ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 7 month July, year2014 Page No. 7323-7328

Multilabel Associative Text Classification Using Summarization

Prof. P.A.Bailke1, Mrs. Sonal Raghvendra Kulkarni2 ,Dr.(Prof.)S.T.Patil3

1 Computer Science Department, Vishwakarma Institute Of Information Technology, Pune
pabailke@gmail.com

2 Computer Science Department, PES Modern College Of Engineering Pune 05
Kulkarni.sonal28@gmail.com

3 Computer Science Department, Vishwakarma Institute Of Information Technology, Pune
Stpatil77@gmail.com

Abstract: ***This paper deals with the concern of curse of dimensionality in the Text Classification problem using Text Summarization. Classification and association rule mining can produce well-organized as well as precise classifiers than established techniques [1]. However, associative classification technique still suffers from the vast set of mined rules. Thus, this work brings in advantages of Automatic Text Summarization. Since text summarization is based on identifying the set of sentences that are most important for the overall understanding of document(s). These techniques use the dataset to mine rules and then filter and/or rank the discovered rules to help the user in identifying useful ones. Finally, for experimentation, the Reuter-21578 dataset are used and thus the obtained outputs have ensured that the performance of the approach has been effectively improved with regards to classification accuracy, number of derived rules and training time.***

Keywords: *text summarization, associative classification, dimensionality reduction, multi-label classification*

1.INTRODUCTION

Considering excess of electronic text information, the need for us to rapidly identify significant documents is much more vital than ever before. That's the reason the technique of automatic text classification is crucial for organizing text data.

In this paper, implemented a multi-label classification, in which a data instance can have several labels [2]. However, none of earlier works considered the effect of redundant information [7], this work also makes use of another text mining task, i.e. automatic text summarization [3].

Text summarization is a rising technique, refines the essential information from a source and final summaries will be a snippet of the original text [3]. In extractive summarization one technique is to assign weights to each sentence by considering some important characteristics of the sentence. Then all terms in the document are reweighted after summarization [4]. This work considers text sentences as basic fragments, because a sentence is usually used to express content in summarization.

Summary of the document contains the sentences, which portray all main topics of the text. Therefore, this paper also studies the applicability of document summaries instead of original texts in multi-label classification task. As the document can belong to more than few category i.e. may have more than one topic, the multi-label classification task has been preferred as a more general approach in comparison to traditional multi-class classification[6].

In this work, redundancy is reduced in the summarization

process to examine its effect on classification performance of Multilabel Associative Classifier. The developed method has been experimentally verified on Reuters-21578 dataset.

The remainder of this paper is organized as follows: Section 2 presents a proposed text summarization method and its experimental method. Section 3 is devoted to experimental investigation of paper approach, where each full text document is replaced with its summary, in multi-label classification task. Finally, in Section 4, explains about the working of MMAC (Multiclass Multilabel Associative Classifier) and in section 5, concluded the paper.

2.DIMENSINALITY REDUCTION VIA TEXT SUMMARIZATION METHOD

2.1 Document Collection

For this work, all the results of multi-label classification of full texts and their summaries are evaluated on Reuters-21578 dataset. This is one of the most popular benchmark datasets for multi-label classification. Reuters-21578 documents are presented in SGML (**S**tandard **G**eneralized **M**arkup **L**anguage) format. The TOPICS node contains one or more elements; attribute TYPE of TEXT node have value NORM. Divided the obtained dataset on training and test using values of attribute LEWISSPLIT. The value TEST of attribute LEWISSPLIT is a sign of test document, the other values of this attribute specifies the document was used for training.

An algorithm designed in such a way that, the length of the summary is defined by a percentage of initial text information amounts and the short documents were removed for which summarization does not make sense.

2.2 Summarization Process

Text Summarization is compressing the source text into a shorter version conserving its information content and meaning. It is monotonous for human beings to summarize large documents of text by hand. The difficulty of efficient data organization is an important concern in data management. By means of implementing dimensionality reduction via summarization, gist of text which gives information about the data. Therefore, this paper introduces summarization for dimensionality reduction purpose [4].

2.2.1 Key word

This summarization system takes input in " .txt " or " . doc " format. Rate of recurrence of some word in an article provides an enhanced idea of its importance. So here as the first step, words were stemmed and stop words are removed. Then each remaining word treated as a keyword.

2.2.2 Identical Cosine Similarity

Next step calculates synonymous cosine similarity which will help in removing redundant sentences. Here it is necessary to set threshold value.

