ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

IJIRSET

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

PMI Based Clustering Algorithm for Feature Reduction in Text Classification

P.Jeyadurga, Prof. P. R. Vijaya Lakshmi, J.S.Kanchana

Department of CSE, KLN College of Engineering, Sivagangai, Tamil Nadu, India.

Department of CSE, KLN College of Engineering, Sivagangai, Tamil Nadu, India.

Department of IT, KLN College of Engineering, Sivagangai, Tamil Nadu, India.

Abstract— Feature clustering is a feature reduction method that reduces the dimensionality of feature vectors for text classification. In this paper an incremental feature clustering approach is proposed that uses Semantic similarity to cluster the features. Pointwise Mutual Information (PMI) is widely used word similarity measure, which finds Semantic similarity between two words and is an alternative for distributional similarity. PMI computation requires simple statistics about two words for similarity measure, that is number of co-occurrences or correlations between two concepts of fixed size are computed. Once the words from preprocessed documents are fed, clusters are formed and one feature (head word) is identified for each cluster which are used for indexing the document. PMI assumes that a word have single sense, but clustering can be optimized further if polysemies of words are considered. Hence PMI may be combined with PMI$_{max}$, which estimates correlation between the closest senses of two words also, thereby better feature reduction and execution time compared with other approaches.

Keywords— Feature reduction, feature clustering, Semantic similarity, Pointwise Mutual Information (PMI).

I. INTRODUCTION

Text classification is the problem of estimating true class label for a new document. High dimensionality of feature vectors of the document can be an obstacle for classification algorithms. Therefore feature reduction approach is applied before text classification algorithm is applied to document. Feature extraction and Feature selection are earlier approaches that were developed for feature reduction of which feature extraction worked better. But all these feature reduction algorithms had higher computational complexity. Feature clustering is an effective approach for reducing the dimensionality of the document. The basic idea is to group the features into clusters that are highly related and a single feature is extracted from each cluster thereby reducing number of features.

Initially Backer and McCallum [3] proposed a feature reduction technique based on clustering, that uses distributional similarity to cluster. Later clustering based on this similarity combined with learning logic for text classifier was proposed by Al-Mubaid and Umair [2].Bekkerman et al [4] and Dhillon et al [5] proposed various methods for feature clustering, but all those had various disadvantages.

A new fuzzy similarity based self-constructing algorithm was proposed by Jung-Yi Jiang el al [8] for clustering. This method is an incremental feature clustering approach, where each word in the document set are represented as distributions and the similar words are grouped into clusters. Also each cluster is represented by a Membership function with statistical mean and deviation. The feature extracted from each is weighted combination of all the features in the cluster. This method was faster than other approaches and the extracted features were also better.

We propose a Pointwise Mutual Information (PMI) [10] scheme for determining word similarity. Thus the words with high PMI similarity are grouped into a cluster. The reminder of this paper is detailed as follows: In Section 2, we explain the existing schemes for feature reduction and fuzzy similarity measure for clustering in detail and its working principle. In Section 3, presents the proposed PMI based feature clustering algorithm.

Experimental results are given in Section 4 and finally, conclusion is given in section 5.PMI is a semantic similarity measure where similarity between two concepts or words is found using their context overlap.

II. RELATED WORK

A. Fuzzy clustering algorithms for mixed feature variables

Symbolic variables may present human knowledge, nominal, categorical and synthetic data etc. Since1980s, cluster analysis for symbolic data had been widely studied. Miin-Sheneta l[11] created a FCM objective function for symbolic data and then proposed the FCM clustering for symbolic data. They connected fuzzy clustering to deal with symbolic data. Fuzzy clustering algorithms for mixed features of symbolic and fuzzy data were proposed Numerical examples and comparisons are also given. Numerical examples illustrate that the modified dissimilarity gave better results.

B. Dimension Reduction in Text Classification with Support Vector Machines

Support vector machines (SVMs) have been recognized as one of the most successful classification methods for many applications including text classification by Hyunsoo Kim et al [6]. Even though the learning ability and computational complexity of training in support vector machines may be independent of the dimension of the feature space, reducing computational complexity is an essential issue to efficiently handle a large number of terms in practical applications of text classification. A novel dimension reduction method to reduce the dimension of the document vectors dramatically was adopted. And also decision functions for the centroid-based classification algorithm and support vector classifiers to handle the classification problem where a document may belong to multiple classes were introduced. Substantial experimental results shows that with several dimension reduction methods that are designed particularly for clustered data, higher efficiency for both training and testing can be achieved without sacrificing prediction accuracy of text classification even when the dimension of the input space is significantly reduced.

