RESEARCH ARTICLE OPEN ACCESS

A General Survey on Associative Classification Techniques of Data Mining to Predict Diabetes Diseases

MD.SHAKEEL AHAMAD* ,Dr. N. SUDHAKAR**
(* Research Scholar, Dept. of CSE, AcharyaNagarjuna University, Guntur,.)
(** Principal &Professor inCSE,Bapatla Engineering College, Baptla.)

----------**********----------

Abstract:

Databases are prosperous with hid data which can be used for wisechoice making. Classification and affiliation rule mining are crucial to such sensible applications. Thus, if these two methods are somehowbuilt-in would result in wonderful savings and conveniences to the user. Such an integrated framework is referred to as associative classification (AC). This integration is carried out through focusing on a specific subset of association regulations whose consequent incorporates only categoryattribute. Several studies in statistics mining have proven that AC is ultimate to different usual classification algorithms due to its several favourable traits such as readability, usability, training efficient and extraordinary accuracy. Hence, a variety of AC methods for diabetes diseases are studied with its professionals and cons. However, AC suffers from a drawback that massive quantity of guidelines is produced as an output. Now, utilizing all these rules for evaluation would be computationally expensive. This paper studies a number of pruning and contrast methods that are employed to produce qualitative rules. Further, the paper empirically evaluates associative classification approach thinking about quite a number of parameters.

Keywords: Data Mining, Classification, Association, Associative Classification, Diabetic disease

----------**********----------

I. INTRODUCTION

Classification is one of the key methods in data mining that allocates objects in collection to target class. Various classification strategies such as Naive Bayesian classifier, Decision Tree, Neural Network, Associative Classification etc. [1]. The main goal of all such methods is to construct quickly and correct classifier.The classifier ought to be constructed such that a subset of the generated classification guidelines can be able to classify new objects or instances. It is regularly discovered that compared to different classification methods, associative classification outperforms in terms of accuracy which is extremely indispensable parameter [2].

Associative classification (AC) is a supervised classification method integrating association rule mining and classification [1]. The integration is done in order to get a special subset of association rules whose right-hand is restricted to classification class attribute [3], [4]. These subsets of rules are referred as Class Association Rules (CARs).

Diabetes is one of the serious health problems and there is growth average of infection people with this disease according to World Health Organization WHO in report 2016, with different kinds, children, women, men, young, old, everybody could be infected. One of the most importance issues to fight this serious disease is

the early fast diagnose, there is a set of a precise tests to diagnose diabetes, and if there are a lot of patient records many classification algorithms play great role to discover whether a person have diabetes or not.

Associative classification process basically comprises of following three steps

i) Association Rule Generation: In general, association rule mining can be viewed as a two-step process:

a) Finding all frequent itemsets from a given dataset satisfying predefined minimum support count for that particular dataset. Dataset consisting of categorical attributes can only be considered; in case of numerical or continuous attributes it requires to get converted into discretized form.

b) Generation of strong association rules from the frequent itemsets. Rules can be considered as strong association rules when they satisfies minimum support and minimum confidence. Various association rule algorithms namely Apriori, FP-Growth etc. exists for association rule generation [1].

ii)) Class Association Rules (CAR) Generation: Class association rules can be formed from association rules wherein right hand side of the rule consists of classification class attribute. Several associative classification algorithms subsist for CAR generation specifically Classification based on Association, Classification based on multiple class association Rules, Classification based on Predictive Association Rules etc. [2]

iii)) Pruning and Evaluation of CAR: Large numbers of CAR's are generated using associative classification algorithms. However, it would be computationally complex as well as ineffective too if all such rules are utilized for analysis. Different methods of pruning considering combination of confidence, support, cardinality, coverage, correlation methods etc. exist as well as various evaluation methods namely accuracy, robustness, scalability, interpretability etc. are available to assess and produce qualitative rules.

