FAST FUZZY FEATURE CLUSTERING FOR TEXT CLASSIFICATION

Megha Dawar[1] and Dr. Aruna Tiwari[2]

Department of Computer Engineering,
Shri Govindram Seksaria Institute of Technology and Science
23 Park Road, Indore, India
[1]meghacsgs@gmail.com
[2]atiwari@sgsits.ac.in

ABSTRACT

Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. In this paper, Fast Fuzzy Feature clustering for text classification is proposed. It is based on the framework proposed by Jung-Yi Jiang, Ren-Jia Liou and Shie-Jue Lee in 2011. The word in the feature vector of the document is grouped into the cluster in less iteration. The numbers of iterations required to obtain cluster centers are reduced by transforming clusters center dimension from n-dimension to 2-dimension. Principle Component Analysis with slit change is used for dimension reduction. Experimental results show that, this method improve the performance by significantly reducing the number of iterations required to obtain the cluster center. The same is being verified with three benchmark datasets.

KEYWORDS

Feature Clustering, Text Classification, Principle Component Analysis (PCA)

1. INTRODUCTION

The goal of text classification [1] is the classification of documents into a fixed number of predefined categories. The first step in text classification is to transform documents, which typically are strings of characters, into a numeric representation suitable for the classification task. This numeric representation scheme leads to very high-dimensional feature spaces containing 10000 dimensions and more. One issue for text classification [1] is the creation of compact representations of the feature space and the discovery of the complex relationships that exist between features, documents and classes. In this context feature selection [2], [3], [4] and feature extraction [5] have been used for feature reduction. Problem in Feature selection is that only a subset of the original words is used. Useful information that can be provided by the unused words may be ignored. And complexity of feature extraction is high. Thus feature clustering is used for dimensionality reduction. It is used as a feature reduction method; features are clustered into groups based on a similarity measure. Feature, which fulfil same similarity measure criteria, are grouped into single event and this create new, reduced-size event feature spaces. Featured dimensionality can be drastically reduced.

In this paper Fast Fuzzy Feature clustering for text classification is proposed. The work presented in this paper is based on the framework proposed by Jung-Yi Jiang, Ren-Jia Liou and Shie-Jue Lee in A Fuzzy Self Constructing feature clustering algorithm for text classification [6]. In this

Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 167–172, 2012.

DOI : 10.5121/csit.2012.2317

paper we proposed fast feature clustering which reduce clusters center dimension from n-dimension to 2-dimension using principle component analysis. This method improves the performance by significantly reducing the number of iterations required to obtain the cluster center.

2. Motivation Of The Algorithm

The first feature extraction method based on feature clustering was proposed by Baker and McCallum [7], which was derived from the "distributional clustering" idea of Pereira et al. [8]. Al-Mubaid and Umair [9] used distributional clustering to generate an efficient representation of documents and applied a learning logic approach for training text classifiers. The Agglomerative Information Bottleneck approach was proposed by Tishby et al. [10], [11]. The divisive information-theoretic feature clustering algorithm was proposed by Dhillon et al. [12], which is an information-theoretic feature clustering approach. Jung-Yi Jiang, Ren-Jia Liou and Shie-Jue Lee used A Fuzzy Self Constructing feature clustering algorithm for text classification [6], which is an incremental feature clustering approach to reduce the number of features for the text classification task, and is more effective than other feature clustering methods. However, difficulties are associated with these methods is that dimension of clusters center is equal to the number of classes in class set, which degrades its performance when number of classes are more.

The above consideration motivated research to reduce the number of iterations required to obtain clusters center. In Fast Fuzzy Feature Clustering, the number of iteration required to obtain the cluster center is reduced by using Principle Component analysis [13] approach and a transformation algorithm. This improves significant performance as compared to the previous framework and become a significant topic of recent interest.

