Feature Clustering algorithms for text classification-Novel Techniques and Reviews

Bhagyasri. A
M.Tech Scholar
Department of CSE
SCET, Narasapur, AP, India

P. Praneetha
Assistant Professor
Department of CSE
SCET, Narasapur, AP, India

I.Kali Pradeep
Assistant Professor
Department of CSE
VITB, Bhimavaram

G. Ravindra Bharathi
Assistant Professor
Department of CSE
VITB, Bhimavaram

Abstract

Classifying text is a challenging technique. The dimensionality of feature vectors is very huge and number of researchers has found many techniques to reduce the number of dimensions of feature vectors in text classification. In this paper, some of the important techniques for text classification have been reviewed and novel parameters using fuzzy set approach have been discussed in detail. Here, the words are grouped into clusters based on degree of similarity. Mean and standard deviation are considered as membership function for each cluster. At last features are extracted from each cluster. The extracted feature equals to the weighted combination of words in each cluster. Moreover, there is no need to specify the number of cluster in advance by trial and error method.

1. Introduction

Normally, the number of dimensions in feature vector is very large. So, Feature reduction techniques are used for reducing the amount of time and complexity for text classification. The number of features can be reduced by either feature selection or feature extraction. In feature selection, the important dimensions are considered, whereas in feature extraction new dimensions are identified which are more effective than the original dimensions. In practice it has been found that feature extraction is more effective than the feature selection. The feature extraction techniques are basically classified into linear and non linear transformations. Linear transformation includes techniques like, Maximum Margin Criterion, Principal Component Analysis, and Linear Discriminant Analysis etc…. and non-linear transformations include Laplacian Eigen maps, Local Linear Embedding etc…
The main applications of text classification are:

- News: Electronic News articles are generated very frequently. Manual classification of these articles is a very difficult. So, automated methods are useful in this case. This application is known as text filtering.
- Digital libraries: A variety of supervised methods may be used for document organization in domains like digital libraries, web collections and scientific literature. .
- Feedback Mining: Customer Feedback of many products are given on many review sites, E-commerce sites etc. Text classification algorithms are used by companies to review their products automatically.
- E mail classification and spam filtering: In these days E mail service providers can identify a mail as spam by certain words and patterns. And mails are also classified as promotion, social networking and important based on certain word clustering techniques

2. Feature clustering

Feature clustering is one of the famous techniques of feature extraction, where similar features are grouped into a cluster and every cluster is considered as a new feature. Many Techniques were introduced in history. Some of the techniques for text classification through feature clustering are discussed below.

The first feature extraction method based on feature clustering was proposed by Baker and McCallum[5] which were derived from the "distributional clustering". In this approach words are clustered into groups based on the distribution of class labels associated with each word. This method compresses feature space well and maintains high document classification accuracy. But the main disadvantage of this approach is hard clustering.

The Agglomerative Information Bottleneck approach was proposed by Tishby et al [6]. In this approach, text categorization is done by combining distributional clustering of words and a Support Vector Machine. This word-cluster representation is computed using Information Bottleneck method, which generates a compact and efficient representation of documents. . When combined with the classification power of the Support Vector Machine, this method yields high performance in text categorization.

The divisive information-theoretic feature clustering algorithm was proposed by Dhillon et al [9]. In comparison to agglomerative strategies this approach was much faster and achieves comparable or higher classification accuracy. But, this approach uses hard clustering, where a word is assigned strictly to a single subset. Mean and variance were not computed when similarities between the clusters are considered. The main disadvantage of this method is to consider number of new feature which are specified by the user in advance.

3. Co-Clustering

In co-clustering, both feature clustering and document clustering is done in a two stage procedure. Slonim and tishby[8] proposed a double clustering algorithm in year 2000. In this approach, at the first stage feature clustering generates coarser pseudo features, which reduce noise and sparseness that might be exhibited in the original feature space. Then, in the second stage, documents are clustered as distributions over the "distilled" pseudo features, and therefore generate more accurate document clusters.