AC is an efficient method of classification and even several experimental studies [2], [5] have shown that AC is a promising approach due to following reasons:

a) Readability: The output of an AC algorithm is represented in simple if–then rules, which makes it simple, trouble free for the end-user to understand and interpret it.

b) Usability: Unlike decision tree algorithms, one can update or tune a rule in AC without disturbing the complete rules set, whereas the same task requires reshaping the whole tree in the decision tree approach.

c) Accuracy: Performance of associative classification is better than other traditional classification method like C4.5 as decision-tree classifier examines one variable at a time while association rules explores highly confident associations among multiple variables at a time.

d) Time-efficient & Training-efficient: Classification is done in quick manner. Training the data is very efficient regardless of the size of training set. Associative classification has abundant advantages for common people and is a boon to society which covers various applications [6],[7],[8],[9] such as: Recommended system: product recommendation in online shopping, Stock trading data: finding signals to sell and buy, Phishing Detection: distinguish phishing websites from legitimate ones, Automatic credit approval: identifying those transactions that are fraudulent, Medical field: Epidemics detection, Surveillance: Pattern discovery from surveillance systems etc. Some of the common issues of classification[10],[11],[12] are Incremental learning, Imbalance data stream classification, Dealing with non-static, unbalanced and cost-sensitive data , Security, privacy and data integrity, Distributed data classification, Classification of sequence as well as time series data

The organization of this paper is: This section covers the concept of associative classification along with its issues and applications. Section 2 provides overview and comparison of different AC techniques. Section 3 introduces pruning and evaluation methods for associative classification. Section 4 constitutes implementation and analysis. Section 5 includes conclusion and future work.

II.ASSOCIATIVE CLASSIFICATION STRATEGIES

Associative classification strategies differ mainly in the 2nd and 3rd step of associative classification

process depicted in this paper wherein CARs are generated considering different methods and ways of utilizing such rules so as to eliminate redundant or general rules and lead to qualitative or specific rules. Some of the basic techniques for associative classification are as follows:

i) Classification based on Association (CBA): CBA [13] uses an iterative approach to generate association rules considering apriori algorithm. Then, classifier is build by a heuristic scheme in which complete set of CARs are produced satisfying predefined minimum support and confidence and are arranged in a decreasing precedence based on their confidence and support. Rule pruning is carried out considering confidence, support and antecedent part of the rule. To classify a new tuple, decision is made based of the situation whether match is found or not. If match is found then: firstly the rule satisfying the tuple will be used to classify it otherwise the rule having the highest confidence is used. When neither of these is possible, the default rule will be utilized for classification process.

ii) Classification based on Multiple Association Rules (CMAR): CMAR [14] adopts a variant of the FP-growth algorithm to find the complete set of association rules satisfying minimum support and confidence thresholds. These rules are then examined, and a subset is chosen to represent the classifier. Pruning of rules is done based on confidence, correlation, and database coverage. For classification purpose, it considers multiple rules, rather than a single rule with highest confidence. If more than one rule satisfies a new tuple, X, the rules are divided into groups according to class Subsequently, CMAR uses statistical measure weighted chi-square to find the strongest amongst group of rules. As a consequence, biasing will be standing apart while predicting the class label of a new tuple.

iii) Classification based on Predictive Association Rules (CPAR): CPAR [15] follows the basic idea of First Order Inductive Learner (FOIL) [1] algorithm in rule generation. The resulting rules are merged to form the classifier rule set. Rules are pruned considering Laplace accuracy measure. CPAR follows the same approach as CMAR whenever more than one rule satisfies a new tuple for classification. On the other hand, CPAR uses the best k rules of each group to predict the class label of new tuple, based on expected accuracy unlike CMAR so that much better efficiency can be achieved.

iv) Classification based on Association Rules Generated in a Bidirectional Approach (CARGBA): CARGBA [16] is essentially a bidirectional rule generation approach that generates not only general but specific association rules too. General rules are produced by apriori approach and specific rules are generated by considering the larger length of respective rule in order to generate specific details wherein support would apparently lower. Then, classifier is build by the construction of final rule set consisting of essential rules formed by the mutual mixture of both the rules by taking confidence, support and length of the rule into consideration. Measure such as correlation coefficient is used for pruning of such rules. When a new tuple is to be classified, the classifier classifies according to the first rule in the final rule set is formed that covers the new tuple.