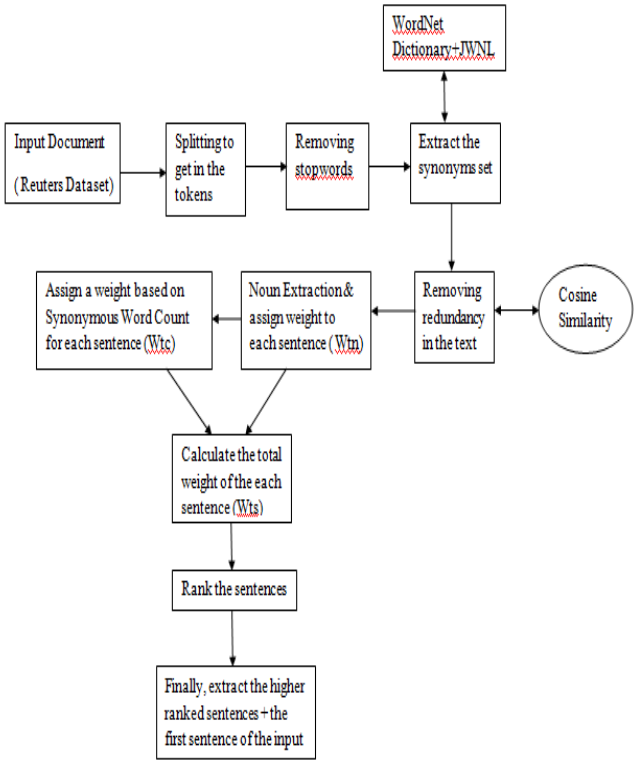

Figure1: Process Of Summarization

2.2.3 Noun Selection

Several times, in documents the word and its synonyms are replicated persistently. So these all words get higher weightage and that will be regarded as a part of the summary document. Word Net Dictionary using Java Word Net Library (JWNL) is used to get the proper noun phrase selection.

2.2.4 Location Of Sentence

Every time first sentence will be considered as a main part of a paragraph or document. Considering the same, this system also includes the first statement as it is in final summarized document. Then all sentences ranked in order of their important factor, and the top ranking sentences are chosen to form summarized document.

3. OVERALL SYSTEM DESIGN

This section brings forward experiments with replacement the full document text by its summary for multi-label classification task. In this paper a traditional vector space representation is used. Therefore, documents and their summaries both use the normalized weighting scheme tf*idf(Term frequency Inverse document frequency), for vector representation. Extraction based summarization chosen to replace the full document text by its summary as a pre-processing step for further classification [8] [9]. The length of the summary is defined by a percentage of initial text information amounts.

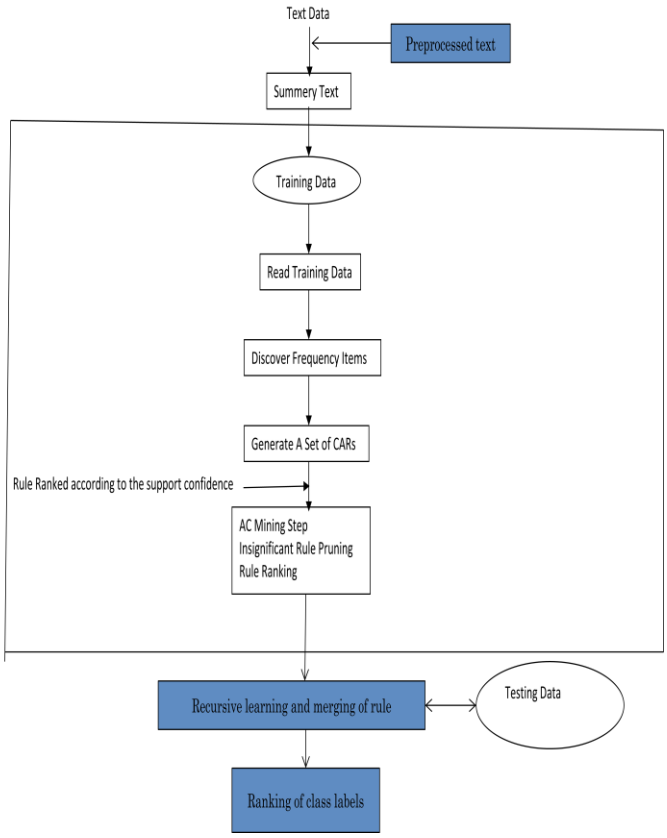

Figure2 Overall Process Of Classification

4. MMAC

This algorithm consists of three segments: rules production, recursive learning and classification. First segment, scrutinizing the training data and preparing complete CAR (Class Association Rules)[11]. Second segment, deals with the remaining unclassified instances and MMAC continues to determine more rules, till no additional frequent items can be created. In the third segment, derived set of rules merged to form a classifier which will bring into play for testing against test data. The distinguishing feature of MMAC is its ability to generate rules with multiple classes from data sets where each data object is associated with just a single class [2].