C. Fuzzy Feature Clustering(FFC)Algorithm

Fuzzy similarity-based self constructing algorithm was an incremental feature clustering approach to reduce the number of features for the text classification task proposed by Jung-Yi Jiang el al [8]. Here Cluster characterization was done by Gaussian distribution.

A document set D of n documents $d_1, d_2, \ldots, d_n$, with the feature vector W of m words $w_1, w_2, \ldots, w_m$ and p classes $c_1, c_2, \ldots, c_p$ are given as input to the classifier. First step is to construct word pattern for each word in W. For word w_i, its word pattern x_i is defined by,

$$x_i = < x_{i1}, x_{i2}, \ldots, x_{ip} >$$

$$= < \quad P(c_1|w_i), P(c_2|w_i), >$$

where

$$P(c_j|w_i) = \frac{\sum_{q=1}^{n} d_{qi} \times \delta_{qj}}{\sum_{q=1}^{n} d_{qi}}$$

for $1 \leq j \leq p$. Note that d_{qi} indicated number of occurrences of w_i in document d_i. Also δ_{qj} is defined as

$$\delta_{qj} = \begin{cases} 1, \textit{if document } d_q \textit{ belongs to class } c_j \\ 0, \qquad\qquad\qquad\qquad \textit{otherwise} \end{cases}$$

The words in W are grouped in to clusters based on these word patterns. Each cluster was characterized by a membership function which is the product of 'p' one - dimensional Gaussian functions. gives fuzzy similarity of a word pattern x to cluster G.

$$\mu_G(x) = \prod_{i=1}^{p} exp\left[-\left(\frac{x_i - m_i}{\sigma_i}\right)^2\right]$$

where m_i and σ_i are mean and deviations of G respectively.

$$m_i = \frac{\sum_{j=1}^{q} x_{ij}}{|G|} \qquad \sigma_i = \sqrt{\frac{\sum_{j=1}^{q} (x_i - m_i)^2}{|G|}}$$

here the value of membership function is between 0 and 1(ie., appr. 1), thus a word pattern close to the mean of a cluster is regarded to very similar to this cluster.

Feature extraction was expressed in the following form, D' = DT where,

$$D = [d_1 \quad d_2 \quad \cdots \quad d_n]^T,$$

$$D' = [d'_1 \quad d'_2 \quad \cdots \quad d'_n]^T,$$

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1k} \\ \vdots & \ddots & \vdots \\ t_{m1} & \cdots & t_{mk} \end{bmatrix}$$

where D is the matrix consisting of original document of m features and D' was the matrix consisting of converted documents with new k features. T was the weighting matrix. The goal of feature reduction was achieved by finding an appropriate T such that k is smaller than m. The elements of T is binary ie., if a word pattern belongs to a cluster then the value of t_{ij} is 1, otherwise 0. The elements of T are derived based on the obtained clusters, and feature extraction was done.

Since there are k clusters, there are k extracted features. Three weighting approaches were proposed:

- The hard-weighting approach, each word was only allowed to belong to a cluster, and so it only contributes to a new extracted feature
- The soft-weighting approach, each word was allowed to contribute to all new extracted features, with the degrees depending on the values of the membership functions.
- The mixed-weighting approach was a combination of the hard-weighting approach and the soft-weighting approach.

Given a set D of training documents, text classification was done in following steps:

- *Step1*: Specify the similarity threshold ρ for Membership function, and apply clustering algorithm. Assume that k clusters are obtained for the words in the feature vector W.

- *Step 2*: Find the weighting matrix T and convert D to D'.
- *Step 3*: Using D' as training data, a classifier based on support vector machines (SVM) is built. A SVM can only separate apart two classes (+1 or -1). Therefore for p classes, p SVMs are created. Thus classifier is the aggregation of these SVMs.
- *Step 4*: (Classifying unknown documents) suppose, d is an unknown document, first convert d to d'. Then feed d' to the classifier and get p values, one from each SVM, d belongs to those classes with 1 appearing at the outputs of their corresponding SVMs.

The proposed system replaces the existing Fuzzy similarity for Semantic similarity using PMI for Feature clustering, which worked better than existing algorithm.

III. PMI SIMILARITY

In the existing methods the similarity between two words for clustering are calculated based on their distributions among the documents. For two given words, distributional similarity obtains collective contextual information for each word and computes how similar their context vectors are. That is for two concepts[1] to be similar it does not require the concepts to co-occur in the same contexts[2]. Distributional similarity has the ability to find "indirect" similarity but it cannot classify them accordingly. In contrast semantic or PMI [10] similarity determines how much commonality they share for two concepts to be similar.

Typically a concepts meaning can be determined by its context in which it occur. Therefore considering this fact, two concepts are said to be semantically similar if their contexts overlap as explained in Fig. 1.

