Fuzzy Self-Constructing Feature Clustering (FFC) Method for Text Classification

PILLI SUDHEER,
Associate Professor,
Department of Computer Science and Engineering,

NARASIMHA DEVANI,
Assistant Professor,
Department of Electronics and Communications Engineering,

Siddhartha Institute of Technology and Sciences,
Narapally, Hyderabad, Telangana – 500 088.

Abstract

Feature clustering is an effective approach for text classification that reduces the dimensions of features extracted. For feature clustering, a fuzzy similarity-based self-constructing method is presented. Based on a similarity test, the words in a document set's feature vector are classified into clusters. Words with similar meanings are put together in a cluster. A membership function with statistical mean and deviation distinguishes each cluster. When some of the phrases have been entered, the system automatically creates the required number of clusters. There is one extracted feature for each cluster. The extracted feature, which corresponds to a cluster, is a weighted combination of the cluster's words. The resulting membership functions from this technique closely resemble and accurately depict the true distribution of the training data. Furthermore, the user does not need to define the number of extracted features in advance, which eliminates the requirement for trial-and-error to determine the suitable number of extracted features. Experiments demonstrate that it can extract characteristics faster and more accurately than previous approaches.

1. Introduction

Data mining is the method of uncovering new patterns in huge data sets using approaches from artificial intelligence, machine learning, statistics, and database systems. The main purpose of the data mining process is to extract information from a data collection in a human-understandable structure, which includes database and administration steps in addition to the raw analysis stage.

The fundamental data mining job is to analyse vast amounts of data automatically or semi-automatically in order to identify previously discovered interesting patterns such as groupings of data records (cluster analysis), atypical records (anomaly identification), and relationships (association rule mining). Typically, this entails the use of database technologies such as spatial indexing.

Clustering is a data mining (machine learning) approach for categorising data pieces without prior knowledge of the group definitions. K-means clustering and expectation maximisation

(EM) clustering are two popular clustering algorithms. Data clustering is a technique for grouping things that have similar features. The criterion for determining similarity depends on the implementation.

Data clustering is a strategy for physically storing information that is conceptually comparable. The amount of database should be reduced in order to improve database system efficiency. Objects with comparable features are grouped together in one group of objects, and a single disc request renders the overall class accessible.

2. Literature survey

The work in machine learning on approaches for dealing with data sets containing a substantial quantity of irrelevant information is reviewed. It is divided into two parts: the problem of picking relevant attributes and the challenge of finding relevant instances. It describes the advancements achieved on these subjects in both empirical and theoretical machine learning studies, and it presents a broad framework for comparing different techniques.

Text categorization, which includes assigning documents to a set of categories automatically, frequently necessitates the administration of a large number of characteristics. The majority of them are inconsequential, while some create noise that might cause the classifiers to be misled. As a result, feature reduction is frequently used to improve the classification's efficiency and efficacy. It is recommended that relevant features be selected using a set of linear filtering measures that are simpler than the commonly used measures.

For several applications, including text classification, support vector machines (SVMs) have been considered as one of the most successful classification methods. Despite the fact that the learning ability and computational complexity of training in support vector machines are independent of the dimension of the feature space, reducing computational complexity is a critical issue in practical applications of text classification to efficiently handle a large number of terms. It uses cutting-edge dimension reduction techniques to drastically reduce the size of document vectors.

The substantial findings show that, even though the measurement of the input space is substantially decreased, high performance for both training and validation can be achieved using several dimension reduction methods designed specifically for clustered data, without having to compromise prediction accuracy of text classification.

3. Methodology

The fuzzy similarity-based self-constructing feature clustering technique is an incremental feature clustering strategy for reducing the amount of features for the text classification job. The words in a document set's feature vector are represented as distributions and processed one after the other. Words that are similar to one another are put together in the same cluster. A

membership function with a statistical mean and deviation characterises each cluster. If a term does not match any of the current clusters, a new cluster is constructed for it.

The similarity between a word and a cluster is defined by taking into account both the clusters mean and variance. After all of the words have been entered, an appropriate number of clusters are produced automatically. The extracted feature corresponding to a cluster is a weighted mixture of the cluster's words. One of the most efficient strategies for feature reduction in text classification is feature clustering.

The aim behind feature clustering is to organise the original features into regions with a high degree of pairwise semantic relatedness. Because each cluster is viewed as a whole new feature, feature dimensionality may be dramatically decreased. The original approach of feature extraction was based on feature clustering.

- Because each cluster is handled as a distinct new feature, the dimensionality of the feature may be minimised.
- A subset of the original words is used to produce each new feature.
- The Feature vectors do not need to be specified ahead of time.

➢ **Self-Constructing Clustering**

Word patterns are examined one at a time. The user does not need to know the number of clusters ahead of time. The similarity of each word pattern to each existing cluster is computed for each word pattern to determine whether it is integrated into an existing cluster or a new cluster is generated. When a new cluster is built, the membership function associated with it should be started. On the contrary, when the word pattern is integrated into an existing cluster, the membership function of that cluster should be modified correspondingly.