3. Our Method

Our proposed approach is an agglomerative clustering algorithm approach. The words in the feature vector of a document set are represented as distributions, and processed one after another. Initially each word represents a cluster. Suppose a document set D of n documents $\{d_1, d_2, \dots, d_n\}$ together with the feature vector W of m words $\{w_1, w_2, \dots, w_m\}$ and p classes $\{c_1, c_2, \dots, c_p\}$, then word pattern $\vec{w} = \langle x_1, x_2, \dots, x_m\rangle$ for each word in W constructed. Based on these word pattern Clusters are created. For word w_i its word pattern x_i is defined, similarly as in [12], by

$$x_i = P(c_i|w), 1 \le i \le p$$

$$x_i = \langle x_{i1}, x_{i2}, \dots, x_{ip}\rangle \quad \text{where}$$

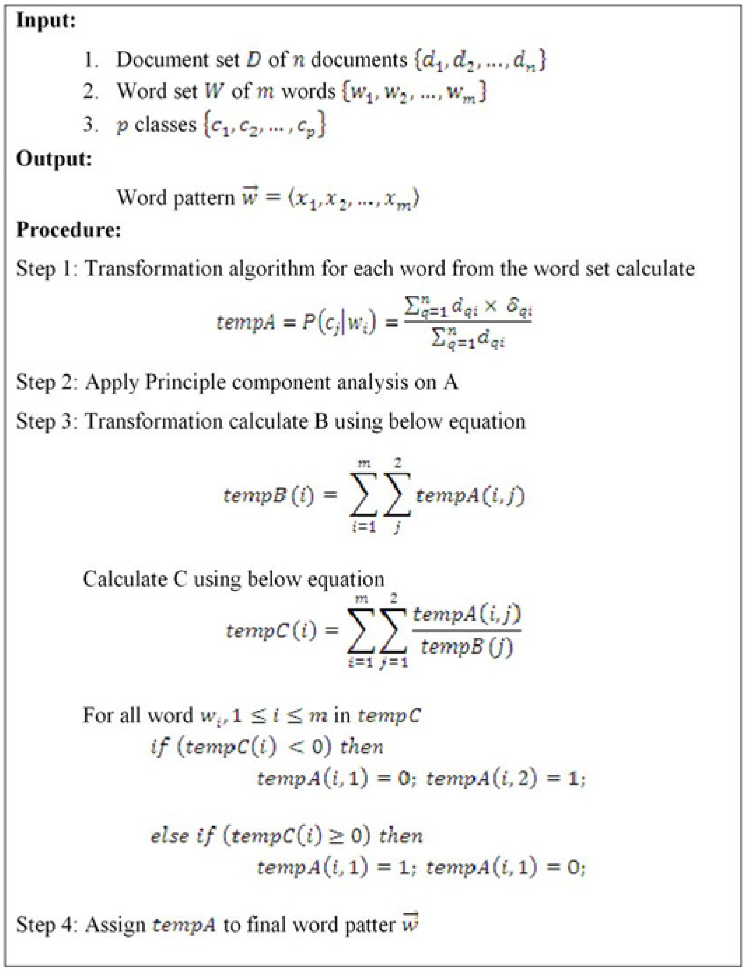
$$P(c_j|w_i) = \frac{\sum_{q=1}^{n} d_{qi} \times \delta_{qi}}{\sum_{q=1}^{n} d_{qi}}$$

$for\ 1 \le j \le p$. d_{qi} Indicates the number of occurrences of w_i in document d_q and δ_{qi} is defined as:

$$\delta_{qi} = \begin{cases} 1, & if\ document\ d_q\ belong\ to\ class\ c_j; \\ 0, & otherwise. \end{cases}$$

It is these word patterns on which our proposed clustering algorithm works. Principle Component analysis is used to reduce this word pattern x from p - dimension to 2 - dimension. All center coordinated should be positive and within the range from 0 to 1, since it is fuzzy based approach. Therefore transformation algorithm is used for this purpose and finally word pattern $x_i = \langle x_{idm}, x_{idm} \rangle$ is obtained where $dm = 2$. Figure 1 represent transformation algorithm.