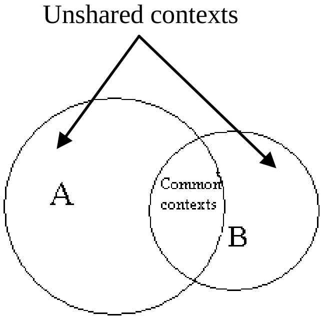

Fig. 1 Context overlap between concepts A and B

If the overlap between contexts of two concepts is larger, then co-occurrences of two concepts are also large. Therefore it is clear that the extent of commonality of context between two concepts is determined by the number of co-occurrences.

PMI is a normalized measure of co-occurrences to represent the similarity.

$$PMI(c_1, c_2) \approx \log\left(\frac{f_d(c_1, c_2}{f_{c1} \cdot f_{c2}}\right. \quad (1)$$

Where f_{c1} and $f_{c2}$are individual frequencies of two concepts c_1 and $c_2$in the corpus and $f_d(c_1, C_2)$ is the co-occurrence frequency of two concepts c_1 and c_2 in the context window of size *d*. *N*is the total number of words in the corpus[3].The optimal value of *d* was between 16 - 32 obtained by Terra et al [16].

A. Incremental Feature Clustering approach

In our proposed approach, initially there is no cluster and a new cluster is created if needed. For each pair of word PMI similarity is calculated using (1) and compared with a threshold value (ρ). If the calculated PMI value is high then both the words are placed in the same cluster otherwise a new cluster is created for the dissimilar words. Threshold (ρ) is a predefined value based on the preferences of the user. If the value of ρ is high, then the number of clusters formed is also large, thus resulting more extracted features. Hence for smaller value of ρ, minimum features are extracted. Also care should be taken, that the value of ρ should not be too high or too low, which may result in false similarity measures.

B. Feature Extraction and Classification

For each cluster, a head word is selected based on occurrences (f_c) of the word in the corpus. Head word is the extracted feature for each cluster and has higher f_c value among all the other features in the cluster. This head word is compared with the new incoming word for PMI similarity.

Given a set of Documents D and their corresponding classes as training dataset for designing a classifier, the steps are,

- Preprocess the documents to obtain set of features.
- Fix the threshold (ρ) and apply the proposed PMI based feature clustering approach for clustering.
- Extract on feature per cluster to obtain D'
- Using D' as the training data construct classifier that identifies the class label for the new document.

Classifier is designed based on comparison between the distributions of features of the training document set among the clusters with the distribution features from the new document, for which the class label to be identified. If more number features from the new document is grouped in a cluster G, then the class label for the new document will be the class of the document for which there is more features grouped into this cluster G.

[1] A concept is referred as a particular sense of word.

[2] A context is a part of text or statement that surrounds a word and determines its meaning.

[3] A corpus is a large collection of writings of a specific kind or on a specific subject.

IV. EXPERIMENTAL RESULTS

The dataset selected is 20 Newsgroups [1] which contains a collection of 20,000 articles and are distributed over 20 classes. After preprocessing there were 25,7198 features in the dataset. Fig. 2 shows the execution time for the feature reduction method that uses our PMI approach and FFC algorithm. The x-axis in the graph represents the number of features extracted for various threshold ρ. The y-axis represents the execution time for the feature reduction procedure. Our proposed PMI based Feature Clustering approach is abbreviated as PMIFC in the figure.

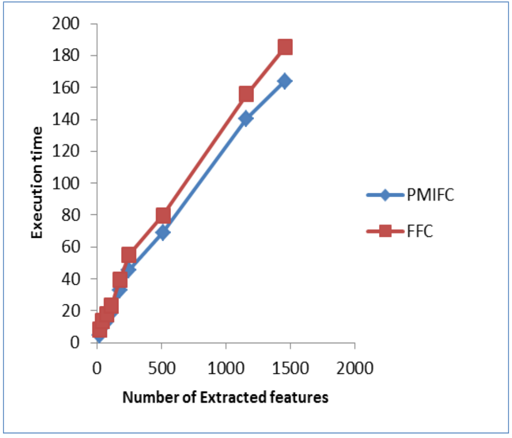

FIG.2 EXECTION TIME(SEC) FOR PMIFC AND FFC ALGORITHM

Table. 1 gives the values of points in the Fig. 2 for different values of ρ. It is clear from the graph that as the threshold value increases, the number of extracted features and execution time also increases. It if found that execution time is reduced for PMIFC approach.

