senior	KOTTUS	Yes	No	No
Youth	KOTTUS	No	No	Yes
Senior	KOTTUS	No	Yes	Yes
Junior	Bruce	No	No	Yes

Table 3.3 Test Instance

Youth	KOTTUS	Yes	No	?

The below rules are extracted by the proposed lazy method for the given test instance as follows:

{Age=Youth}
{Age=Youth, Exc. Protocol=KOTTUS}
{Age=Youth, Exc. Protocol=KOTTUS, Smoking=Yes}
{Age=Youth, Exc.Protocol=KOTTUS, Smoking=Yes, Chol=No}

Table 3.4 Sample class computation for Age Attribute

ItemSet	**Class**	**Count**
Youth	Yes	2
	No	2
KOTTUS	Yes	4
	No	2
Smoking	Yes	3
	No	4
Chol	Yes	6
	No	2

Table 3.4 represents the sample computation of different class distribution of ‘Age’ attribute.

Probability of yes class=0.3*0.4*0.3*0.1= 0.232

Probability of no class = 0.4*0.360,2*0=0

Based on the above probability rate of both ‘yes’ and ‘no’ class, the yes class will be assigned for the testing instance since it has maximum harmonic mean.

EVALUATION OF THE ALGORITHM

To evaluate the efficiency and accuracy of the gain ratio based proposed algorithm, comparing the result with conventional popular associative classification assessment with various in balanced datasets of UCI [21] illustrate in table 3.5. For fare assessment different min_supp and cogency threshold was used. RapidMiner tool was used for data preprocessing and Java version 15 was used as implementation software. Various Holdout methods were applied in the chosen datasets for fare evaluation [26]. The algorithm's accuracy is illustrated in table 3.6.

Table 3.5 Description of the Datasets

UCI Datasets	**Number of Transactions**	**Number of Class labels**
Breast Cancer	286	2
Balance Scale	625	5
Breast-w	699	2
Credita	690	2
Diabetes	768	2
Glass	214	7
Ionosphere	351	2
Flare	1389	3
Iris	150	3
Heart Disease	303	3

Table 3.6 Accuracy Computation

UCI Datasets	**CBA**	**LLAC**	**GRLLAC**
Balance Scale	70.29	80.95	85.71
Breast Cancer	68.54	72.41	79.31
Breast-w	83.6	82.85	83.05
Credita	74.65	74.02	75.89
Diabetes	70.54	82.85	85.71
Glass	81.30	88.88	91.66
Ionosphere	90.56	73.33	80
Flare	74.68	79.71	84.05
Iris	90	93	95
Heart Disease	89	89	95.89
Average	**79.3**	**81.7**	**85.62**

In figure.1 shows the number of rules generated with heart disease dataset through proposed algorithm and other classification algorithms. The x axis represents different associative classification and y axis represents number of rules generated by each algorithm.

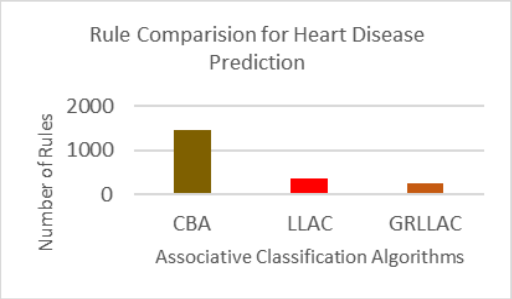

FIG. 1. COMPARISONS OF DIFFERENT RULE MINING ALGORITHMS

3. CONCLUSION

We proposed an innovative gain ratio based lazy learning associative classification in this paper. The algorithm has several notable features: (1) It reduces the number of subset generation, which leads to overall performance (2) it considers some interesting rules during the prediction phase. The proposed GRLLAC was tested with different unbalanced as well as minority classes datasets [21]. The experiment result shows that our proposed algorithm is notably effective and better accuracy in comparison with other Apriori based CBA and other lazy methods. The experiment results prove that the proposed attribute selection based lazy learning outperform all the traditional algorithms in terms of accuracy. As a future direction the proposed work may be tested with rare rule prediction model.

4. REFERENCES

[1] Agrawal R. and Srikant R. "Fast Algorithms for mining association rule" Proceedings of the 20th International Conference on Very Large Data Bases. pp. 487-499. 1994.

[2] Elena Baralis and Silvia Chiusano "A Lazy Approach to Associative Classification" IEEE Transaction on knowledge and Data Engineering, vol 20, No.2, pp. 156-171. 2008.