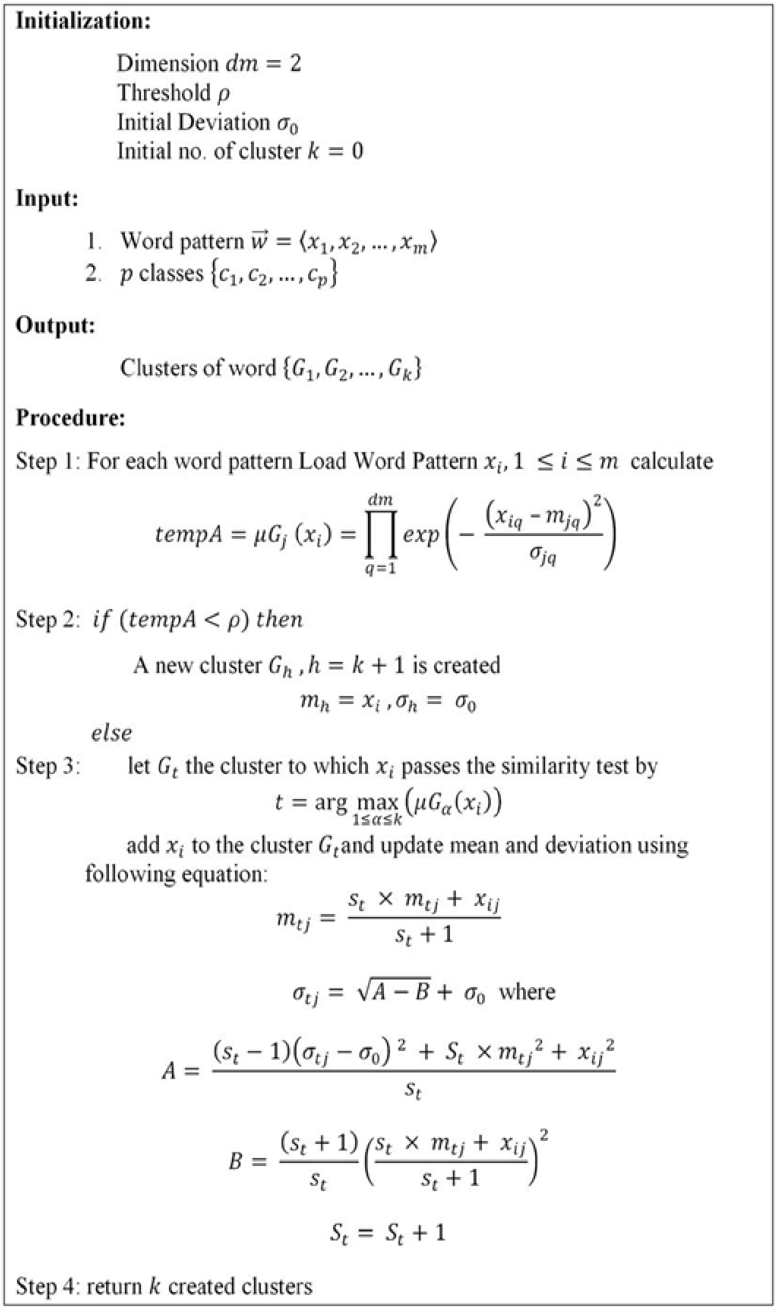
Figure 1.



```
Input:
        1. Document set $D$ of $n$ documents $\{d_1, d_2, \dots, d_n\}$
        2. Word set $W$ of $m$ words $\{w_1, w_2, \dots, w_m\}$
        3. $p$ classes $\{c_1, c_2, \dots, c_p\}$
Output:
        Word pattern $\vec{w} = \langle x_1, x_2, \dots, x_m \rangle$
Procedure:
Step 1: Transformation algorithm for each word from the word set calculate
        $tempA = P(c_j|w_i) = \frac{\sum_{q=1}^{n} d_{qi} \times \delta_{qi}}{\sum_{q=1}^{n} d_{qi}}$
Step 2: Apply Principle component analysis on A
Step 3: Transformation calculate B using below equation
        $tempB(i) = \sum_{i=1}^{m} \sum_{j}^{2} tempA(i,j)$
    Calculate C using below equation
        $tempC(i) = \sum_{i=1}^{m} \sum_{j=1}^{2} \frac{tempA(i,j)}{tempB(j)}$
    For all word $w_i, 1 \leq i \leq m$ in $tempC$
        $if\ (tempC(i) < 0)\ then$
            $tempA(i,1) = 0;\ tempA(i,2) = 1;$
        $else\ if\ (tempC(i) \geq 0)\ then$
            $tempA(i,1) = 1;\ tempA(i,1) = 0;$
Step 4: Assign $tempA$ to final word patter $\vec{w}$
```

