Lakshmi Kumari / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 2, March -April 2013, pp.928-931

Improved Graph Based K-NN Text Classification

Lakshmi Kumari
*(Department of Computer Science, HMR INSTITUTE OF TECHNOLOGY AND MANAGMENT, IP, India)

ABSTRACT

This paper presents an improved graph based k-nn algorithm for text classification. Most of the organization are facing problem of large amount of unorganized data. Most of the existing text classification techniques are based on vector space model which ignores the structural information of the document which is the word order or the co-occurrences of the terms or words. In this paper we have used the graph based representation of the text in which structural information of the text document is taken into consideration. Feature selection phase plays a very important role in classification. The emphasis has been on effective feature selection methods using both standard as well globalized methods of feature selection which are MI+Chi, RMI+Chi (standard method) and WT (localized method). The dataset that had been used is self made English text document of five different categories. The final result had shown that it is not always that a standard method of feature selection will improve the categorization but a localized method that is Weight Of Terms [WT] can also improve the classification.

Keywords – **Feature Selection, K-NN, Text Classification, Vector Space Model**

I. INTRODUCTION

With the development of internet, a large amount of data in any organization needs efficient classification. The traditional text classification technique have been based on vector space model which ignored the structural information of the document that is word order and co-occurrence of the words in the document. Therefore the paper has used graph based technique which takes into account the structural information of the document. Feature selection phase always play a major role in classifying the document as it removes the redundant, irrelevant and noisy words from the documents. The feature selection methods that the paper has used are RMI+Chi, MI+Chi and WT method.

The Graph based technique has been recently developed by Schenker A. and Zhaotao et al.

Morework on graph based technique has been done by Wei Jin , Rohini k and Maual and Alxander.

The paper had used graph based model to capture the structural information of the document by using a centroid matrix representation of the document [1].

The algorithm which the paper has used is K-nn text classification algorithm for classifying the document.

II. FEATURE SELECTION AND PREPROCESSING

The text classification begins with preprocessing which includes the stop word removal and stemming of the document. The stemming algorithm that the paper has used is porter stemmer. After preprocessing the next phase is feature selection. The feature selection that the paper has used are MI[2] with Chi, RMI[3] with Chi [4] and WT .

RMI

Rgularized mutual information measures the relevance of a term in a category. It is effective than mutual information and do not takes into account the numerical values.

RMI = 2MI(t,c) / H[t] +H[c]

Weight Of Terms

It is formed by replacing the IDF [inverse document frequency] in TF-IDF. It is used to measure the weight of terms appearing frequently as well as rarely in the document.

WT = TF (t).MI (t,c)

MI

It is used to measure the mutual dependence of the two terms in a paragraph or in whole document.

$I(t,c) = \log P(t,c) / P(t/c) * P(t)$

$I(t,c) = \log P(t/c) - \log P(t)$

Where P(t,c) is the probability of the term t in the category c, P(t) is the probability of the term t and P(c) is the probability of the category c.

Chi Square Statistics

It is used to measure the lack of independence between the term t and the category c.

$CHI(w,c) = N*(P(w,c)* P(\overline{w},\overline{c}) - P(w,\overline{c})*P(\overline{w},c))/P(w)* P(\overline{w})* P(c) * P(\overline{c})$

P(w) is the probability of w in the document d and P(c) is the probability when the text belong to category c. $P(\overline{w}, c)$ is the probability that word do not occur in the category , $P(\overline{w}, \overline{c})$ is the probability that word w and category do not appear .similarly the meaning of rest of the terms can be known.

Combined Feature Selection Method

In this paper a combined method of feature selection [5] has used.The features whose frequency is very low in a document may play a vital role in classifying a document and also the feature which is occurring

many a times may not be efficient in determining a document. So the paper the proposed the combination of various feature selection method so to find the words or the features which are more efficient in determining a document. So the paper has used first MI+Chi then RMI+Chi and finally a localized method WT.

II.GRAPH BASED K-NN TEXT CLASSIFICATION

After the feature selection step the whole text converted into graph based on the features selected.

2.1. Graph Based Text Representation Model [6]

A graph is 3 tuple G=(V,E,F,W,M), where V is a set of nodes , E is a collection of weighted edges connecting nodes. FWM (Feature Weight Matrix) [7] is defined as the feature weight matrix of the edges.

- Node:

Unique feature terms obtained from the train set using feature selection methods.

- Edges:

Constructed based on order and co-occurences relationship between feature words.

- Feature Weight Matrix:

The FWM is defined as a diagonal matrix to represent the structural information of the text with the help of centroid feature vector W shown below :

$W=(W_{11}, W_{22}\ldots W_{ii}\ldots W_{nn})$

We assign the frequency of feature terms f_i which appears in a text to W_{ii} defined as the i-th diagonal weight of the matrix. This matrix form shows more feature of text such as frequency, order and co-occurrence of the terms in the text.

