ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

Study of Effect of Modified Feature Selection Method and Re-Occurrences of Features on Performance of Multi-Label Associative Classifier

Prof. Mrs. P. A. Bailke[1], Ms. Shweta Kambare[2], Prof. Dr. S. T. Patil[3]

Professor, Computer Department, Vishwakarma Institute of Technology, Pune, India[1,3]

M. Tech Student, Computer Department, Vishwakarma Institute of Technology, Pune, India[2]

Abstract: In this paper, the effect of modified feature selection method and re-occurrences of features on performance of multi-label associative classifier is studied. In the proposed approach, important words (keywords) from each document in the training dataset are selected by using two methods. One is to select words having mutual Information (MI) value greater than given threshold as keywords and second is to select a limited number of words from a document having maximum values of MI as keywords. The method to select keywords is decided by comparing the maximum MI value of word from a document with limit value. If the maximum MI value is greater than limit value then first method is used for keyword selection, otherwise second method is used for keyword selection. This method ensures that keywords are selected from each and every document and unnecessary keywords are avoided. Association rules are generated by using the extracted keywords. Re-occurrences of features i.e. keywords are considered while calculation of supports of rules. In the proposed approach, multiple minimum support threshold method is used for rule pruning to handle the rare class problem. The classifier assigns multiple labels for a single document. If no label is found for a document from the generated rules, then the class label with highest support in the dataset is assigned to the document. The classifier built by using the proposed approach provides good accuracy as compared to traditional associative classifiers.

Keywords: Multi-label associative classifier, text classification, multiple minimum support thresholds, mutual information, association rules.

I. INTRODUCTION

About 80% of data in organizations is in unstructured format. As the use of internet is increasing, it is necessary to classify the text documents and organize the information. Building fast and accurate classifiers for text documents is an important task in text mining. An associative classifier is robust and can deal with impure dataset and it also takes less time than the traditional techniques as it has to only check the rule schema and not the whole dataset. So, Integrating classification and association rule mining can produce more efficient and accurate classifiers than traditional techniques.

In this approach, classification of documents is done by using multi-label associative classifier and reoccurrences of terms is also considered. The classifier built consists of rules with keywords as antecedent (LHS of the rule) and classes as consequent (RHS). The documents are represented with keywords. If keywords of a document to be classified don't match any antecedent of the rules then the class with highest support is given as default class to the document. The documents are classified into previously defined classes based on the rules generated from training dataset. Using this accuracy of the classifier can be increased.

II. RELATED WORK

A new approach for multi-class multi-label classification rules has been proposed in [1]. This approach has following distinguishing features:

(1) It produces classifiers that contain rules with multiple labels.
(2) It presents four evaluation measures for determining accuracy that are applicable to a wide range of applications.
(3) It employs an efficient method for discovering rules that requires only one scan over the training data.
(4) It employs a detailed ranking method, which prunes redundant rules, and ensures only effective ones are used for classification.

This approach doesn't provide an effective classifier for text documents. The rules selected can be uninterested for the user. This approach doesn't provide good accuracy. In the proposed approach, the rules will be applied based on similarity between the keywords and the antecedents of the rules. The documents will be classified in different classes based on the similarity with the antecedent of the rule. So the proposed approach provides more accuracy.
In [2], development of a novel system for automated classification of MEDLINE article references is described.

DOI 10.17148/IJARCCE.2015.4795

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

Here, associative classifier with reoccurring items (ACRI), to assign MeSH keywords to article references is used. This method is capable of performing challenging multi-label classification. It is a novel system for automated classification of MEDLINE documents to MeSH keywords based on ACRI, which was modified to accommodate multi-label classification. This approach doesn't consider semantics of the terms. In the proposed approach, the similarity between keywords and the antecedents of the rules is estimated using Knowledge base. Thus the proposed technique can improve the accuracy of the classifier.