Transformation Algorithm

Now it is these reduce dimension word pattern on which fuzzy feature clustering algorithm works. For each word pattern, the similarity of this word pattern to each existing cluster is calculated using Gaussian function, to decide whether it is combined into an existing cluster or a new cluster is created. Once a new cluster is created, the corresponding membership function should be initialized. On the contrary, when the word pattern is combined into an existing cluster, the membership function of that cluster should be updated accordingly. Figure 2 represent Fast Fuzzy Feature Clustering Algorithm. After obtaining the clusters Feature extraction is performed according to [6]. Three weighting methods are used for feature extraction Hard weighting, Soft weighting, and Mixed weighting. After applying these features extraction methods reduced dimension input dataset is obtained. Further Text classification can be performed on these reduced dimension datasets using any classification algorithm such as Support Vector Machine (SVM).



```
Initialization:
        Dimension $dm = 2$
        Threshold $\rho$
        Initial Deviation $\sigma_0$
        Initial no. of cluster $k = 0$
Input:
    1. Word pattern $\vec{w} = \langle x_1, x_2, \dots, x_m \rangle$
    2. $p$ classes $\{c_1, c_2, \dots, c_p\}$
Output:
        Clusters of word $\{G_1, G_2, \dots, G_k\}$
Procedure:
Step 1: For each word pattern Load Word Pattern $x_i$, $1 \le i \le m$ calculate
            $tempA = \mu G_j\,(x_i) = \prod_{q=1}^{dm} exp\left(-\frac{\left(x_{iq} - m_{jq}\right)^2}{\sigma_{jq}}\right)$
Step 2: $if\ (tempA < \rho)\ then$
            A new cluster $G_h$ , $h = k + 1$ is created
                $m_h = x_i$ , $\sigma_h = \sigma_0$
    $else$
Step 3:     let $G_t$ the cluster to which $x_i$ passes the similarity test by
                $t = \arg\max_{1 \le \alpha \le k} \left(\mu G_\alpha(x_i)\right)$
            add $x_i$ to the cluster $G_t$and update mean and deviation using
    following equation:
                $m_{tj} = \frac{s_t \times m_{tj} + x_{ij}}{s_t + 1}$
                $\sigma_{tj} = \sqrt{A - B} + \sigma_0$ where
                $A = \frac{(s_t - 1)\left(\sigma_{tj} - \sigma_0\right)^2 + S_t \times {m_{tj}}^2 + {x_{ij}}^2}{s_t}$
                $B = \frac{(s_t + 1)}{s_t}\left(\frac{s_t \times m_{tj} + x_{ij}}{s_t + 1}\right)^2$
                $S_t = S_t + 1$
Step 4: return $k$ created clusters
```