The iterative double clustering algorithm was proposed by (Yaniv & Souroujon). This approach is an extension of simple double clustering algorithm. But, Number of iterations is made to improve clustering quality. This high data is achieved due to the generation of progressively less noisy data representations

On the other hand Dhillon[7], proposed an information-theoretic co clustering algorithm that intertwines both row (feature) and column (document) clustering. . Here, the row-clustering incorporates column-clustering information and vice versa. The algorithm iterates until it almost accurately reconstructs the original distribution discovers the natural row and column partitions and recovers the ideal compressed distribution. This method is especially useful when the data is sparse.

4. Fuzzy Approach

In the above approaches we need to specify the number of clusters in advance. And these methods are based on hard clustering concept(sharp cutoffs). In fuzzy approach, Rather than having a precise cutoff between categories or sets, fuzzy logic uses truth values between certain ranges to represent the degree of membership that a certain value has in a given category. Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" but "fuzzy" .I.e. data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

4.1Text classification by self constructing fuzzy approach

As, discussed above there are three disadvantages in the above method
1) The number of features to be extracted should be specified by the user in prior to the clustering. 2) Variance measure is not considered in clustering similarity analysis. 3) Hard clustering.

The detailed process is explained below.

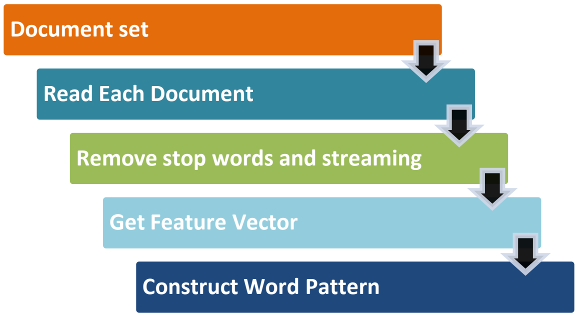

Figure1. The Initial Processing

4.1.1 Word Pattern construction: This is mainly the starting step of the process. Consider,

- m words (feature vectors) as w1,w2,…..,w3
- Document set D with n documents d1, d2, d3,… …,dn
- P classes c1.c2.c3… cp

Here, Word patterns are constructed by taking into considerations the probability for a class with presence of word. This is denoted as

Xi=< X_{i1}, X_{i2} ,…………. X_{in} >
= < $P(C_1/W_i)$, $P(C_2/W_i)$ ……………., $P(C_n/W_i)$ >

Where,

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

$d_{qi=}$ number of occurrences of w_i in document d_q

δqj= 1, If document dq belongs to class Cj
0, Otherwise

4.1.2 Automatic cluster formation: The two main advantages of this approach is 1) No initial number of clusters need to be specified in advance. And 2) Self Constructing I.e., Word patterns are considered one after another.

Initially, there are no clusters and can be created if necessary. Each word pattern is compared with existing clusters for similarity. If similar clusters are found then, the word pattern is grouped into cluster. Otherwise, a new cluster is formed. The similarity between two word patterns is calculated by:

$$\mu_{G_j}(\mathbf{x}_i) = \prod_{q=1}^{p} \exp\left[-\left(\frac{x_{iq} - m_{jq}}{\sigma_{jq}}\right)^2\right]$$

Where,
$m_{jq=}$ Mean of j^{th} Cluster I.e., m_j = <m_{j1},m_{j2}, . . .,m_{jp}>
σ_{jq}= Standard Deviation of j^{th} cluster I.e., σ_j= <σ_{j1}, σ_{j2},σ_{j3},…,σ_{jp}>
The membership function should be updated when a cluster is added to a cluster. A membership function should be initialized when a new cluster is found

