ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 2, February 2017

Survey of Associative Classification Techniques for Text Mining

[1]Asst. Prof. Ms. Shweta M. Kambare, [2]Asst. Prof. Mrs. Varsha A. Jujare, [3]Asst. Prof. Mr. Atul A. Kumbhar
.Sharad Institute of Technology College of Engineering

***Abstract*- Amount of unorganized text data is increasing day by day as the use of internet is increasing. Proper classification and knowledge discovery from these documents is an important area for research. Approximately 80% of the information of an organization is stored in unstructured textual format, in the form of reports, email, views and news etc. So there is need of automatic retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis. Associative classification is one of the most efficient techniques for text classification. Associative classification is integration of association rule mining and classification rule mining. In this paper, different techniques of associative classification are discussed in brief.**

***Index Terms*— Associative classification, Association rule mining, Class association rules, Text classification, Text mining**

I. INTRODUCTION

Text classification

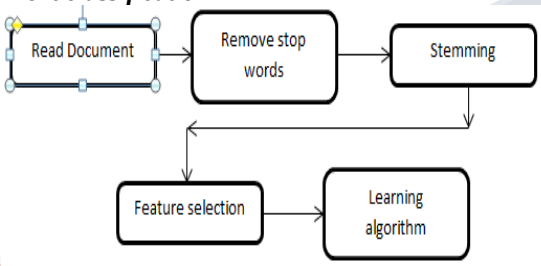

Fig. 1.Text Document Classification block diagram

Figure 1 describes steps involved in text document classification. Stop words are words which don't have much importance. E.g. am, is, are, was etc. They are removed from the document. Then remaining words are converted to their root forms by stemming. E.g. Removed – remove. In the feature selection step, important words are chosen to represent the document in vector form. In the last step, learning algorithm is applied to build classifier.

Text classification techniques

There are various techniques for text document classification such as Machine learning algorithms, semantic classification, Associative classification etc. Associative classification is one of the most efficient techniques for text classification. In Association rule mining, all the rules existing in the database that satisfy some minimum support and minimum confidence thresholds are found. For association rule mining, the target of discovery is not pre-determined, while for classification rule mining there is one and only one predetermined target. Integrating these two techniques gives associative classifier. The integration is done by focusing on mining a special subset of association rules, called class association rules (CARs). The classifier built this way is, in general, more accurate than traditional classification techniques.

II. ASSOCIATIVE CLASSIFICATION

In associative classification, the training data set T has m distinct attributes A_1, A_2,...., A_m and C is a list of classes. The number of rows in T is denoted by |T|. An attribute may be categorical (where each attribute takes a value from a finite set of possible values) or continuous where each attribute takes a value from an infinite set (e.g. reals or integers). For categorical attributes, all possible values are mapped to a set of positive integers. In the case of continuous attributes, a discretization method can be used.

Definition 1. An item can be described as an attribute name A_i and its value a_i, denoted (A_i, a_i).

Definition 2. The j th row or a training object in T can be described as a list of items (A_{j1}, a_{j1}), . . ., (A_{jk}, a_{jk}), plus a class denoted by cj.

Definition 3. An itemset can be described as a set of disjoint attribute values contained in a training object, denoted {(A_{i1}, a_{i1}), . . ., (A_{ik}, a_{ik})}.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 2, February 2017

Definition 4. A ruleitem r is of the form {cond, c_i} where condition cond is an itemset and c ε C is a class.

Definition 5. The actual occurrence (actoccr) of a ruleitem r in T is the number of rows in T that match r's itemset.

Definition 6. The support count (suppcount) of ruleitem r = {cond, c_i} is the number of rows in T that matches r's itemset, and belongs to a class c.

Definition 7. The occurrence (occitm) of an itemset I in T is the number of rows in T that match I.

Definition 8. An itemset i passes the minimum support (minsupp) threshold if (occitm(i) / |T|) >= minsupp. Such an itemset is called a frequent itemset.

Definition 9. A ruleitem r passes the minsupp threshold if, suppcount(r)/ |T| >= minsupp. Such a ruleitem is said to be a frequent ruleitem.

Definition 10. A ruleitem r passes the minimum confidence (minconf) threshold if suppcount(r) / actoccr(r) >= minconf.

Definition 11. A rule is represented in the form: cond → $c_1 \vee c_2 \vee \ldots c_j$, where the left hand side of the rule (antecedent) is an itemset and the right hand side of the rule (consequent) is a list of class labels.

Rules are generated from training dataset by using the frequent rule-items. Frequent rule-items are rule-items which satisfy minimum support threshold. The rules which satisfy minimum confidence threshold are included in the classifier.

In associative classification, the item to be classified is matched with the rule's antecedents and is given the label of matching antecedent.

III. DIFFERENT ASSOCIATIVE CLASSIFICATION TECHNIQUES

CBA (Classification Based On Associations)

Liu proposed an algorithm called CBA [1] that integrates association rule mining with classification. CBA operates in three main steps. First, it discretizes real/integer attributes and second, it uses the apriori approach of Agrawal and Srikant to discover frequent itemsets and generate the rules. Finally, a subset of the rules produced is selected to represent the classification system. The discovery of frequent item sets is the most resource- and time-consuming step in CBA, because it requires multiple passes over the training data. In each pass, the seed of the rule items found in the previous pass are used to generate potential rule items in the current pass. The experimental results show that CBA scales well with regard to error rate if compared with decision trees.