The new model consists of five main phases: the summarization, pre-processing phase, learning the rules, making of the classifier, and classifying new test data. Figure 2 represents a general description of our proposed method. It consists of all the phases of classification and Summarization also.

4.1 Pre-processing Phase

Preparation of the input data for mining is an important phase in Text Classification. Since the input text data are frequently unstructured, and might contain noise like records redundancy, imperfect transactions, absent values, etc. Therefore, the quality of the constructed models is significantly affected by the quality of the input data set.

4.1.1 Stopwords Filtering

In text documents many of words that are not helpful for the learning algorithms such as ('is', 'that', 'the'). Such words should be removed within pre-processing phase because they negatively have an effect on classifier construction. In this model, the most popular technique which is SMART stop word list has used since it is effective and has been used in many earlier works on text classification.

4.1.2 Tokenization

It is a method that comprises separating the chain characters into more meaningful Tokens. In Text Classification, the text document is divided into sentences, and words.

4.1.3 Stemming

It is the process of reducing derived words to their origin, for example, 'construction' to 'construct'. This model implemented using a popular technique which is porter stemmer.

4.1.4 TID Representation

This work adopted a data format in which input data is set as a group of columns where every column has a key identifier, it is called an item identifier (IID) and a group of transaction identifiers (TIDs) [19]. The approach uses the TID of two or more different items of the same level to find locations where they occur together and this determines whether the new item is frequent.

4.1.5 Feature Selection

The main idea of feature selection is to select a group of frequent terms that appears in the training set and make use of this collection as features in Text Classification. In this model Information Gain technique is used to transform the high dimensionality of the Reuter text collection into a numeric matrix and then used simple TID list intersections to compute the frequent items. Used Information Gain concept before rule generation in order to reduce the number of produced rules. The information gain required to be calculated for every attribute, the attribute with the highest value is considered as the best splitting attribute which will be used to produce the rules.

The weighted class association rules are generated based on the two quality measurement factors of the rule. These are: weighted support for a rule R is the ratio of the number of occurrences of R, given all occurrences of rules and weighted confidence of X-->Y, is the ratio of the number of occurrences of Y given X, among all other occurrences given X. The ruleitems that pass the predefined weight support and weighted confidence are added to the frequent weighted ruleitems set, while the others are added to infrequent.

4.2 Building the Classifier

4.2.1 Frequent Items Discovery and Rules Generation.

This method involves in scrutinizing and counting the occurrence of single items from training data, from which it decides those that pass MinSupp and MinConf thresholds, and stores them along with their occurrences (rowIds) inside fast

access data structures. Then, by intersecting the rowIds of the frequent single items discovered to this point. The rowIds for frequent single items are useful information, and can be used to locate items easily in the training data in order to obtain support and confidence values for rules involving more than one item. Once an item has been identified as a frequent item, the algorithm checks that if the item confidence is larger than MinConf, then it will be generated as a candidate rule in the classifier else discarded. Thus, all items that survive MinConf are generated as candidate rules in the classifier.

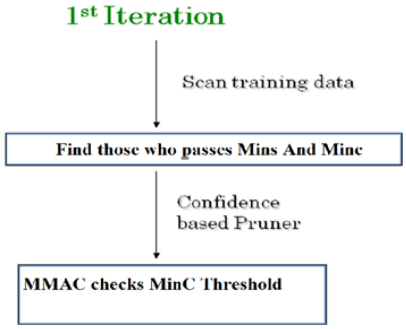

Figure3 Rule Generation

e.g. (d,c) = (The U.S. And China are the two largest consumers of energy in the world,Mr.Obama said, "Politics\","Economy\") Where (d,c) are document and class labels respectively.

4.2.2 Ranking of Rules and Pruning

After rules get generated, ranking them is important because the top sorted ranking rules take part in classifying test data. The precedence of the rules is decided by a number of measures like confidence, support and rule antecedent length[10]. For this model rule's confidence taken into consideration first, then support and lastly rule's length.

