CLASSIFICATION OF TEXT DATA USING FEATURE CLUSTERING ALGORITHM

Avinash Guru[1], Asma Parveen[2]

[1]*MTech 4th sem,Department of Computer Science and Engineering,KBN College of EngineeringGulbarga,Karnataka, India*

[2]*HOD, Department of Computer Science and Engineering, KBN College of Engineering Gulbarga, Karnataka, India*

Abstract

Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. Generally clustering means the collection of similar objects or data in groups. In this paper, we propose a feature clustering algorithm for classifying the text data. The document set contains number of words; these words are grouped into clusters based on the similarity. Words that are similar to each other are grouped into the same cluster, and the words that are not similar are grouped in another cluster. Each cluster is characterized by a membership function with statistical mean and deviation. When all the words are fed in the document then the clusters are formed automatically. Then the extracted feature starts functioning as it is based on the weighted combination of the words. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Earlier, the user has to specify the extracted feature in advance but now it is not required as the clusters are formed automatically and the trial and error method can be avoided. The experimental results show that our method can run faster and obtain better extracted features than other methods.

Keywords:*Feature clustering, feature extraction, feature reduction, text classification*.

--*--**

1. INTRODUCTION

In text classification, generally the dimensionality of feature vector is huge, and it is difficult to classify the large dimensional data. Hence to reduce this difficulty the feature reduction approaches is applied. There are two major approaches used in this feature reduction. They are; feature selection and feature extraction. This dissertation contributes to the subject area of Data Clustering, and also to the application of Clustering to Image Analysis. Data clustering acts as an intelligent tool, a method that allows the user to handle large volumes of data effectively. The basic function of clustering is to transform data of any origin into a more compact form, one that represents accurately the original data. The compact representation should allow the user to deal with and utilize more effectively the original volume of data. The accuracy of the clustering is vital because it would be counter-productive if the compact form of the data does not accurately represent the original data. One of our main contributions is addressing the accuracy of an established fuzzy clustering algorithm.

Typically, a set of numeric observations, or features, are collected of each object.The collected feature-sets are aggregated into a list which then acts as the input to achosen computational clustering algorithm. This algorithm then provides a descriptionof the grouping structure which it has discovered within the objects.

1.1 Fundamental Concepts of Clustering

Generally clustering means the combination of similar objects or data in a group. Based on the similarity test we classify the data into different clusters. The words that are similar are grouped in one cluster and the words which are different are grouped in another cluster. The computing revolutionof the sixties and seventies gave momentum to this new field because, for the first time,Computers enabled the processing of large amounts of data and took the burden of thevery large amounts of computation generally involvedif translated to modern formalisms, Linnaeus's quotation is very relevant to theclustering problem. Linnaeus uses the term natural distinction; this is the much sought after goal of clustering finding an “intrinsic classification” or an “inherent structure”in data. The better we are at finding an inherent structure in data,the more knowledge we possess about it. As the bigger the volume of data is more numerous objects, the more necessary it is todevelop better clustering methods.

1.2 Contributions

- We studied and investigated the FCMalgorithm (Fuzzy c-Means Clustering Algorithm) thoroughly and identified its main strengths and weaknesses.
- We developed a systematic method for analyzing FCM's classification accuracy when it is used to cluster data sets that contain clusters of very different sizes and populations.

- We proposed a new algorithm, based on FCM, which performs far more accurately than FCM on data sets like those described above. We also investigated performance properties of our new algorithm.
- The feature clustering algorithm is used to reduce the dimensionality of the features in text classification.
- By applying this algorithm, the derived membership function matches closely and provides the exact results.

2. EXISTING SYSTEM

In the existing system we have the Bottleneck approach. These approaches provide the divisive information-theoretic feature clustering, In this system we have some set of original words present in the document. Each time when we want to form a new cluster we have to compare the words with the original words. Hence when the words matches then only the cluster is formed otherwise no cluster. Hence this system works on the concept of trial and error method; this is one of the major disadvantages of the existing system.

