Optimized Association Rule Based Text Categorization with (OARTC) Algorithm

[1]K.Gayathri, [2]Dr.A.Marimuthu MCA, M.Phil, Ph.D

Abstract

Text based classification is the process to classify the documents into pre-defined categories based on their content. The task of data mining is to classify the documents automatically into predefined classes based on their content. Existing supervised learning algorithms are to classify text need sufficient documents to learn accurately. This paper presents a new algorithm for text classification using data mining that requires fewer documents for training. Instead of using words, word relation, i.e the association rules from these words, is used to derive feature set from pre-classified text documents. News Group data set consists of twenty thousand message is one which is widely used. Calculate the distance to classify the test samples. Before classification initially the reduced feature set is received from TF/IDF method which was discussed already in the our earlier work. Finally, the association rules are formed by Aprior Algorithm and term sets are also formed.

Keywords : Text Categorization, Association rule, Aprior Algorithm.

[1]Research Scholar of Computer Science
Karpagam University, Coimbatore, India
Sasmithagp@gmail.com
[2]PG & Research Dept. of Computer Science,
Government Arts College,
Coimbatore, India mmuthu2005@gmail.com

I. Introduction

The volume of information is continued to increase, better finding, filter and manage these resources. Text categorization is the assignment of natural language documents to one or more predefined categories based on their semantic content which is an important component in many informational organization and management tasks. Automatic text categorization task can play an important role in a wide variety of more flexible dynamic and personalized tasks as well as real time sorting of email or files, the document management systems, search engines, digital libraries. Many classification methods have been applied to Text Categorization, for example, Naïve Bayes Probabilistic Classifiers [1], Decision tree classifiers [2], Regression methods [3], Neural Network [4], Knn classifiers [3,5] & Support Vector Machine(SVM)[6]. In many applications, dynamically mining large web repositories, the computational efficiency of these schemes is often the key element to be considered. Sebastiani pointed out in his survey on text categorization [7] Association Rule Mining can also play an important role in discovering knowledge [8]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [9] Association Rule of data mining involves picking out the unknown interdependence of the data and finding out the rules between those items [10]. Agrawal introduced

Association rules Point of Sale (POS) systems in supermarkets. A rule is defined as an implication of the form A=>B, Where A&!B'"Ø the left-hand side of the rule is called as antecedent. The right-hand side of the rule is called as consequent. [10]

This paper is organized as following. The second section gives an overview of Text Categorization with Association Rule. In the third section, introduce a new Text Categorization approach with apriori Algorithm. Experimental results are described in Section four, the research and future work are summarized the fifth section.

II. Association Rule Mining

Since the presentation of Association Rule Mining by Agrawal, Imielinski and Swami in their paper "Mining association rules between sets of items in large databases" in 1993 [11], this area remained one of the most active research areas in machine learning and knowledge discovery. Presently, Association Rule Mining is one of the most important tasks in data mining. It is considered as a strong tool for market basket analysis that aims to investigate the shopping behavior of customers in hoping to find regularities [11]. Association Mining is one of the most important data mining's functionalities and it is the most popular technique which has been studied by researchers. Extracting association rule is the hub of data mining [12]. It is mining for association rules in database of sales transactions between the items which is important field of the research in dataset. The profits of these rules are detecting unknown relationships, producing results which can perform basis for decision making and prediction [12]. The discovery of association rules is divided into two phases [13]. In the first phase, every set of items is called item set, if they occurr together greater than the minimum support threshold [14], this item set is called frequent item set. Finding regular item sets is easy but costly, so this phase is more important than second phase. In the second phase, it can generate many rules from one item set as in form, if item set {I1, I2, I3}, its rules are {I1_I2, I3}, {I2_I1, I3}, {I3_I1, I2}, {I1, I2_I3}, {I1, I3_I1}, {I2, I3_I1}, number of those rules is n2-1 where n = number of items. To validate the rule (e.g. X_Y), where X and Y are items, based on confidence threshold which determine the ratio of the transactions which contain X and Y to the transactions A% which contain X, this means that A% of the transactions which contain X also contain Y. The minimum support and the confidence are defined by the user which represents constraint of the rules. So the support and confidence thresholds should be applied for all the rules to prune the rules which values less than thresholds values. The problem that is addressed into association mining is finding the association among different items from large set of transactions efficiency [12].

III. Introduction Of An Apriori Algorithm

This is also referred as a fast Algorithm in mining frequent item sets and associations. The main objective of apriori Algorithm is to uncover hidden information that is the major goal of data mining. It was first introduced in 1993, Association Rules Mining is a very popular data mining technique and

it finds relationships among the different entities of records (for example, transaction records). Since the introduction of frequent item sets in 1993 by Agrawal et al. [15], it received a great deal of attention in the field of knowledge discovery and data mining. Association mining using apriori Algorithm is fundamentally based on the principle of

- ★ Numerous item set generation: Generate all item sets with support>= min-support
- ★ Rule generation: Generate high confidence rules from each frequent item set.