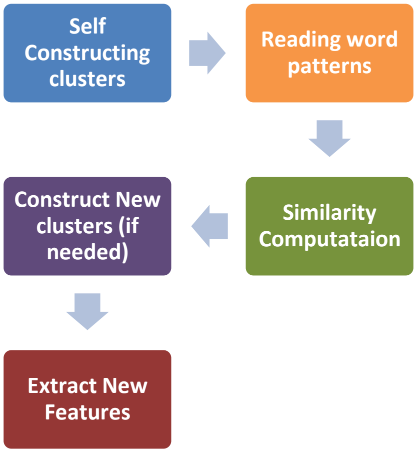

Figure2.The process of new feature discovery

4.1.3 Extraction of features: This part is the most important module. The extraction of features is done by grouping word patterns into clusters by self constructing approach which is discussed above and extracting unique features from each cluster.
Basically, the extraction function is given by the formula S'=ST, where S is the weighing matrix. The elements of T are derived based on the obtained clusters, and feature extraction will be done. Here, three weighing approaches are used.

- **Hard clustering**: In hard clustering each word pattern is exactly in one cluster.
- **Soft Clustering**: In soft clustering every word pattern has a membership function in every cluster.
- **Mixed Clustering**: In mixed clustering, is a combination of hard and soft clustering, where user threshold is used to control its degree of hardness or softness.

4.1.4 **Text Classification:** From above, we can obtain weighing matrix T. Now, classifier is constructed based on new training data I.e., $S^{|}$ by using techniques like naïve Bayesian theorem, Decision tree induction, or support vector machine. If we consider support vector machine, SVM are constructed for each class and aggregated. At this stage we can classify unknown documents. For Example, consider a unknown document 'S' we first construct a modified document '$S^{|}$' from S. Now this modified document '$S^{|}$' is used as training data for classification. The SVM is obtained for each class and original document 'S'is assigned class to one which is the output of corresponding SVM.

5. Conclusion

Unlike other types of data, Text data have may features. In this paper we have discussed about different types of feature extraction techniques using clustering method. At first, text documents like agglomerative technique, divisive techniques and distributive clustering were discussed. Then, we have reviewed about co-clustering techniques which combine both text and document classification. In the next section we have highlighted fuzzy based clustering approach for classifying the text data. This fuzzy approach considers mean and standard deviation as membership function of each cluster. Moreover, there is no need to specify initial number clusters in advance.

6. References

[1] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification" IEEE transactions on knowledge and data engineering, vol. 23, no. 3, march 2011

[2]Sainani Arpitha, Dr.P.Raja Prakash Rao "Clustering Algorithm for Text Classification Using Fuzzy Logic"
International Journal of Advanced Research in Computer Science and Software Engineering. Volume 2, Issue 8, August 2012

[3] Dinesh Kavuri , Pallikonda Anil Kumar , Doddapaneni Venkata Subba Rao "Text and Image Classification using Fuzzy Similarity Based self constructing algorithm" international journal of engineering science & advanced technology Volume-2, Issue-6, pg:1572 – 1576

[4] A.Kavitha, Y.Sowjanya Kumari, Dr.P.Harini "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification" IOSR Journal of Engineering Volume 2, Issue 9 (September 2012), Pg 36-44.

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

[6] 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.

[7] Inderjit S. Dhillon " Co-clustering documents and words using Bipartite Spectral Graph Partitioning" KDD 2001 San Francisco, California.

[8] Noam Slonim, Naftali Tishby " Document clustering using word clusters via the information bottleneck method" Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval Pages 208-215

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

[10] Ran El-yaniv ad Oren Souroujon Iterative double clustering for unsupervised and semi-supervised learning in In Advances in Neural Information Processing Systems conference.

[11] A. Krishna Mohan, V.V.Narasimha Rao, MHM Krishna Prasad "A Novel Fuzzy Based Clustering Algorithm for Text Classification" IJCSNS International Journal of Computer Science and Network Security, VOL.13 No.5.

[12]T. Joachims, "Text Categorization with Support Vector Machine: Learning with Many Relevant Features," Technical Report LS-8- 23, Univ. of Dortmund, 1998.