CMAR (Classification Based Multiple Association Rules)

F. A. Thabtah et al. Li et al. developed the CMAR [2] algorithm that uses the FP-growth approach to find frequent item sets. CMAR differs from other associative methods since it uses more than one rule to assign a class to a test object and stores the classification rules in a prefix tree data structure, known as a CR-tree. When classifying a new object, CMAR accumulates the subset of classification rules matching the new object and looks at their class labels. In the case where all rules have a common class, CMAR simply assigns that class to the test object. In cases where the classes of the accumulated rules are not identical, CMAR divides the rules into separate groups based on their class values and compares the effects of every group to identify the strongest one.

The effectiveness of a rule-based classifier on test data sets that contain missing values has been studied by Li. The authors introduced two simple methods, one extends decision trees and the other extends optimal class association rules. The predictive power of the two methods was compared with popular classification methods, like CBA and decision trees, on test data that contain missing values. It was claimed that proposed association rule method generally derives a smaller set of rules that is able to predict test data objects with missing values. The k-optimal rule set has been introduced to obtain further minimal effective rule-based classifiers. The results reported on four data sets showed that the optimal association rule method is competitive with traditional classification methods based on decision trees.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 2, February 2017

CPAR (Classification based on Predictive Association Rules)

A greedy associative classification algorithm called CPAR [3], which adopts the FOIL strategy to generate rules, was proposed by Yin and Han. CPAR seeks the best rule condition that brings most FOIL gain among the available ones in the data set. FOIL gain is used to measure the information gained from adding a condition to the current rule. Once the condition is identified, the weights of the positive examples associated with it are reduced by a multiplying factor, and the process repeats until all positive examples in the training data set are covered. The search for the best rule condition is the most time-consuming process of CPAR, since the gain for every possible item needs to be calculated to determine the best overall gain. In the rule-generation process, CPAR derives not only the best condition but also all similar ones, since there is often more than one attribute item with similar gain. It is claimed that CPAR improves the efficiency of the rule-generation process when compared with popular associative classification methods such as CBA.

A recently proposed approach for building classification systems based on both positive and negative rules has been introduced by Antonie and Zaiane. The "interestingness" of the rules for the proposed algorithm is based on the correlation coefficient that measures the strength of the linear relationship between pairs of variables. In addition to confidence and support thresholds, the correlation coefficient has been used to prune the final rules in the classifier, giving a much reduced rules set. The way in which the algorithm generates the rules is similar to an a priori method where multiple scans are required to discover the rules. Ranking occurs in a similar way as the CBA rules ranking method, e.g. confidence, support. Experimental tests on six data sets from the UCI data collection showed that negative association rules are useful when used with positive rules for producing competitive classification systems.

IV. MULTIPLE LABELS ASSOCIATIVE CLASSIFICATION

The proposed algorithm in [4] can be considered an associative classification technique since it uses the core concepts of association rule mining (support and confidence), in a classification framework. MMAC presents the idea of extracting rules with multiple labels and has many distinguishing features over current associative classification algorithms. It generates set of rules in each iteration and later combines the rule set to give a multi-label associative classifier.

ACRI (Associative Classifier with Reoccurring Items)

ACRI [5] consists of two modules: Rule generator and classifier. The algorithm is base for mining associations with reoccurring items on Apriori-based MaxOccur. The building of the classification model follows ARC-BC approach. Other associative classification methods are biased towards dominant classes in the case when rare classes exist. Rare classes are classes with very few representatives in the training set. MaxOccur run on transactions from each known class separately makes the core of rule generator module. It mines the set of rules with reoccurring items from the training set. These rules associate a condition set with a class label such that the condition set may contain items preceded by a repetition counter. Different classification rules may match, thus the classifier module applies diverse strategies to select the appropriate rules to use. In other associative classifiers, a default rule is applied, either the rule with the highest confidence in the model or simply assigning the label of the dominant class in case of the antecedent of the rules don't match. In ACRI approach partial matching or closest matching by modeling antecedents of rules and new objects in a vector space is allowed.

Considering repetitions of observed features is beneficial. ACRI is less sensitive, with respect to accuracy, to the support threshold. In this approach the similarity between terms is not considered for classification.

V. CONCLUSION

Basics of Associative classification and different techniques of associative classification such as CBA(Classification Based on Associations), CMAR, CPAR, MMAC, ACRI are discussed above. Associative classifier provides a great classifier for unstructured data such as text. All the above discussed techniques for associative classification are sensitive to the support threshold except ACRI. Since the accuracy depends on the threshold. ACRI is less

sensitive to the support threshold with respect to accuracy.

REFERENCES

[1] Liu, B., Hsu, W., Ma, Y. "Integrating classification and association rule mining" Proceedings of the KDD, (pp. 80-86). New York, NY. (1998)

[2] Wenmin Li Jiawei Han Jian Pei" CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules" Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference

[3] Yin, X., Han, J "CPAR: Classification based on predictive association rule" Proceedings of the SDM p. 369-376. San Francisco, CA. (2003).

[4] Thabtah, F., Cowling, P., Peng, Y "MMAC: A new multi-class, multi-label associative classification approach" Proceedings of the Fourth IEEE (2004). International Conference on Data Mining (ICDM '04), (pp. 217-224). Brighton, UK. (Nominated for the Best paper award).

[5] Rafal Rak, Wojciech Stach, Osmar R. Za¨ıane, and Maria-Luiza Antonie "Considering Re-occurring Features in Associative Classifiers" Springer-Verlag Berlin Heidelberg(2005).

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