A. Classical Apriori Algorithm

Join Step: - Ck is generated by joining Lk-1 with itself.

Prune step: - Any (k-1) –itemset that is not frequent cannot be a subset of a frequent k-1 itemset

Where,

C_k: Candidate itemset of size k

L_k: Frequent itemset of size k

L_1= {frequent items};

for(k= 1; L_k != f; k++) do begin

C_{k+1}= candidates generated from L_k;

for each transaction t in database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1}= candidates in C_{k+1} with min_support

end

return $\dot{E}_k L_k$

B. Draw Backs in the Apriori Algorithm:

- ★ Over fitting
- ★ Poor comprehensibility
- ★ Both these issues arise because of the increase in the number of Association Rules generated and result in low classification accuracy.
- ★ Solution – Pruning Technique.

IV. Methodology

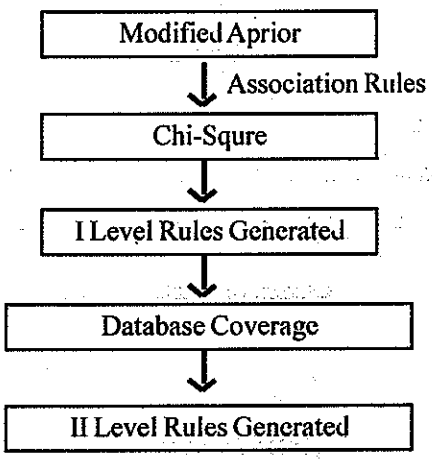

Figure1 Steps in Pruning

IV. Pruning The Set Of Association Rules

The number of rules that are be generated in the Association Rule Mining phase could be very large. There are two issues that must talk to in this case. One of them is that such a huge amount of the rules could contain noisy information which would mislead the classification process.

Another is that a huge set of the rules would make the classification time longer. This could be an important problem in the applications where fast responses are required. The pruning methods that deals in this paper are the following: eliminate the precise rules and keep only those

that are more general and with high confidence, and prune unnecessary rules by database coverage. It introduces the notions used in this subsection by the following definitions:

Algorithm Pruning the set of association rules

Input : The set of Association Rules that are found in the association rule mining phase (S) and the training text collection (D)

Output: A set of rules used in the classification process

Method:

(1) for each rule in the set S

(2) prune those that have lower confidence

(3) a new set of rules S' is generated

(4) for each rule R in the set S'

(5) go over D and find those transactions that are covered by the rule R

(6) if R classifies correctly at least one transaction

(7) select R

(10) remove those cases that were covered by R

5.1 Prediction Of Classes Associated With New Documents

The set of rules that are selected after the pruning phase, which represents the actual classifier. This categorizer can be used to predict to which class's new documents are attached.

Given a new document, the classification process searches in this set of rules for finding those classes are the closest to be attached with the document presented for categorization.

This subsection discusses the approach for labeling new documents based on the set of association rules that forms the classifier. A trivial solution would be to attach to the new document, the class that has the most rules matching this new document or the class associated with the first rule that applies to the new object. However, in the text categorization domain,

multi-class categorization is an important and challenging problem that needs to be solved. In this approach give a solution to this problem by introducing the dominance factor.

By employing this variable, allow the system to assign more than one category. The dominance factor ä is the proportion of rules of the most dominant category in the applicable rules for a document to classify. Given a document to classify, the terms in the document would yield a list of applicable rules. If the applicable rules are grouped by category in their consequent part and the groups are ordered by the sum of rules' confidences, the ordered groups would indicate the most significant categories that should

be attached to the document to be classified. This order category dominance factor is ä. The dominance factor allows to select the candidate among the categories only the most significant. When ä is set to a certain percentage, a threshold is computed as the sum of rules' confidences for the most dominate category times, the value of the dominance factor. Then, only those categories that exceed this threshold that are selected. Take K Classes(S, ä) function selects the most k significant classes in the classification algorithm.[16]

VI. Experimental Results And Performance Study

Text Corpora

In order to independently evaluate the Algorithm in respect of other approaches, like other researchers in the field of automatic text categorization, we used the Reuters-21578 text collection [17] as benchmarks. This text database is described below. Text collections for experiments are usually split into two parts: one part for training or building the classifier and a second part for testing the effectiveness of the system. There are many splits of the Reuters collection; it is chosen to use the ModApte version. The evaluation of a classifier is done using the precision and recall measures. To derive a robust measure of the effectiveness of the classifier, it is able to calculate the break event point, the 11-point precision and average precision to evaluate the classification for a threshold ranging from 0 (recall=1) up to a value where the precision value equals 1 and the recall value equals 0, incrementing the threshold with a given threshold step size. The break event point is the point where recall meets precision and the eleven values 0.0, 0.2,..0.9. "Average precision" refines the eleven point precision as it approximates the area "below" the precision / recall curve.