4.2.3 Rules Assessment

If any rule r is considered as a significant then it is necessary that it covers as a minimum one training instance. After rule generation and ranking process, an assessment step come into the picture to remove the redundant rule.

After checking the necessary condition of rule assessment, that if a rule correctly classifies at least a single instance, then that will be considered as a survivor, and a good candidate rule. On top of this, all instances in the approved manner classified by it will be removed from the training data. If a rule has not classified any training instance, it will then be deleted from the rules set.

4.2.4 Recursive Learning

In the process of formation of multi-label classifier, MMAC obtains more than one rule set D, for given training instances and merging of rules takes place. For D, it creates a first-pass rules set wherein each one rule is linked with the most obvious class label. After generation of the rule set, all training instances associated with it are removed and the remaining unclassified instances turn into a new training data, D'

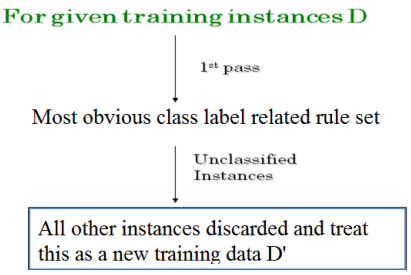

Figure4 Recursive Learning

4.2.4 Ranking of Class Labels

While the class label l1 weighing against l2, if class label l1 has a larger representation than l2 in training data then l1 precedes l2 in multilabel rule [2][5].

4.3 Classification

In classification, generated rules and the training data is taken into consideration. In this method set of high confidence rules in R(Rule set) to cover T(Training data) will be preferred. While classifying a test object, the foremost rule in the set of rules that go with the test object condition classifies it. This process guarantees that only the highest ranked rules classify test objects.

5. Experimental Results

Assessment of learning algorithm is a best criteria to judge how far the learning system's predictions from the actual class labels, tested on some unseen data. In classification task efficiency, performance and results are checked by using precision, recall and accuracy. As multilabel prediction has a supplementary details of being partially accurate, therefore calculation of the average difference between the predicted and actual labels is done independently for each test instance and averaged across the test set[12]. This approach is known as "Example Based Evaluation".

Accuracy (A): Accuracy for each instance is defined as the proportion of the predicted correct labels to the total number (predicted and actual) of labels for that instance. Overall accuracy is the average across all instances.

$$A = 1/n \sum_{i=1}^{n} \frac{|Yi \cap Zi|}{|Yi \cup Zi|}$$

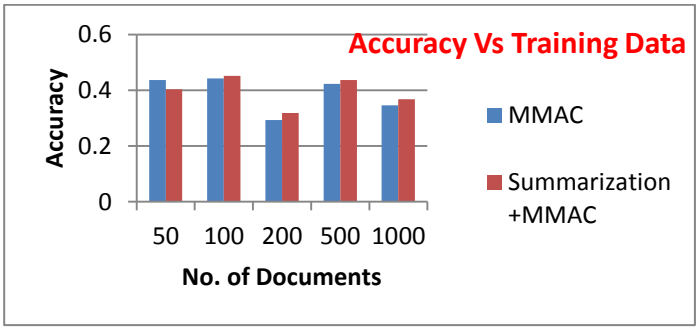

Figure5 Accuracy Results

Precision (P): Precision is the proportion of predicted correct labels to the total number of actual labels, averaged over all instances.

$$P = 1/n \sum_{i=1}^{n} \frac{|Yi \cap Zi|}{|Zi|}$$

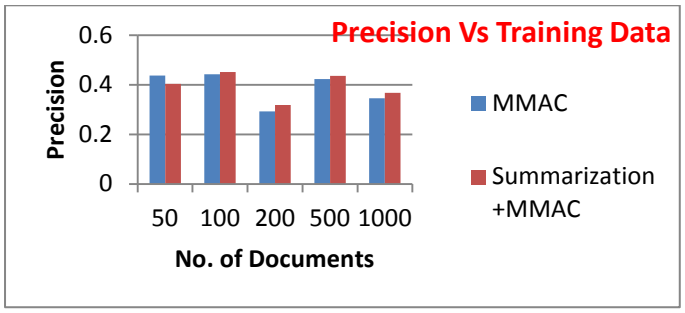

Figure6 Precision Results

Recall (R): Recall is the proportion of predicted correct labels to the total number of predicted labels, averaged over all instances.