3. PROPOSED SYSTEM

We propose a feature clustering algorithm, which is mainly used to reduce the number of features in the text classification. The words in the feature vector of a document set are represented as distributions, and processed one after another. Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. If a word is not similar to any existing cluster, a new cluster is created for this word.

4. SYSTEM ARCHITECTURE

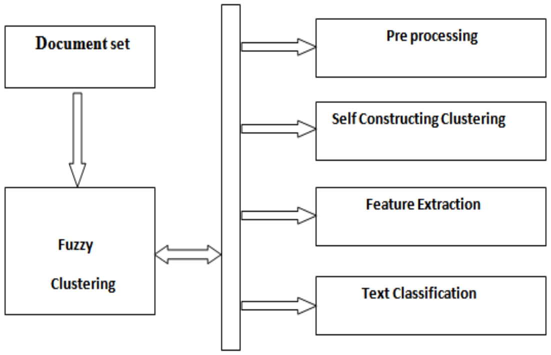

Fig: Architecture Diagram

4.1 Preprocessing

In this module we construct the word pattern of training document set. Read the document set and remove the stop words and perform stemming process. Get the feature vector from the training document .Next we construct the word pattern.

4.2 Self-Constructing Clustering

In this module, we use the self-constructing clustering algorithm. First we read each word pattern, then we compare the similarity based on the original words. If the word matches with given set of words then the word is grouped in the existing cluster and if the word does not match then it is grouped in a new cluster.

4.3 Feature Extraction

Feature extraction module begins; here we compute the cluster in three different ways: hard weight, soft weight, mixed weight, In the hard weight clustering the data is divided into crisps, where the data indicates exactly one cluster. Degree of membership is either 0 or 1 and this hard clustering method leads to local optimum In the soft-weighting approach, each word is allowed to contribute to all new extracted features, with the degrees depending on the values of the membership functions. The mixed-weighting approach is a combination of the hard-weighting approach and the soft-weighting approach.

4.4 Text Classification

Given a set D of training documents, text classification can be done as follows: Get the training document set and specify the similarity threshold ρ. Assume that k clusters are obtained for the words in the feature vector W. Then find the weighting matrix T and convert D to D`. Using weka we classify the text. Weka is a collection of machine learning algorithms for data mining tasks.

5. CONCLUSIONS

In this work, we have presented a feature clustering algorithm. By using this algorithm each cluster is used as an extracted feature and this reduced the dimensionality of data.

REFERENCES

[1].J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi,and Z. Chen, “Effective and Efficient Dimensionality Reductionfor Large-Scale and Streaming Data Preprocessing,” IEEETrans.Knowledge and Data Eng., vol. 18, no. 3, pp. 320-333, Mar. 2006

[2].G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining Multi-LabelData,” Data Mining and Knowledge Discovery Handbook, O. Maimonand L. Rokach eds., second . Springer, 2009

[3]. H. Kim, P. Howland, and H. Park, “Dimension Reduction in Text Classification with Support Vector Machines,” J. Machine Learning Research, vol. 6, pp. 37-53, 2005.

[4]. F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[5]. B.Y. Ricardo and R.N. Berthier, Modern Information Retrieval. Addison Wesley Longman, 1999.

[6]. E.F. Combarro, E. Montan˜ e´s, I. Dı´az, J. Ranilla, and R. Mones, "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, Sept. 2005.
[7]. K. Daphne and M. Sahami, "Toward Optimal Feature Selection," Proc. 13th Int'l Conf. Machine Learning, pp. 284-292, 1996.
[8]. R. Kohavi and G. John, "Wrappers for Feature Subset Selection," Aritficial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997
[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]. D. Ienco and R. Meo, "Exploration and Reduction of the FeatureSpace by Hierarchical Clustering," Proc. SIAM Conf. Data Mining,pp. 577-587, 2008.