TABLE. 1 : SAMPLE EXECUTION TIMES (SECONDS) OF OUR APPROACH ON 20 NEWSGROUPS DATASET

No. of extracted features	**20**	**45**	**80**	**114**	**180**	**250**	**515**	**1160**	**1460**
Threshold (ρ)	2	3	4	5	6	7	8	9	10
PMIFC	6.51	11.3	15.8	21.4	36.8	53.7	77.9	140.4	178.8
Threshold (ρ)	0.01	0.02	0.03	0.06	0.12	0.19	0.23	0.32	0.36
FFC	8.1	13.9	17.6	23.4	39.3	55.3	79.7	155.9	185.2

V.CONCLUSION

The PMI based clustering approach is proposed which finds similarity between the features based on their semantic similarity. PMI similarity works on the basic idea that if two words are similar then their contexts overlap and the context determines the words meaning. Hence the amount of co-occurances of the concepts in the context is determined for the similarity calculation. Based on this similarity the features are cluster. For each cluster a head word is selected based on its number of occurrences on the corpus. This extracted feature issued for indexing the document. Classifier is designed for this training dataset and used to find class label for new document. The execution time for proposed system is comparatively minimum for PMIFC and also better feature reduction was achieved.

VI. FUTURE ENHANCEMENT

PMI similarity assumes that a word has single sense, but reality better feature reduction can achieved if polysemy of words are considered. PMI_{max} is an enhancement on PMI that considers polysemy of word to find similarity. PMI_{max} estimates correlation between the closest senses of two words also. Thus PMI_{max} can combine PMI in the future for better results.

REFERENCES

[1] Http://people.csail.mit.edu/jrennie/20Newsgroups/, 2010.

[2] H. Al-Mubaid and S.A. Umair, "*A New Text Categorization Technique Using Distributional Clustering and Learning Logic*", IEEE Trans. Knowledge and Data Eng., vol. 18, no. 9, pp. 1156-1165, Sept. 2006.

[3] L.D. Baker and A. McCallum, "*Distributional Clustering of Words for Text Classification*", Proc. ACM SIGIR, pp. 96-103, 1998.

[4] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "*Distributional Word Clusters versus Words for Text Categorization*", J. Machine Learning Research, vol. 3, pp. 1183-1208, 2003.

[5] I.S.Dhillon, S.Mallela, R.Kumar, "*A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification*", Journal on Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

[6] Hyunsoo Kim, Peg Howland, Haesun Park, "*Dimension Reduction In Text Classification With Support Vector Machines*", Journal of Machine Learning Research, vol.6, pp. 37–53, 2005.

[7] Jun Yan, Benyu Zhang, Ning Liu, Shuicheng Yan, "*Effective And Efficient Dimensionality Reduction For Large-Scale And Streaming Data Preprocessing*" IEEE Transactions on Knowledge And Data Engineering, vol. 18, no. 3, pp. 320 – 333, Mar 2006.

[8] Jung-Yi Jiang, Ren-JiaLiou, and Shie-Jue Lee, "*A Fuzzy Self-Constructing Feature Clusterig Algorithm for Text Classification*", IEEE Transaction on Knowledge and Data Engineering, vol. 23, no. 3, pp. 335 – 349, Mar. 2011.

[9] Kaur, A.J. Hornof, "*A Comparison of LSA, Wordnet and PMI-IR for Predicting User Click Behavior*", Proceedings of Conference on Human Factors in Computing Systems, pp. 51-60, 2005.

[10] Lushan Han, Tim Finin, Paul McNamee, "*Improving Word Similarity by AugmentingPMI with Estimates of Word Polysemy*", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1307-1322, June 2013.

[11] Miin-Shen Yang, Pei-Yuan Hwang, De-Hua Chen, "*Fuzzy Clustering Algorithms for Mixed Feature Variables*", Elsevier - Fuzzy Sets and Systems, vol. 141, no. 2, pp. 301–317, 2004.

[12] Monica Rogati, Yiming Yang, "*High-Performing Feature Selection For Text Classification*", CIKM '02 - Proceedings of the 11th International Conference on Information and knowledge management, pp. 659-661, 2002.

[13] Shady Shehata, FaKhriKarray, "*A Concept-Based Model for Enhancing Text Categorization*", KDD '07 - Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 629 – 637, 2007.

[14] N. Slonim, N. Tishby, "*The Power of Word Clusters for Text Classification*", Proceedings of 23rd European Colloquium on Information Retrieval Research (ECIR), 2001.

[15] P. Turney, "*Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL*", Proceedings of 12th European Conference on Machine Learning, pp. 491-502, 2001.

[16] E. Terra and C.L.A. Clarke, "*Frequency Estimates for Statistical Word Similarity Measures*", Proc. Human Language Technology and North Am. Chapter of the ACL Conf., pp. 244-251, 2003.