ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 6, Issue 1, January 2017

A Survey On Different Association Classification Technique For Constructing Efficient Classifier

S.P. Siddique Ibrahim, M.Pavithra, Dr. M. Sivabalakrishnan

Abstract— **Associative classification (AC) is a data mining approach it's a combination of both association rule and classification to build classification models (classifiers).More number of associative classification algorithms have been stated such as Classification based Association (CBA), Classification based on Multiple Association Rules (CMAR), etc .This paper surveys major AC algorithms and compares the steps and methods performed in each algorithm with CBA, CMAR and MCAR.**

Index Terms-- **associative classification, data mining, classification**

I. INTRODUCTION

DATA mining is the process of examine large set of data and extracting hidden patterns from different data types in order to find previously unknown design. The discovery process can be an automatic or semi-automatic [1]. For decision making data mining is the knowledge discovery in the database and the KDD main steps are the data selection, data pre-processing, transformation, data mining, and evaluation. Data mining tasks including classification, clustering, association rule discovery, pattern recognition, regression, etc. [2]

There are two type of learning model available in data mining such as supervised and unsupervised. In supervised learning, it contain the class label .For example, in credit card scoring application, the goal is to whether the financial institution should issued a credit card to the client or not to the client. On the other hand, training data set with no class attribute is considered as unsupervised learning. Association rule discovery task [3] is a common example of unsupervised learning where the aim is to discover the correlations among items. AC is often capable of building efficient and accurate classification systems, since it utilizes association rule discovery methods in the training phase [4] which finds all possible relationships among the attribute values in the training data set.

S.P. Siddique Ibrahim *,Department of Computer Science and Engineering, Kumaraguru college of Technology, Coimbatore-641049.*
M.Pavithra, *Department of Computer Science and Engineering, Kumaraguru college of Technology, Coimbatore-641049.*
Dr. M. Sivabalakrishnan **,** *Associate Professor,School of Computing Science and Engineering, VIT University, Chennai.*

AC algorithms have been proposed including: CBA [4], CMAR [7], CPAR [8] MCAR [9]. This paper deal with the main concept of the associative classification, and also gives an example to establish its main steps.

Classification rule mining aims to discover a small set of rules in the database that forms an accurate classifier. In association rule mining rule will be constructed based on the support and confidence . There is no predetermined class. Main concept of this paper is used use the CBA algorithm and compare with C4.5[4].

For unstructured data handling c4.5 and CBA algorithm used ; for huge datasets its not efficient so CMAR[7] algorithm is used and compare with c4.5 and CBA. It extends an FPtree, frequent pattern mining method. This CP tree applied this method to store and retrieve. It contain two phases as Rule generation and classification.

Rule generation: R: P→C P pattern, C class label this rule will be based on the support and *confidence* , high quality rules will be taken.
Classification: Classification based on the given data object, subset of rule will be extracted and predict the class label of the object by analyzing these subset of rules. After rule generation new object class label will be assigned and to measure the high values of rule.

II. BASIC CONCEPT OF SUPPORT AND CONFIDENCE

A. Support and Confidence

Support [5] is defined as the percentage of transactions in the data that contain all items in both the antecedent and the consequent of the rule,

$$S=P(X\cap Y) = \{X\cap Y\}/\{D\}$$

Confidence is estimates of the conditional probabilities of *Y* given *X*, i.e. $P(X\cap Y)/P(X)$.

$$C= P(X\cap Y)/P(X)$$

The support of a rule is also important since it indicates how frequent the rule is in the transactions. A rule that has a very high confidence (i.e., close to 1.0) is very important because it provides an accurate predict ion on the association of the items in the rule.

III. DATA MINING APPLICATIONS

Data mining is used in different ranges such as retail industry, telecommunication industry, healthcare industry, financial data analysis, intrusion detection, sports etc[6].