➢ **Feature Extraction**

Word patterns have been organised into clusters in the information theoretic feature clustering method mentioned in applying clustering technique, and words in the feature vector W have also been clustered correspondingly. There is an extracted feature for one cluster. There are three weighing methods: hard, gentle, and mixed. Each word is only permitted to belong to one cluster in the hard-weighting strategy, thus it can only contribute to a new extracted feature.

➢ **Feature Reduction**

Prior to performing document classification tasks, feature reduction methods are used. The feature selection strategy typically employs Information Gain. It calculates the decreased uncertainty using an information-theoretic metric and assigns a weight to each word. One of the most efficient strategies for feature reduction in text classification is feature clustering. The aim behind feature clustering is to organise the original features into clusters with a high degree

of pairwise semantic relatedness. Because each cluster is viewed as a whole new feature, feature dimensionality may be dramatically decreased.

- **Text Classification**

The dimension of the feature vector in text classification is frequently big real-world data sets; both include more than 15,000 features. Such high dimensionality can provide a significant challenge to categorization algorithms. The fuzzy similarity-based self-constructing feature clustering technique is an incremental feature clustering strategy for reducing the amount of features for the text classification job. The words in a document set's feature vector are represented as distributions and processed one after the other. Words that are similar to one another are put together in the same cluster.

4. Result and discussion

The proposed system replaces the old system during the system implementation phase. In the system implementation process, the FUZZY feature clustering extraction method is applied. By introducing the suggested system, the FUZZY feature clustering extraction technique replaces the current system. For feature clustering, the current system is integrated and the new system is used.

In the suggested system, feature reduction approaches for text categorization can be used. The program is written in Java and has a SQL server as a backend. The system is designed using the Java Development Kit 1.7. The JDK is installed, and the environmental variables' path settings are configured. To begin the program's execution, use the Java main menu command. The user can choose from a menu of choices and run the software.