III.PROPOSED APPROACH

A. Building Multi-label Associative Classifier

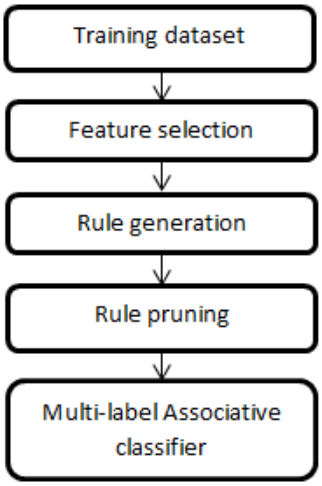

Fig.1.Building multi-label associative classifier

The First step in the classification is building classifier. RCV1-v2 dataset is used for building the classifier. Fig. 4.1 shows the steps involved in building classifier. The steps are explained below.

1) Training dataset:
RCV1-v2 dataset is used for training. The dataset is in tokenized form. So there is no need for pre-processing.

2) Feature selection:
In this step, the keywords of documents are extracted. For choosing keywords, feature selection method used is Mutual Information.

3) Mutual Information:
MI measures how much information the presence/absence of a term contributes to making correct classification decision on c. The MI is given by

$$I(u;c) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(u = e_t, c = e_c) \log_2 \frac{P(u = e_t, c = e_c)}{P(u = e_t)P(c = e_c)}$$

where u is a random variable that takes values e_t=1 (the document contains term t) and e_t = 0 (the document does not contain term t), and C is a random variable that takes values e_c = 1(the document is in class c) and e_c = 0 (the document is not in class c.

$$I(u;c) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_1 N_1} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_0 N_1} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_1 N_0} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_0 N_0}$$

where the Ns are counts of documents that have the values of e_t and e_c that are indicated by the two subscripts. For example, N_{10} is the number of documents that contain term t (e_t =1) and are not in class c (e_c = 0). $N_1 = N_{10} + N_{11}$ is the number of documents that contain t (e_t =1) and documents independent of class membership ($e_c \in \{1,0\}$). $N = N_{10} + N_{11} + N_{01} + N_{00}$ is the total number of documents.

MI reaches maximum value if the term is perfect indicator for a class. The keywords of a document are chosen by two methods. The method to be used is decided by limit value. If Maximum MI value of a term in a document is greater than limit value, then terms satisfying threshold are chosen as keywords for the document. Otherwise the terms with descending order of MI values are chosen as keywords for the document. In this case the number of keywords will be limited. These keywords are used to represent document in the further process.

4) Rule generation:
Rules are generated from the feature selection matrix. LHS of rules that is antecedents of rules are keywords of the documents and the RHS of the rules are class labels. Support and confidence of each rule is calculated. Re-occurrences of terms are considered while calculation of support.

The frequent rule-items are chosen to generate rules. Frequent rule-items generation is done as follows. A rule-item gives a rule. It is in the form $\{o_{11}k_1, c_1\}$. This gives rule $o_{11}k_1 \rightarrow c_1$

1. In the first step, the candidate 1 rule-items C_1 is generated. For candidate 1 rule-items, the term and class combinations chosen are depending upon the number of occurrences of the term and the classes of the documents in which the term is present.
2. The rule-items satisfying minimum support threshold are saved in frequent 1 rule-item L_1.
3. The minimum support threshold for each class is calculated by method discussed in 4.1.5.
4. In the next step $C_2 = L_1$ join L_1 is generated. The rule-items satisfying minimum support threshold are saved in frequent 2 rule-items L_2.
5. In the candidate 3 rule-items generation, Apriori property is used. The 2 rule-items which have infrequent subsets are pruned from the candidate set.
6. This process continues till the candidate set is empty.

The rule-items give rule schema. In the next step the rule pruning is done to reduce number of rules.

5) Multiple Minimum Support threshold method:
Multiple minimum support method is used for calculation of minimum support threshold for different classes. The minimum Support (MS) for a class c_i is calculated as described below.