A. *RETAIL INDUSTRY*

data mining is huge application in retail industry as it collects big amount of data which includes transportation, sales and consumption of goods and services Data mining helps to identify customer's buying devices and direction that lead to enhanced the condition of customer service and customer's satisfaction.

B. *TELECOMMUNICATION INDUSTRY*

Telecommunication industry is the most spreading industry as it procure various services such as fax, pager, cellular phones and e-mails. Data mining helps to identify telecommunication design, fraud activities, make better use of effects and improve quality of service.

HEALTHCARE INDUSTRY:

C. Data mining is very useful in healthcare industry in diagnosis of heart diseases, breast cancer and diabetes. It helps in analyze designs and trends in patient's records having same risk factor and helps in decision making.

D. *FINANCIAL DATA ANALYSIS:*

It helps in loan payment prediction and customer credit policy analysis. It also helps in collecting of customers for target marketing.

E. *INTRUSION DETECTION:*

intrusion is any kind of action that enforce the confidentiality or integrity of network effects from any outside party. With the increased usage of internet and opportunity of the tools and fraud of intrusion and blasting network, intrusion detection has become an critical issue for network administration. Data mining helps in the development of data mining algorithm for intrusion exposure and investigation of stream data so that intrusion threats can be evade.

F. *SPORTS:*

In sports, broad amount of statistics are collected for each player, team, game and season. Data mining is used in the prediction of completion of players, selection of players and forecast of future events

IV. ASSOCIAIOTION RULE MINING

An association is an indicating expression of the form X→Y, where X and Y are disjoint item sets. Support determines how frequently a rule is applicable to given datasets, while confidence determines how frequently items in Y appear in transaction that contain X.

Example 3.1

Television→setup box [supp = 5%, confidence = 80%]

80% of customers who buy a television also buy a setup-box and 5% of customers buys all these products together

V. ASSOCIATIVE CLASSIFICATION MINING

Associative Classification (AC) is a natural classification learning advance in data mining that adopts association rule encounter methods and classification to build the classification models. Associative classification trust in mainly on two important thresholds called minimum support (*MinSupp*) and minimum confidence (*MinConf*). Minimum support produce the frequency of the attribute value and its associated class in the training data set from the capacity of that data set. Whereas minimum confidence produce the frequency of the attribute value,

Row id	**A**	**B**	**C**	**Class label**
1	**A1**	**B1**	**C1**	**A**
2	**A1**	**B2**	**C1**	**B**
3	**A2**	**B3**	**C2**	**A**
4	**A1**	**B2**	**C3**	**A**
5	**A1**	**B2**	**C1**	**C**

Table 1 training data set

A. RULE GENEARTION

The key operation of CBA-RG is to find all ruleitems that have support high minsup. A rule item is of the form:

<condset, y>

where condset is a set of items, y € Y is a class label. The support count of the condset (called condsupCount) is the number of compact in D that contain the condset The support count of the rule item (called rulesupCount) is the number of cases in D that contain the condset and are labeled with class y. Each ruleitem essentialy represents a rule:

condset → y,

Rule items that satisfy minsup are called frequent rule items, while the rest are called infrequent rule items.

B. BUILDING A CLASSIFIER

This section instant the CBA-CB algorithm for construction a classifier using CARs . To yield the good classifier out of the whole set of rules would involve calculating all the possible subsets of it on the training data and choosing the subset with the right rule sequence that gives the least number of errors. [4]

Step 1: among the rule generation which rule having a highest confidence that would be considered as a best rule generation(L1>L2)

Step 2: if the two rule having same confidence then considered the support count (L1=L2)

Step 3: If the support count is also same consider which rule will appeared first, whether L1 is earlier than L2.

C. PREDICTION

Data mining automates the process of awarding predictive information in huge databases and achievement was positive. To create association rules for each of them particularly would give rise first, the memory space employed by these rules can be many times larger than the original database, second, identifying the most applicable rules and connecting their sometimes conflicting predictions may easily obtain excessive computational costs. A decision support system encourage a medical doctor about which other diseases may lead the ones already recognize can help in the selection of the most important new tests.