III.PRUNING AND EVALUATION METHODS

AC provides better accuracy as compared to other classification techniques. However, it suffers from a drawback that large numbers of class association rules are generated. Thus, it is extremely indispensable to produce qualitative rules amongst bulky amount of rules. As a consequence, pruning is performed to produce effectual rules and to reduce computational overhead. Various pruning strategies [9], [17], [18] are as follows:

a) Confidence and Support: Most common pruning strategy firstly considers confidence threshold. If there are two rules having the same confidence level then support is taken into concern. Subsequently, if confidence and support are same then the rule which is generated earlier or first is taken into contemplation.

b) Confidence, Support, and Rule Cardinality Procedure: In this strategy, along with confidence and support, rule cardinality is considered. Thus, the rule having longest antecedent part will be used to resolve the conflicts in case of same confidence and support threshold.

c) Database coverage: Given a tuple X, from a class labelled data set D, let n_{covers} be the number of tuples covered by R and |D| be the number of tuples in D.

Database coverage can be defined as : Coverage(R) = n_{covers} / |D|

d) Correlation method: For categorical (discrete) data, a correlation relationship between two random variables can be revealed by aχ^2 (chi-square) test. The χ^2value (also known as the Pearson χ^2 statistic) [1] is computed as:

$$\chi 2 = \sum_{i=1}^{c} \sum_{j=1}^{r} (oij - eij)2 / eij$$

Where oi j is the observed frequency (i.e., actual count) and ei j is the expected frequency. Besides pruning, evaluation of rules is also required for testing the effectiveness of each rule and hence forming an efficient rule set. Several different criteria's have developed for assessment of rules so as to find the most significant set of rules. Moreover, an essential task in classification is to measure the quality of classifiers which can be conceded by various ways for producing the effective results. Some of the parameters [1][12] are as follows:

a) Accuracy: Rule's accuracy can be measured by looking the tuples that it covers and make out what percentage of them, the rule can correctly classify. Accuracy [1] can be defined as: Accuracy = n_{covers} / $n_{coverage}$

b) Robustness: An ability to make correct predictions given noisy data or data with missing values.

c) Speed: It refers to the computation costs involved in generating the rules and utilizing it.

d) Scalability: producing more efficient rules given large amount of data can be termed as scalability.

e) Interpretability: It can be termed as a capability to understand easily and straightforwardly exclusive of any dilemma.

IV. PROCEDURE OF IMPLEMENTATION

In AC, association rules and subsequently class association rules are produced. In this paper, association rules are generated using apriori algorithm and CARs are produced utilizing associative classification method namely CBA. However, large numbers of CARs are produced by CBA and therefore pruning is carried out utilizing parameters such as confidence, support and coverage to find the reduced set of rules. The work has been implemented using Windows operating system utilizing MATLAB tool. The generated set of CARs is evaluated by measuring the accuracy to build the appropriate classifier. Then, the induced classifier will be tested on unseen instances to assess the performance of a classifier. Various real datasets available in UCI data repository [19] are taken into consideration for implementation of associative classification technique and its details are described in Table 2 below:

TABLE 2: DATASET COMPOSITION

Name of Datasets	# Attributes	# Instances	# Class
Mammography	6	961	2
Contraceptive Method Choice (CMC)	9	1473	3
Adult	11	30719	2
Bank Marketing	17	45211	2

Dataset description :

a) Mammography is the most effective method for breast cancer screening available today. This dataset can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age.

b) In CMC, the problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socioeconomic characteristics.

c) The Adult dataset predicts whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

d) Bank Marketing: The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Pre-processing techniques namely discretization and sampling were carried out on these dataset. The discretization process has been conceded on the attributes having continuous values as the algorithm used for associative classification considers only discrete values. Discretization has been done using WEKA tool. Moreover, to solve the problem of imbalance class in certain above datasets, sampling techniques specifically undersampling or over-sampling was used as and when required.

The first step for associative classification process i.e. generation of association rule is carried out using

apriori algorithm. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases [1]. Then, class association rules are produced using algorithm namely CBA as a second step of associative classification process by varying the two measures namely support and confidence.