For every class c_i, the MS (c_i) is calculated as follows

DOI 10.17148/IJARCCE.2015.4795

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

$$MS(c_i) = \beta S(c_i); \text{ if } \beta S(c_i) > LS = LS \text{ else}$$

where, β is a user-specified proportional value which can be varied between 0 to 1,$S(c_i)$ refers to support of an item equal to $f(c_i)/N$, ($f(c_i)$ represents frequency of c_i and Nis the number of transactions in a transaction dataset) and LS corresponds to user-specified least support value.

6) Rule Pruning:

Confidence of each rule is calculated by following formula.

$$\text{Conf}(o_{11}k_1 \rightarrow c_1) = \frac{\text{support count of rule}}{\text{support count of LHS of rule}} * 100$$

The rules which satisfy minimum confidence threshold satisfied by user are included in the classifier. Remaining is pruned.

7) Multi-label associative classifier:

After rule pruning, the remaining rules are included in the classifier.

B. Multi-label associative classification of text documents

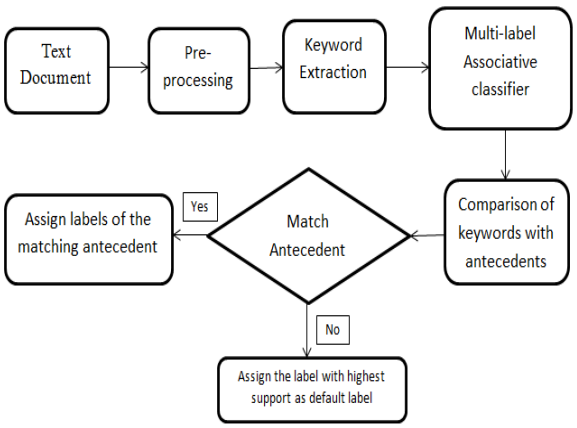

Fig.2. Classification of text documents

Text document is given as input to the classifier. Before classification, keyword extraction from the document is done. The document is represented by keywords.

1) Keyword extraction:

For keyword extraction, the association between term and given classes is considered. The terms having higher association with the given classes are extracted as keywords.

2) Classification of Document:

For classification of the document, the keywords representing the document are compared with antecedents of rules in the classifier. The document is assigned the label based on percentage of coverage of antecedent. It is calculated as below

$$\% \text{ coverage} = \frac{\text{number of common words}}{\text{Total number of words in the antecedent}} * 100$$

3. If the keywords of the document don't match with any of the antecedents of the rules then the label with highest support is assigned to the document.

IV. EXPERIMENTAL RESULTS

The algorithm is tested on machine with Intel(R) core(TM) i3-2328M CPU @ 2.20 GHz with installed RAM of 8GB on 1000 documents as training documents and 500 documents as testing documents from RCV1-V2 dataset. A multi-label associative classifier is built without considering re-occurrences of features and using traditional MI based feature selection method (traditional approach) for comparison purpose. In the traditional MI feature selection method, words having MI value greater than given MI threshold are considered as features. The results of the classifier built by using the traditional method and the proposed approach are shown in the following tables.

Table I Accuracy of classifier built without considering re-occurrences of features and using traditional mi feature selection method

Least support value used for calculation of multiple minimum support thresholds (%)	Mutual Information threshold value	Accuracy (%)	Precision (%)	Recall (%)	Number Of Rules in the classifier
5	0.1	61.96	88.71	56.49	7737
5	0.3	61.96	88.71	56.49	7609
5	0.5	61.59	86.70	53.39	7552
5	0.7	60.14	82.75	48.33	7305
5	1.0	58.87	92.90	57.61	7196
5	2.0	56.96	85.86	47.85	9534
10	1	49.61	86.43	55.13	6018
10	0.7	50.34	79.29	47.75	7694
10	0.5	51.25	75.36	58.88	5027
10	0.3	51.61	69.64	60.63	5079
10	0.1	51.61	75.48	50.63	8586