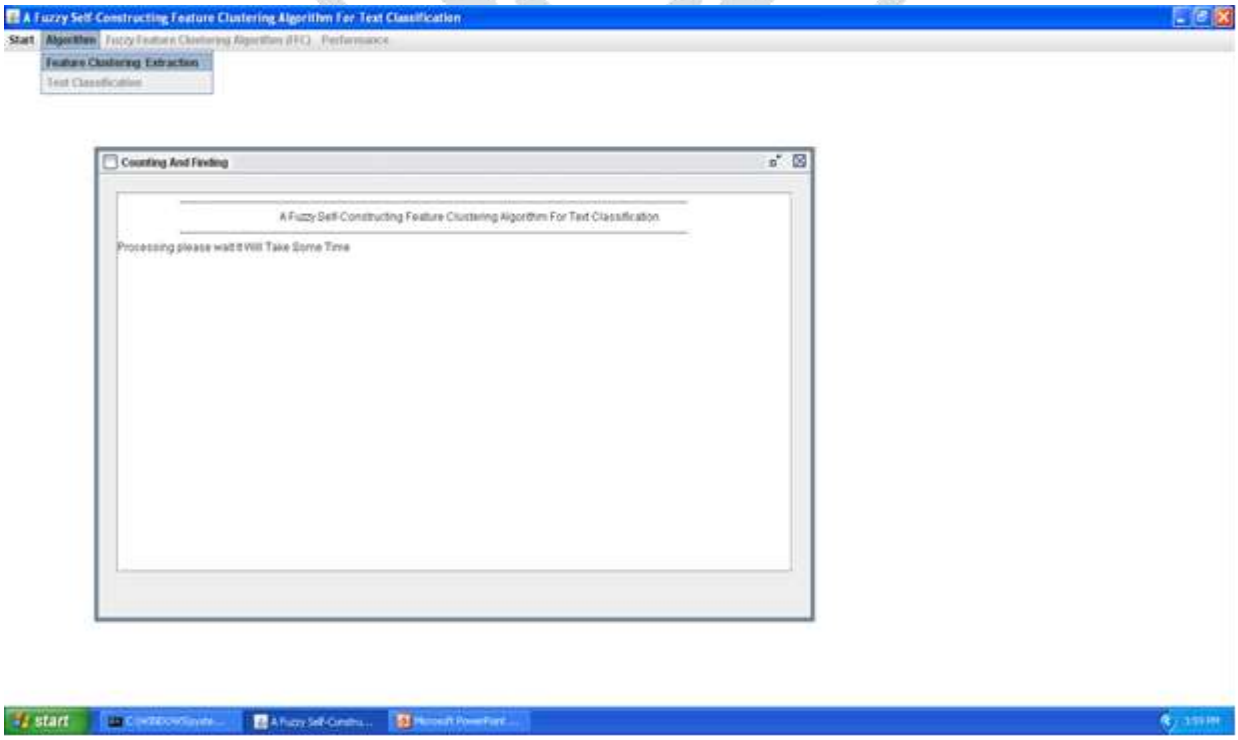

Fig 1.Feature Clustering Extraction Algorithm Start page

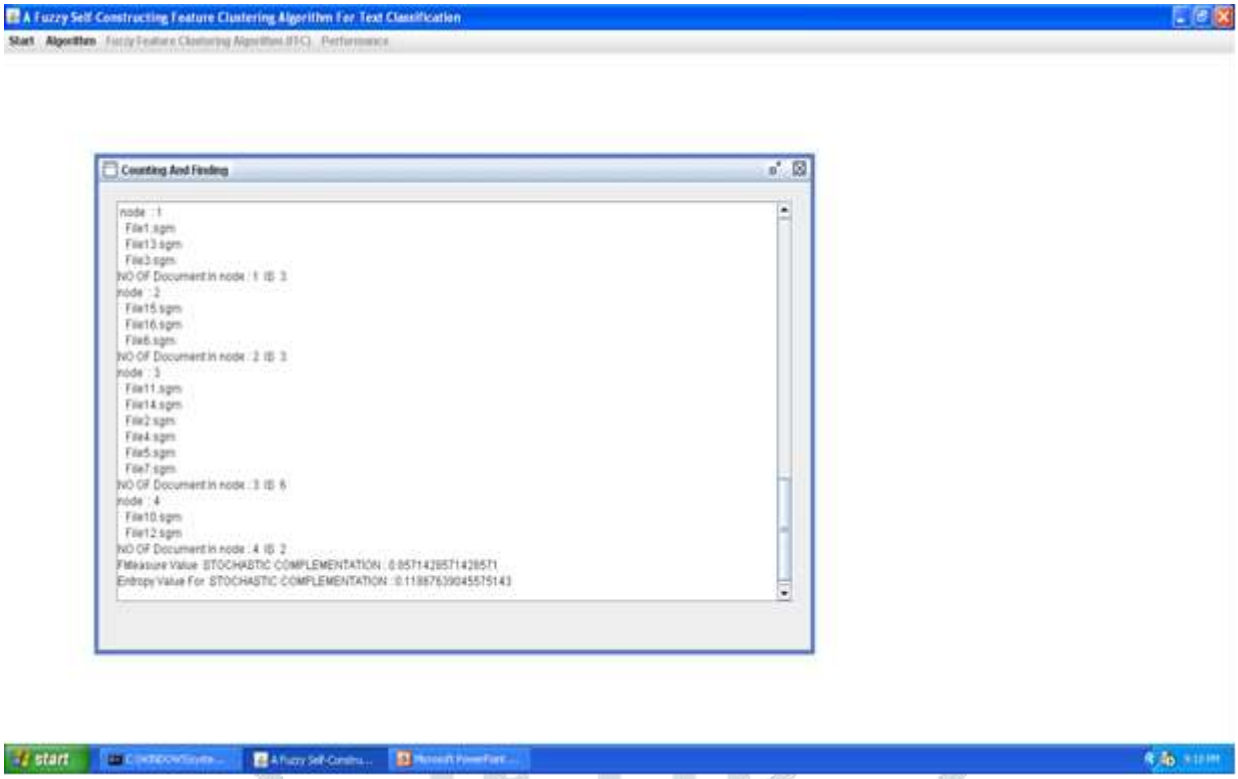

Fig 2. Clustered Features

5. Conclusion

The fuzzy self-constructing feature clustering (FFC) method is an incremental clustering strategy for text categorization that reduces the dimensionality of features. Feature pairs that are comparable are placed together in the same cluster. A membership function with statistical mean and deviation characterises each cluster. If a word does not belong in any of the current clusters, a new one is constructed for it. When comparing a word to a cluster, both the cluster's mean and variance are taken into account. The extracted feature that corresponds to a cluster is a weighted mixture of the cluster's words. The resulting membership functions from this technique closely resemble and accurately depict the true distribution of the training data. Furthermore, the user does not need to define the number of extracted features in advance, which eliminates the requirement for trial-and-error to determine the suitable number of extracted features. Experiments on three real-world data sets have shown that it can extract features faster and more accurately than previous approaches.

References

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

2. Combarro.E.F, Montan.E, Dı´az.I, Ranilla.J, and Mones.R(Sept. 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.

3. Daphne.K and Sahami.M(1996), "Toward Optimal Feature Selection," Proc. 13th Int'l Conf. Machine Learning, pp. 284-292.

4. Jolliffe.I.T(1986), "Principal Component Analysis" Springer-Verlag.

5. Kim.H, Howland.P, and Park.H (2005), "Dimension Reduction in Text Classification with Support Vector Machines," J. Machine Learning Research, vol. 6, pp. 37-53.

6. Kohavi.R and John.G(1997), "Wrappers for Feature Subset Selection," Aritficial Intelligence, vol. 97, no. 1-2, pp. 273-324.

7. Lewis.D.D(1992), "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217.

8. Li.H, Jiang.T, and Zang.K(2004), "Efficient and Robust Feature Extraction by Maximum Margin Criterion," Sebastian.T, Lawrence.S, and Bernhard.S eds. Advances in Neural Information Processing System, pp. 97-104, Springer.

9. Martinez.A.M, and Kak.A.C(2001), "PCA versus LDA," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2 pp. 228-233.

10. Oja.E(1983), "Subspace Methods of Pattern Recognition" Research Studies Press.