Firstly, class association rules (CARs) are generated by keeping the confidence threshold uniform to 0.6 and varying the support threshold on four different datasets. Then, CARs are generated by keeping the support threshold uniform to 0.2 and varying the confidence threshold. The results obtained are as shown below:

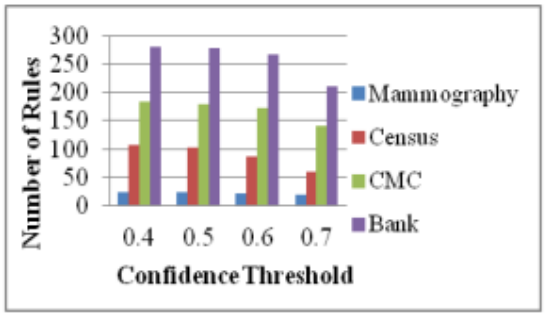

Figure 2: Confidence threshold and number of rules

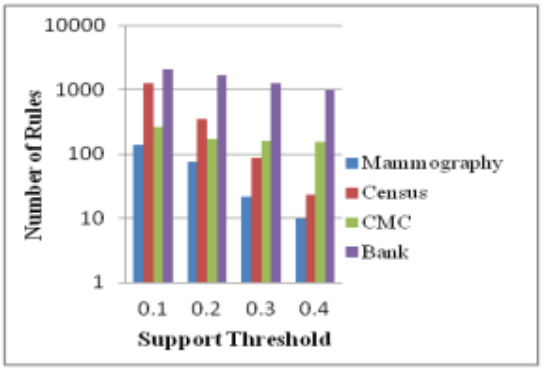

Figure 3: Support threshold and number of rules

In figure-3, it is clearly depicted that as the support threshold increase, the number of rules generated also increases for all the four datasets. Moreover, it is seen that a dataset having large number of instances produces more number of rules. However, it can be concluded that at support threshold 0.2 neither too bulky nor too minute but appropriate numbers of rules are produced required for classification of data on various datasets.

Then, pruning of rules is done considering support as well as confidence and accuracy has been considered as an evaluation parameter. However, it has been found that redundant rules reduce the accuracy. So, such rules are eliminated and below results are taken after elimination.

Now, accuracy has been determined by varying the support threshold and keeping the confidence level constant as 0.6 on all the datasets. Also, by varying the confidence threshold and keeping the support threshold constant as 0.2 evaluation of all the dataset is done considering the accuracy parameter. The results obtained are as shown below

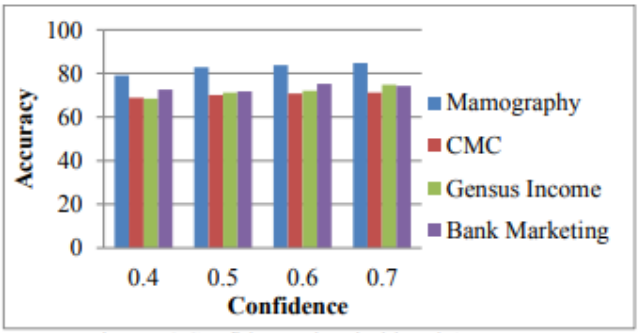

Figure: 4 Confidence threshold and Accuracy

In figure-4, it can be seen that considering the threshold level of confidence as 0. 6 and support set as 0.2, appropriate level of accuracy has been achieved as well as reasonable number of CARs is produced.

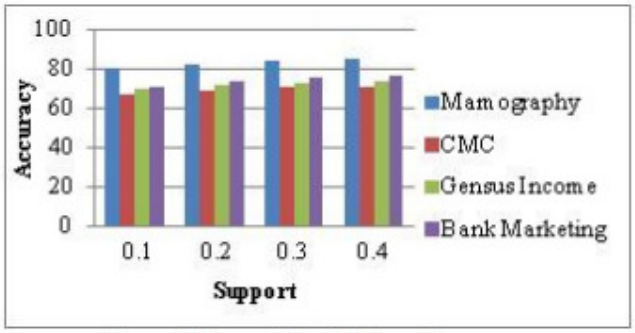

Figure 5: Support threshold and Accuracy

In figure-5, it can be seen that considering the threshold level of support as 0.2 and confidence set as 0.6, appropriate level of accuracy has been achieved as well as reasonable number of CARs is produced. Optimization approaches for data mining [18] essentially includes genetic algorithms, evolutionary search, simulated annealing, branch-and-bound, logical analysis of data, and mathematical programming Finding the optimal associative classification can be treated as a

combinatorial optimization rather than mathematical programming that can be accomplished by genetic algorithms [19]. However, as the associative classification considers discrete values genetic algorithms are preferred for optimizing the parameters which are considered for evaluation.