Figure 2. Fast Fuzzy Feature Clustering Algorithm

4. EXPERIMENT AND RESULTS

The proposed algorithm is implemented in Matlab 7.8.0(R2009a) and applied on various data sets. Results of these experiments are summarized in Table 1. Datasets used for training are Reuter-21578 R8, 20 Newsgroup and WebKB data set. All of these data sets are obtainable from http://web.ist.utl.pt/~acardoso/datasets/. The input data sets files are pre-processed, which provide require input set for the learning. The input data set files is converted into numeric data on the basis of number of occurrence of words in the documents. Proposed algorithm determines number of iterations in obtaining cluster center. Furthermore we use F-FFC and FFC to represent Fast Fuzzy Feature Clustering and Fuzzy Self Constructing Feature Clustering respectively. Execution time for both F-FFC and FFC is calculated in second.

Table 1. Sample Execution Time (in sec) on Different Dataset

Dataset	**No. of Instances**	**No. of Features**	**No. of Classes**	**F-FFC**	**FFC**
Reuters-21578 R8	391	3125	8	2.734	88.536
20 Newsgroups	400	9005	20	30.498	1792.285
WebKB	400	3491	4	3.110	20.756

5. CONCLUSIONS

The proposed algorithm Fast Fuzzy Feature Clustering improves significant performance as compared to the previous framework. By using this proposed algorithm desired no. of cluster center are obtained in less iteration. Each cluster is used as one extracted feature, which reduced the dimensionality of the input samples.

ACKNOWLEDGEMENTS

The authors would like to thank everyone.

REFERENCES

[1] Antomia Kyriakopoulou , (2008) “Text Classification Aided by Clustering: a Literature Review”, Tools in Artificial Intelligence, pp, 233-252.

[2] A.L. Blum and P. Langley, (1997) “Selection of Relevant Features and Examples in Machine Learning,” Aritficial Intelligence, vol. 97, nos. 1/2, pp. 245-271.

[3] E.F. Combarro, E. Montanes, I. Dıaz, J. Ranilla, and R. Mones, (2005) “Introducing a Family of Linear Measures for Feature Selection in Text Categorization,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 9, pp. 1223-1232.

[4] Y. Yang and J.O. Pedersen, (1997) “A Comparative Study on Feature Selection in Text Categorization,” Proc. 14th Int’l Conf. Machine Learning, pp. 412-420, 1997.

[5] D.D. Lewis, (1992) “Feature Selection and Feature Extraction for Text Categorization,” Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[6] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, (2011) “A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification”, IEEE Transactions on Knowledge and Data Engineering, VOL. 23, NO. 3, pp.335-349.

[7] L.D. Baker and A. McCallum, (1998) “Distributional Clustering of Words for Text Classification,” Proc. ACM SIGIR, pp. 96-103.

[8] F. Pereira, N. Tishby, and L. Lee, (1993) “Distributional Clustering of English Words,” Proc. 31st Ann. Meeting of ACL, pp. 183-190.

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

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

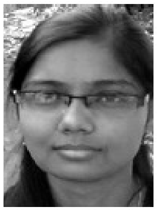
[11] N. Slonim and N. Tishby, (2001) "The Power of Word Clusters for Text Classification," Proc. 23rd European Colloquium on Information Retrieval Research (ECIR).

[12] I.S. Dhillon, S. Mallela, and R. Kumar, (2003) "A Divisive Infomation-Theoretic Feature Clustering Algorithm for Text Classification," J. Machine Learning Research, vol. 3, pp. 1265-1287.

[13] Lindsay I Smith, (2002) "A tutorial on Principal Components Analysis"

Authors

Megha Dawar received her Bachelor of Engineering degree in Computer Engineering from RGPV University, India in 2010. She is currently pursuing Master of Engineering in Computer Engineering from SGSITS, Indore, India. Her research interests include Data mining, and Soft Computing.



Dr. Aruna Tiwari She is currently working as Associate Professor in Computer Engineering Department at SGSITS Indore, India. Her research interest areas are Data mining, Computational Learning and Soft Computing.