Firstly the rule in the classifier are rated by support, confidence and cardinality . The prediction is based on either a single rule which contest the new Data and has the highest priority or different rules that are all suitable to ne data.

D. ACCURACY CALCULATION

Accuracy measures the ratio of correct prediction to the total number of cases evaluated. Calculations of accuracy for model is

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}$$

TN is the number of true negative calls
FP is the number of false positive calls
FN is the number of false negative calls
TP is the number of true positive calls

Example :

[1]	[2] Predictive negative	[3] Predictive positive
[4] Negative cases	[5] 9700	[6] 150
[7] Posi4ive cases	[8] 50	[9] 100

$$\text{Accuracy} = \frac{9700 + 100}{9700 + 150 + 150 + 100}$$

$$= 98.0\%$$

VI. VALIDATION TYPES

There are various types of validation technique available in data mining such as

1. Holdout method
2. Cross validation
3. Leave one out method

A. HOLD OUT METHOD

The hold out method store a secure amount for testing and benefits the remainder for training , usually one third for testing , the rest for training.

REPEATED HOLD OUT METHOD

Hold out measure can be made more honest by duplicating the process with different subsambles. In each Iteration a certain ratio is nevertheless selected for training this is called repeated hold out method.

B. CROSS VALIDATION

The dataset is divided into k subsets of identical size and each subset in turn is used for testing and the remainder for training this is called k fold cross validation.

C. LEAVE ONE OUT

Loo is a simple cross validation each learning set is generated by taking all the samples rejecting one , the test set being the sample abandoned. We have n different training set and n different test set. This cross validation steps does not decay much data as only one sample is left from the training set.

VII. CONCLUSION

This paper surveys major AC algorithms and prove that it produces more accuracy than c4.5.
Many AC algorithms have been successfully used to generate accurate classifiers, such as CBA, MCAR, CMAR and CPAR, where only the most clear class interact to a rule is created and other classes are simply neglected. The main challenge of the researcher is to minimize the association rules and classification error in different datasets.

VIII. REFERENCES

1) Suzan Wedyan "review and comparison of associative classification data mining approaches", International Journal of Computer, Control, Quantum and Information Engineering Vol:8, No:1, 2014.

2) Fayyad, U., and Irani, K. (1993) Multi—interval discretization of continues-valued attributes for classification learning. Proceedings of IJCAI, pp. 1022-1027. 1993.

3) Agrawal, R., Imielinski, T., Swami, A (1993) Mining Association Rules between Sets of Items inLarge Databases. In: Proc. of the ACM SIGMOD Conference, pp. 207–216, 1993.

4) Liu, B., Hsu, W., and Ma, Y. Integrating classification and association rule mining. Proceedings of the Knowledge Discovery and Data Mining Conference-KDD, pp. 80-86. New York, NY 1998.

5) Varshali Jaiswal and Jitendra Agarwal"The evolution of the association rules",International Journal of Modeling and Optimization, Vol. 2, No. 6, December 2012

6) Divya kundra, er. navpreet kaur"review on prediction system for heart diagnosis using data mining techniques",international journal of latest research in engineering and technology (ijlret) issn: 2454-5031,volume 1 issue 5| pp 09-14|october 2015.

7) Li, W., Han, J., and Pei, J. (2001) CMAR: Accurate and efficient classification based on multiple-class association rule. Proceedings of the IEEE International Conference on Data Mining –ICDM, pp. 369- 376, 2001.

8) Yin, X., and Han, J. (2003) CPAR Classification based on predictive association rule. Proceedings of the –the SIAM International Conference on Data Mining -SDM, pp. 369-376, 2003.

9) Thabtah, F., Cowling, P., and Peng, Y. (2005) MCAR: Multi-class classification based on association rule approach. Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications (pp. 1-7).Cairo, Egypt.