.

V .CONCLUSION AND FUTURE WORK

This paper studies existing techniques for Associative Classification which is an active area of research in prediction of diabetic using data mining. The paper also studies pruning methods for eliminating inefficient rules and various measures used to evaluate the associative classification rules. It has been concluded that CBA is a simple and efficient technique that produces accurate rules. Further, the paper tries to find optimal support and confidence parameters for associative classification with respect to number of rules as several applications demand interpretable results which is dependent on the number of rules. The obtained parametric values can be used to further improve associative classification. As a part of future work, classification in modern data mining fields such as privacy-preserving data mining, data stream mining, spatial data mining, etc. can be addressed using associative classification technique. However, optimization of associative classification can be carried out by considering genetic approach.

REFERENCES

1) J. Han and M. Kamber, "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers.

2) P. Garach, D. Patel, R. Kotecha, " Privacy-Preserving Associative Classification" in International Conference on Information and Communication Technology for Intelligent Systems (ICTIS), Springer; (March 2017).

3) S. Gambhir, N. Gondaliya, "A Survey of Associative Classification Algorithms", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 9, (2012).

4) X. Li, D. Qin, C. Yu, "ACCF: Associative Classification Based on Closed Frequent Item sets", Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery– FSKD, 380-384 (2008).

5) P. Shekhawat, S. Dhande, "A classification technique using associative classification", International journal of computer application; vol. 20-No.5, pp 20-28 (2011).

6) Y. Pan, X. Ding, "Anomaly Based Web Phishing Page Detection", Proceedings of the 22nd Annual Computer Security Applications Conference; IEEE, 381-392 (2006).

7) Y. Chien, Y. Chen,"Mining associative classification rules with stock trading data – A GA-based method", Knowledge-Based Systems; vol.23, 605–614 (2010).

8) Y. Jiang, J. Shang, Y. Liu "Maximizing customer satisfaction through an online recommendation system: A novel associative classification model", Decision Support Systems, Vol.48, 470- 479 (2010).

9) D. Sasirekha, A. Punitha, "A Comprehensive Analysis on Associative Classification in Medical Datasets", Indian Journal of Science and Technology (IJST), vol. 8, 1-9 (2015).

10) Q. Yan., X. Wu, "10 Challenging problems in data mining research" International Journal of Information Technology & Decision Making, World Scientific vol. 5, 597-604, (2006).

11) B. Krawczyk, "Learning from imbalanced data: open challenges and future directions", Progress in Artificial Intelligence, Springer vol. 5, 221-232 (2016).

12) F. Thabthah, "A review of associative classification mining", Knowledge Engineering Review vol. 22, 37-65, (2007).

13) B. Liu, W. Hsu, Y. Ma, " Integrating Classification and Association Rule Mining" ,Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, AAAI, New York; 80- 86 (1998).

14) W. Li, J. Han, J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules" Proceedings of the IEEE International Conference on Data Mining, IEEE, 369-376, (2001).

15) X. Yin, J. Han, "CPAR: Classification based on Predictive Association Rules", Proceedings of the SIAM International Conference on Data Mining, San Francisco 369-376, (2003).

16) K. Gourab, M. Sirajum, F. Islam, M. Murase, K. Ishikawa, M. Doya, K. Miyamoto, H. Takeshi "A Novel Algorithm for Associative Classification", International Conference on Neural Information Processing, (ICONIP), Kitakyushu-Japan, Springer; vol.14, 13-16(2007).

17) S. Wedyan, "Review and Comparison of Associative Classification Data Mining Approaches", International Journal of Computer, Electrical, Automation, Control and Information Engineering; World Academy of Science, Engineering and Technology vol.8 (2014).

18) F. Thabtah, P. Cowling, S. Hammoud "Improving rule sorting, predictive accuracy and training time in associative classification" Expert Systems with Application, Elsevier, vol. 31,414-426, (2013).

19) M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2013).