III. IMPROVED KNN CLASSIFICATION BASED ON GRAPH

The method for classification had used KNN as it is the simplest method with great precision capability. Although KNN is an example based algorithm therefore the classification speed of KNN is slow but owing to its great precision the paper had used KNN algorithm.

As a graph consists of nodes ,edges and the weight of the edges, we can define the similarity measure of two graphs by those elements. Three different algorithms [7] have been used for classification. First has been used to convert the text into graph and then two improved classification measures have been used in calculating the similarity between two graphs and finally classifying the document.

Input :

Training set $D=\{d_1,d_2\ldots d_i\ldots d_n\}$ D_i is a text after segment and stop words filtering $D_i=\{f_1,f_2,..f_i..f_m\}$, f_i is the i-th word of text f_i = Feature selected, N_i =Node w_i = Weight

Output :

Training set $G=\{ g_1,g_2,\ldots g_i, \ldots g_n \}$ g_i is the i-th text represented by graph

Procedure:

1. For each d_i in D
2. Initialize the node set N_i , edge set E_i and Feature Weight Matrix FWM_i to be empty.
3. For each f_i in d_i
4. If($f_i \in N_i$)
5. create a new node n_i representing f_i,
6. add n_i to N_i,set $w_{ii}=1$ // w_{ii} is defined in (3)
7. End If
8. End for
9. For each f_i in d_i
10. Create a new edge e_i connecting f_i and f_{i+1}
11. find the node n_k which representing f_i
12. If($e_i \in E$)
13. add e_i to Ei, set weight e_i =1
14 w_{kk++} ; // w_{kk} represents the frequency of n_k
15. Else If($e_i \in E_i$)
16. weight e_i ++;
17. w_{kk++}
18. End If
19. End For

Algorithm1. Text to graph conversion

FW(feature weight): It describes the similarity between two graphs by weight of both nodes and edges appear in both two graphs. It can be calculated as follows.

Testing set graphs $G=\{g_1,g_2,\ldots g_i, \ldots g_n\}$

Training set graphs $CG=\{cg_1,cg_2,\ldots cg_i, \ldots cg_n\}$

w_{ij} = weight of the edge

Fw= Feature Weight

Procedure:

1. For each edge in g_i
2. If edge in cg_i
3. If($w_{ij}(g_i)>= w_{ij}(cg_i)$ // w_{ij} is the weight of edge
4. If(j>i)
5. $Fw+=\alpha\ w_{ij}(cg_i)$
6. Else if(j=i)
7. $Fw+= w_{ij}(cg_i)$
8. End if
9. Else If($w_{ij}(g_i)< w_{ij}(cg_i)$
10. If(j>i)
11. $Fw+=\alpha\ w_{ij}(g_i)$
12. Else if(j=i)
13. $Fw+= w_{ij}(g_i)$
14. End if
15. End if
16. End if
17. End for

Algorithm 2. Calculation of feature weight

The following algorithm has been used for the final classification of the document into its category.
Input:
Testing set graphs $G=\{g_1,g_2,\ldots g_i, \ldots g_n\}$, value k =5
Training set graphs $CG=\{cg_1,cg_2,\ldots cg_i, \ldots cg_n\}$
Nfp= Node Fit Percent
Efp= Edge Fit Percent
F_w = Featured Weight
Output :
Result set $R=\{r_1,r_2,\ldots r_i, \ldots r_n\}$

Procedure:
1 For each g_i in G
2. Initial List RL to store F_w and text

category（length is K）
3 For each cg_i in CG
4. If $Nfp(g_i,cg_i)>\alpha$ && $Efp(g_i,cg_i)>\alpha$
5. Calculate Feature weight $F_w(g_i,cg_i)$
6 If RL is not full
7 Add $F_w(g_i,cg_i)$ and category of cg_i to RL
8 Else If RL is full
9 If $F_w(g_i,cg_i)$ >min (F_{wi} in RL)
10 Replace F_{wi} in RL with $F_w(g_i,cg_i)$
11 End if
12 End if
13 End if
14 . End For
15 the category of g_i is the category appears most in RL
16 add the category of g_i to the Result Set R.
17 End For
Algorithm 3. Classification of document

IV. EXPERIMENT

We have taken self made corpus made of five different categories namely business, health, sports education and science. The different features of the testing and training have been chosen whose statistics have been given below in the table.
TABLE 1. Corpus Statistics

Category	Testing Set [words selected]	Training Set
Business	369	515
Health	341	495
Sports	329	540
Education	266	602
Science	375	478

We have used the F1 method to measure the performance of the classification methods.