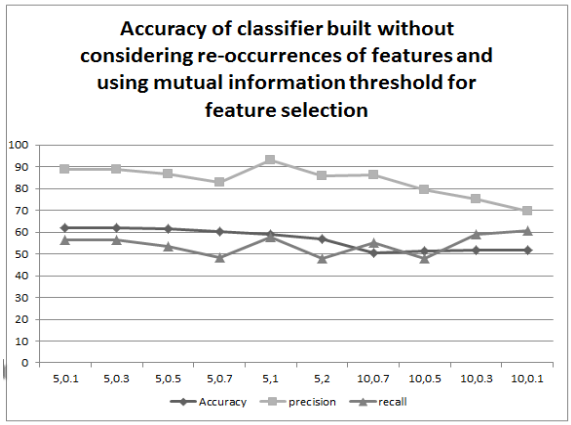

Fig.3.Accuracy of classifier built without considering re-occurrences of features and using mutual information threshold for feature selection

TABLE II Accuracy Of Classifier Built By Using Proposed Approach

Least support value used for calculation of multiple minimum support thresholds (%)	Mutual Information threshold value	Accuracy (%)	Precision (%)	Recall (%)	Number Of Rules
25	0.3	68.49999	79.16	55.75	5479
25	0.7	60.72	63.02	57.78	5121
25	1.0	61.5384	73.79	48.25	5130
25	2.0	73.68	73.79	49.46	5167
10	0.3	65.47	87.65	57.50	14288
10	0.5	64.68	89.75	55.67	10199
10	0.7	59.91	93.03	55.47	7961
10	1.0	70.647	86.13	47.22	7998
10	2.0	81.30	97.14	57.41	6886
5	2.0	82.77	71.17	55.97	14187
5	1.0	77.021	95.18	45.54	22016

DOI 10.17148/IJARCCE.2015.4795

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

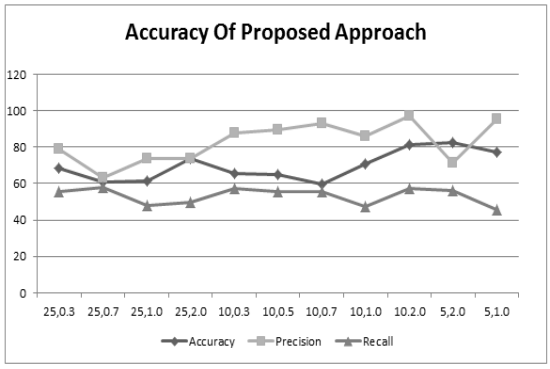

Fig.4.Accuracy of classifier built by using proposed approach

TABLE III Comparison between Two Approaches

Least support value used for calculation of multiple minimum support thresholds (%)	Mutual Information threshold value	Accuracy of classifier built without considering re-occurrences of features and using mutual information threshold for feature selection (%)	Accuracy Of classifier built by using proposed approach (%)
10	1	49.61	70.647
10	0.7	50.34	59.91
10	0.5	51.25	64.68
10	0.3	51.61	65.47
5	2	56.96	82.77
5	1	58.87	77.021

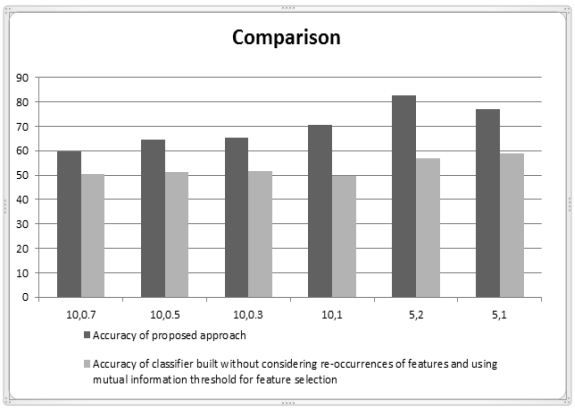

Fig.5. Comparison between the two approaches

TABLE IV 5 Fold Cross Validation for Proposed Approach

Cross fold	Accuracy	Precision	Recall
1	71.67	88.16	56.29
2	60.61	78.57	52.95
3	59.91	77.71	64.83
4	70.67	66	51.79
5	71.79	91.58	47.001