[3] Merschmann, L., Plastino, A.: A bayesian approach for protein classification. In: Proc. of the 21st Annual ACM Symposium on Applied Computing, Dijon, France, 2006.

[4] John G.H. and Langley P. "Estimating Continuous Distributions in Bayesian Classifiers" Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, pp. 338-345. 1995.

[5] Quinlan J. R. See5.0 (Http://www.rulequest.com) viewed on May 2010.

[6] Quinlan J. R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, San Francisco. 1993.

[7] Frank E. and Written I. "Generating accurate rule sets without global optimization" Proceeding of the 5th International Conference on Machine Learning, pp. 144-151. 1998.

[8] Syed S, and Chandran KR., "LACI: Lazy Associative Classification using Information Gain" International Journal of Engineering and Technology, vol.4, no.1, pp. 1-6, Feb 2012.

[9] Cendrowska J. "An Algorithm for inducing modular rules" International Journal of Man-Machine Studies. vol.27, pp. 349-370. 1987.

[10] Yeh C. and Lien C. "Cosmetics purchasing behavior-An analysis using association reasoning neural networks" Expert Syst. Appl., vol.37, no.10, pp.7219-7226, Oct. 2010.

[11] Furnkranz, J. and Separate-and-conquer rule learning. Technical Report TR-96-25, Austrian Research Institute for Artificial Intelligence, Vienna. 1996.

[12] Yin X. and Han J. " CPAR: Classification based on predictive Association rule" in SDM-2003.

[13] Li W. and Han J. " CMAR: Accurate and Efficient Classification based on Multiple-class Association Rule" in ICDM-2001; pp. 369-376. 2001.

[14] Thabtah F. and Cowling P. " MMAC: A New Multi-Class, Multi-Label Associative Classification Approach" 4th International Conference on Data Mining. 2004.

[15] Baralis E. and Torino P. “A Lazy approach to pruning classification rules” Proceedings of the IEEE International Conference on Data Mining. pp. 35-42. 2002.

[16] Liu, B., Hsu, W., and Ma, Y. “Integrating Classification and association rule mining” In KDD ’98, New York, NY, Aug. 1998.

[17] Li, W., Han, J., and Pei, J. CMAR: Accurate and efficient classification based on multipleclass association rule. In ICDM’01, pp. 369-376. 2001.

[18] David E. and Goldberg. “Genetic Algorithms in Search, Optimization and Machine Learning” 1989.

[19] Bahi M. “Parallel mining of association rules with a Hopfield type neural network” IEEE 2000.

[20] Yeh I. and Lien C. “Cosmetics purchasing behavior-An analysis using association reasoning neural networks” Expert System Application. Vol.37, pp. 7219-7226. 2010.

[21] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[22] Hecht-Nielsen R. “The mechanism of thought” International Proceeding of the Joint Conference Neural Network. pp. 419-426. 2006.

[23] Solari S. and Smith A. “Confabulation theory” Phys. Life Rev., pp. 106-120. 2008

[24] Siddique Ibrahim S.P., Dr.M. Sivabalakrishnan., “Lazy Learning Associative Classification in Map Reduce Framework” International Journal of Recent Technology and Engineering (IJRTE) (Scopus), Vol. 7, Issue 4S, pp. 168-172, Nov 2018.

[25] Azadeh Soltani M.R. and Akbarzadeh T. “Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets” IEEE Transactions on Neural Networks and learning systems. Vol.25, pp. 2055-2064. 2014

[26] Reitermanova Z., “Data Splitting” WDS’10 proceeding of contributed papers, part I, pp. 31-36, 2010.

[27] Siddique Ibrahim S.P., Sivabalakrishnan M.,” An Evolutionary Memetic Weighted Associative Classification Algorithm for Heart Disease Prediction” Recent Advances on Memetic Algorithms and its Applications in Image Processing. Studies in Computational Intelligence, Vol 873. Springer, Singapore, 2020.

[28] Siddique Ibrahim S P, Sivabalakrishnan M., "An Enhanced Weighted Associative Classification Algorithm without Preassigned weight based on Ranking Hubs" International Journal of Advanced Computer science and Applications, Vol. 10, No. 10, pp. 290-297, Nov 2019.

[29] Siddique Ibrahim S P, Sivabalakrishnan M., "Rare Lazy Learning Associative Classification using Cogency Measure for Heart Disease Prediction" Intelligent Computing in Engineering, pp. 681-691. Springer Vol. 1125 Apr 2020.