F1 = 2*Precision * Recall / Precision + Recall

We have taken the value of k is taken to be 5. The details of the result are shown below.
TABLE I. The Categorization Result Using

Category	Recall %	Precision %	F1%
Business	85.71	100	92.30
Health	87.5	87.5	87.5
Education	80	100	88.88
Science	83.88	83.33	83.33
Sports	100	83.33	90.90
Average	87.30	90.83	88.58

Improved GKNN Algorithm [RMI+CHi]

Category	Recall %	Precision %	F1%
Business	66.66	80.00	72.72
Health	62.50	83.30	71.42
Education	60	100	75
Science	71.42	100	83.32
Sports	80	80	80
Average	68.11	88.66	76.49

TABLE II. The Categorization Result Using Improved GKNN Algorithm [MI+CHi]

Category	Recall %	Precision %	F1%
Business	83.33	83.33	83.33
Health	75	85	79.6
Education	75	75	75
Science	85.71	100	92.30
Sports	83.33	100	90.90
Average	80.47	88.66	84.22

TABLE III. The Categorization Result Using improved GKNN Algorithm [WT]

The categorization results have shown that the classification accuracy is improved. The graphs clearly shows that the use of WT as feature selection method has improved the classification accuracy of the algorithm. The RMI+CHI has also improved the accuracy but not better than WT method. It is already proved that using single mutual information has inferior performance compared to other methods due to a bias favouring rare terms and strong sensitivity to probability estimation errors. Although use of mutual information with TF-IDF is just an ad hoc approach to improve the efficiency but the results have shown that it is a reliable measure for selecting informative features or words. It can be used instead of any other globalised method of feature selection.
Th The main observations that we had obtained after analysis of the re results are as follows
1. 1.CHI is normalized and scores obtained are comparable across the same category.

2. 2.Using WT had boosted the performance with the fact that rarely occurring words are effective in classifying a document.
3. 3.Combining good methods with little or no correlation improved the results of classification.
4. 4.It is not always the globalised methods of feature selection which will improve the performance, localised methods of feature selection can also improve the performance

V. CONCLUSION

However, the graph based K-NN had improved the categorization to some extend and had also given a compact text representation technique, still there is a need to focus on better text representation methods. The project had worked with only K-NN classifier [8], which is considered to be best classifier, the graph based need to be combined other standard classifiers like SVM and Naive Bayesian. This thesis discussed the following three points related to text classification using machine learning.

1. How to perform a highly precise classification by using a large number of word attributes,
2. How to achieve a highly precise and efficient classification by assuming the existence of sub-categories and using graph based text representation,
3. How to utilize the different feature selection methods effectively to improve the classification accuracy.

With the increasing amount of existing online data as well as the documents in an organization there is a need of an effective text categorization technique. Here in this technique of text categorization the project had concentrated on the compact representation of the document. The feature selection phase plays a vital role in improving the text classification precision because it helps in finding out the relevancy of a particular document in its training categories. The project has used two different methods of feature selection, one is globalised and other one is localised.

The experimental results have clearly shown that it is not always the globalised method which improves the categorization accuracy, the localised method which we introduced WT had significantly improved the classification accuracy.

However, the graph based K-NN had improved the categorization to some extend and had also given a compact text representation technique, still more work need to be done on better text representation methods. The project had worked with only K-NN classifier, which is considered to be best classifier, the graph based need to be combined other standard classifiers like SVM and Naive Bayesian.

REFERENCES

[1] Zonghu Wang , Zhijing Liu ,2010 .Graph-based Chinese Text Categorization. *In Proc. Of Seventh International Conference On Fuzzy systems and Knowledge Discovery*. pp 2363-2366.

[2] Xiang Zhang, Mingquan Zhou, Guohua Geng, Na Ye, 2009. A Combined Feature Selection Method for Chinese Text Categorization. *In Proc. Of International Conference Information Engineering and Computer Science*, Wuhan, pp. 1-4.

[3] Xiang Zhang, Mingquan Zhou, Guohua Geng, Na Ye, 2009. A Combined Feature Selection Method for Chinese Text Categorization. *In Proc. Of International Conference Information Engineering and Computer Science*, Wuhan, pp. 1-4.

[4] Yao- Tsung, Chen, Meng Chang Chen, 2011. Using Chi-Square Statistics To Measure Similarities For Text Categorization. Taiwan. pp – 3085-3090. 14.

[5] Zonghu Wang, Zhijing Liu, 2010. Graph Based K-NN Text Classification**.** *In Proc. Of International Conference on Electrical and control engineering.* pp 1092-1095.

[6] Zhou, Fan Zhang, Bingru Yang ,2005. Towards Graph-based Text Representation Model and Its Realization. *In Proc. Of International Conference on Natural Language and Knowledge Engineering , CNLP-KE*, Beijing ,vol.19, pp. 1-8 .

[7] Zonghu Wang , Zhijing Liu ,2010 .Graph-based Chinese Text Categorization. *In Proc. Of Seventh International Conference On Fuzzy systems and Knowledge Discovery*. pp 2363-2366.

[8] Fabrizio Sebastiani, 2002. Text Categorisation. ACM Computing Survey, Italy, pp. 1-19.