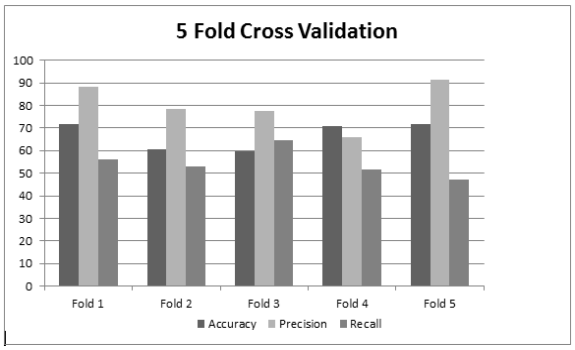

Fig.6.5 Fold cross validation for the proposed approach

V. CONCLUSION

Fig.5 shows the accuracy of two classifiers. In case of traditional classifier, if the MI threshold is set very low, unnecessary features may get selected. If the MI threshold is set high then keywords from some documents might not get selected because they have maximum MI value of word less than the given threshold. Thus the accuracy of classifier increases if the threshold is set very low and decreases if the mutual information threshold increases.

In the proposed method, the features are selected by using two methods. So unnecessary keyword selection is avoided and keywords are selected from each and every document. The number of features selected is not completely dependent on MI threshold set for the mutual information value. Thus accuracy of the classifier built by using proposed approach is not solely dependent on the given MI threshold.

In the proposed approach, re-occurrences of features are considered while calculation of support. The accuracy of the classifier is less sensitive to the support threshold unlike traditional associative classifier. Multiple minimum support method is used for rule pruning. Minimum support threshold for each class is calculated based on its support in the training dataset. The class having less support in the training dataset is assigned less minimum support threshold. Thus the rules having the rare class also get included in the classifier and the rare class problem is solved.

Experimental results show that the proposed approach provides good accuracy compared to other associative classifier.

REFERENCES

[1] Fadi Abdeljaber Thabtah · Peter Cowling·Yonghong Peng "Multiple labels associative classification" Knowl Inf Syst (2006) 9(1): 109–129
[2] Rafal Rak, Lukasz A. Kurgan, and Marek Reformat. "Multilabel Associative Classification Categorization of MEDLINE Articles into MeSH Keywords" IEEE Engineering in Medicine and Biology Magzine. 2007
[3] Rafal Rak, Wojciech Stach, Osmar R. Zaıane, and Maria-Luiza Antonie "Considering Re-occurring Features in Associative Classifiers" Springer-Verlag Berlin Heidelberg 2005
[4] Bing Liu, Wynne Hsu and Yiming Ma "Mining Association Rules with Multiple Minimum Supports" KDD-99 San Diego CA USA ACM 1999.
[5] Bing Liu Wynne Hsu Yiming Ma "Integrating Classification and Association Rule Mining" KDD-98 Proceedings 1998.
[6] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules," in Proc. IEEE Int. Conf. Data Mining, San Jose, CA, 2001, pp. 369–376.
[7] X. Yin and J. Han, "CPAR: Classification based on predictive association rules," in Proc. 3rd SIAM Int. Conf. Data Mining (SDM'03), San Francisco, CA, 2003, pp. 369-376.
[8] O.R. Zaïane and M.-L. Antonie, "Classifying text documents by associating terms with text categories," in Proc. 13th Australasian Database Conf., Melbourne, Australia, 2002, pp. 215–222.