$$R = 1/n \sum_{i=1}^{n} \frac{|Yi \cap Zi|}{|Yi|}$$

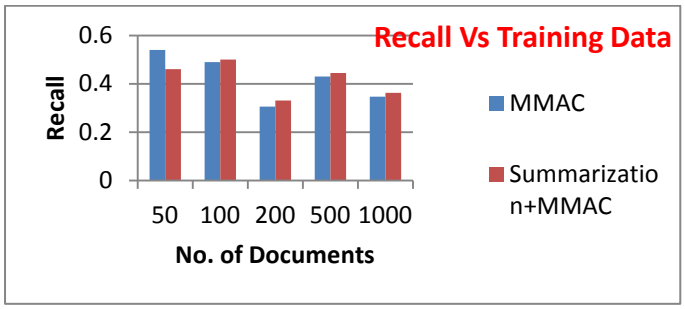

Figure7 Recall Results

Also the time needed for classification is taken into account.

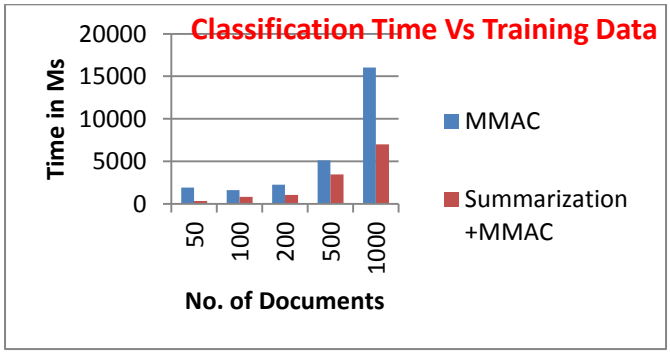

Figure8 Classifier Time Vs Training Instances

6. CONCLUSION

This paper brings in the advantages of using automatic text summarization as a dimensionality reduction technique for classifying documents. In this work, succeed to get the precise rule set and hence it increases the classification accuracy by using the summarized data set as input for the Multiclass Multilabel Associative classifier. Also found that classifier trained using summaries is much better than classifier trained using original full documents, it proves that re-weighting process of all the features in summarization can truly helpful for MMAC classification. These experiments resulted in higher accuracy, precision, and recall, but longer the execution time for summarized documents (in comparison with full documents).

References

[1] Bangaru Veera Balaji, Vedula Venkateswara Rao, "Improved Classification Based Association Rule Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013

[2] Fadi Abdeljaber Thabtah, Peter Cowling, Yonghong Peng,"Multiple labels associative classification", Knowl Inf Syst (2006)

[3] A.Anil Kumar, S. Chandrasekhar, "Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 5, July - 2012

[4] Pranitha Reddya, R C Balabantarayb, "Improvisation of the Document Summarization by Combining The IR Techniques with ' code-Quantity', Memory and Attention' Linguistic Principles", 2nd International Conference on Communication, Computing & Security [ICCCS-2012], ScienceDirect.

[5] Francisco Charte, Antonio Rivera, Maria Jose del Jesus, "Improving Multi-label Classifiers via Label Reduction with Association Rules", Springer-Verlag Berlin Heidelberg, 2012

[6] Tsoumakas, G., Katakis, I., Vlahavas, "Mining Multi-label Data. In: Data Mining and Knowledge Discovery" Handbook, pp. 667–685 (2010)

[7] Iris Hendrickx, Walter Daelemans, Erwin Marsi, Emiel Krahmer, "Reducing redundancy in multi-document summarization using lexical semantic similarity", ACL-IJCNLP ,2009

[8] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[9] Karel Jezek, Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti, 2008, pp.1- 12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[10] Houtao Deng, George Runger, Eugene Tuv, Wade Bannister," CBC: An Associative Classifier with a Small Number of Rules", Decision Support Systems, In Press, Accepted Manuscript, Available online 4 December 2013

[11] S.Kannan, R.Bhaskaran,"Role of Interestingness Measures in CAR Rule Ordering for Associative Classifier: An Empirical Approach", JOURNAL OF COMPUTING, VOLUME 2, ISSUE 1, JANUARY 2010

[12] Mohammad S Sorower, " A Literature Survey on Algorithms for Multi-label Learning", Corvallis, OR, Oregon State University. December 2010

Author Profile

Prof.P.A.Bailke received the B.E. and M.E. degrees in Computer Science Engineering from Pune Institute of Computer Technology and Bharti Vidyapith, Pune respectively in 2002 and 2006, respectively. She is working with Vishwakarma Institute of Technology from last 9 years as a Asst. Professor. Pursuing PhD in Computer Science. She is working on AICTE research project,grant of Rs 6.5 lakhs from 2012 onwards.