VII. Results

Experimental

Results Precision

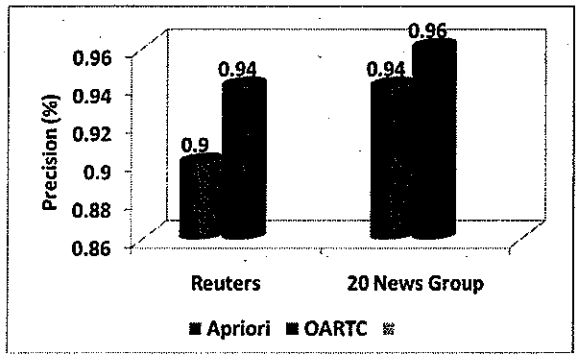

Figure 2.1

Experimental Results Recall

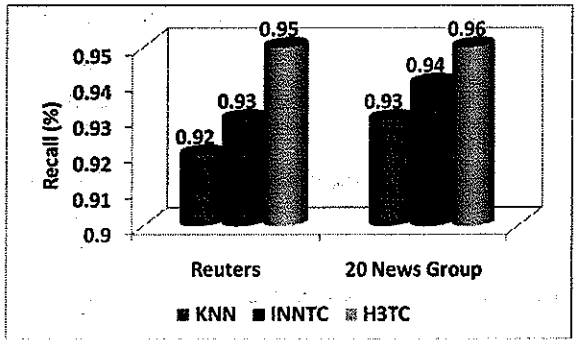

Figure 2.2

Experimental Result of the Speed

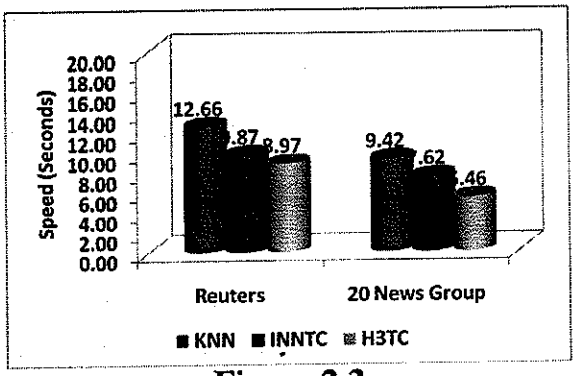

Figure 2.3

Viii. Conclusion And Future Work

★ This research paper proposed techniques to improve the performance of text document categorization. The results obtained to prove that the proposed algorithms are efficient and effective in classifying the new documents and can be applied to all online news applications.

★ In future the classification model can be built, which analyses terms on the concept sentence in document.

IX. References

1. Lewis D(1998) Naïve Bayes at forty: *"The independent assumption in information retrieval"*, In: Proceedings of ECML-98, 10th European conference on machine learning pp 4-15.

2. Cohen W, Singer Y (1999) *"Context-sensitive learning methods for text categorization"*, ACM Trans Inform Syst 17(2):141-173.

3. Yang Y, Liu X (1999) *"A re-examination of text categorization methods"*, In: Proceedings of SIGIR-99, 22nd ACM international conference on research and development in information retrieval, pp:42-49.

4. Ruiz M, Srinivasan P (1999) *"Hierarchical neural networks for text categorization"*, In Proceedings of SIGIR-99, 22nd ACM international information retrieval, pp281-282.

5. Mitchell T (1996) *"Machine learning"*, McGraw Hill, New York .

6. Joachims T (1998) *"Text categorization with support vector machines: learning with many relevant features"*, In : Proceedings of 10th European conference on machine learning, Chemnitz, germany, pp 137-142.

7. Sebastiani F (2002) *"Machine learning in automated text categorization"*, ACM Comput Surv;1-40.

8. Sotiris Kotsiantis, Dimitris Kanellopoulos, *"Association Rules Mining: A Recent Overview"*, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

9. Qiankun Zhao and Sourav S. Bhowmick, *"Association Rule Mining: A Survey"*, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003

10. R. Agrawal, T. Imielinski, and A.N. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 207-216, May 1993.

11. Agrawal, R., Amielinski, T., and Swami, A. (1993). *"Mining association rule between sets of items in large databases"*, In Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, DC, May 26-28. 12 F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, *"An Improved Algorithm for Mining Association Rules in Large Databases,"* Vol. 1, No. 7, 311-316, 2011.13.

3. R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD Record, vol. 22, pp. 207–216, 1993.

14. T. C. Corporation, "Introduction to Data Mining and Knowledge Discovery", Two Crows Corporation, Book, 1999.

15. R. Agrawal, T. Imielinski, and A.N. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 207-216, May 1993.

16. Maria Luiza Antonie , Osmar R.Zaiane " Text Dcoument Categorization by Term Association"

17. The reuters-21578 text categorization test collection http://www.research.att.com/~Lewis/reuters21578.html.

Authors Biography

K. Gayathri, pursuing Ph.D in Computer Science in Karpagam University. She is a Assistant Professor of Computer science in Nirmala College. Her research interests are in machine learning, pattern re-cognition,Feature selection, text categorization & information retrieval.

Dr. A. Marimuthu serving as Govt. Arts College (Autonomous) Coimbatore. He as presented many research papers in international Seminars & published many papers & published books in his research. He has more than two decades of Research.