A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification

KOTHAGATTU RAMU,
Associate Professor,
Department of Computer Science and Engineering,

KETHAVATH SRILATHA,
Assistant Professor,
Department of Electronics and Communications Engineering,

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

Abstract

Feature clustering is an efficient method for classification tasks that minimises the dimensionality of retrieved features. A Divisive Information-Theoretic Feature Clustering Algorithm for feature clustering is provided. The words in a document set's feature vector are sorted into clusters based on a similarity test. A cluster is formed by grouping words with similar meanings. Each cluster is distinguished by a membership function with a statistical mean and deviation. When some of the sentences are entered, the algorithm generates the necessary number of clusters. Each cluster has one extracted feature. A weighted combination of the cluster's phrases is used to extract the feature that corresponds to a cluster. This approach produces membership functions that closely match and properly represent the real distribution of the training data. Furthermore, the user is not required to specify the number of extracted features in advance, removing the need for trial-and-error to find the appropriate number of extracted features. Experiments show that it can extract attributes more quickly and correctly than earlier methods.

1. Introduction

Clustering is a data mining (machine learning) strategy of categorising data bits that does not require prior knowledge of the group definitions. K-means and expectation maximisation (EM) clustering are two well-known clustering methods. Data clustering is a technique for grouping items with similar characteristics. The implementation determines the criterion for detecting similarity.

Data mining is a technique for detecting novel patterns in large data sets by combining artificial intelligence, machine learning, statistics, and database systems. The primary goal of the data mining process is to extract information from a data set and organise it in a human-readable way, which comprises database and administration phases in addition to the raw analysis stage.

Data clustering is a mechanism for effectively organizing functionally related information. To increase database system performance, the volume of database should be minimised. Objects with similar characteristics are clustered together in one set of entities, and a single disc request makes the whole class available.

The basic data mining task is to analyse massive volumes of data automatically or semi-automatically in designed to check previously found unique patterns such as data file groups (cluster analysis), atypical records (anomaly identification), and correlations (association rule mining). This usually includes the use of database systems like spatial indexing.

2. Literature survey

The significant findings show that, despite a significant reduction in input space measurement, high performance for both training and validation can be achieved using several dimension reduction methods designed specifically for clustered data, without compromising text classification prediction accuracy.

Text categorization, which involves automatically assigning documents to a set of categories, typically needs the management of a huge number of attributes. The bulk of them are insignificant, but a few of them generate noise that might lead the classifiers to be deceived. As a result, feature reduction is widely employed to boost classification efficiency and effectiveness. Relevant features should be chosen using a set of linear filtering measures that are less complex than widely used measures.

The work in machine learning on ways for coping with data sets including a large amount of irrelevant material is discussed. It has two parts: the issue of selecting relevant qualities and the challenge of locating relevant examples. It highlights the progress made in both empirical and theoretical machine learning studies on these topics, as well as a comprehensive framework for comparing different methodologies.

Support vector machines (SVMs) have been regarded as one of the most successful classification methods for a variety of applications, including text categorization. Due to the fact that knowing capability and training complexity of the algorithm in support vector machines are irrespective of feature dimensional space, lowering complexity of the algorithm is a crucial challenge in practical applications of text classification to efficiently handle a high number of words. It dramatically reduces the size of document vectors by employing cutting-edge dimension reduction algorithms.

3. Methodology

Words that are related to one another yield better retrieved characteristics than other approaches. Other is clustered 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 divisive information parameter clustering approach is an incremental feature grouping mechanism used to reduce the number of features required for text categorization. The phrases in the feature space of a document collection are expressed as distributions and evaluated one after the other.

The goal of component grouping is to group the original characteristics into areas that have a significant level of pairwise semantic similarity. Because each cluster is treated as a separate feature, attribute dimensionality may be significantly reduced. The initial feature extraction method relied on feature grouping.

Each word pattern's correlation to each established connection is evaluated to decide whether it is incorporated into an area related or a new cluster is produced. When a new cluster is created, the membership function should be activated. When the word pattern is incorporated into an existing cluster, the membership function of that cluster should be adjusted accordingly.

For one cluster, a feature has been retrieved. Weighing methods are classified as firm, mild, or mixed. In the hard-weighting technique, each word is only allowed to correspond to one cluster, therefore it could only provide to one new feature extraction stage.

Attribute clustering is one of the most effective ways for dimension reduction in text classification. The goal of feature clustering is to group input data into clusters that have a high degree of pairwise semantic relations. As each cluster is treated as a separate feature, feature dimensionality may be significantly reduced.

Feature reduction, feature selection, and feature extraction are the two methods used in general. A new feature set W′ ={w1′, w′2......} is produced using feature selection procedures, which is a subset of the original feature set W. W′ is then utilised as an input in classification tasks. In the feature selection process, Information Gain (IG) is widely used. It uses an information-theoretic metric to calculate the decreased uncertainty and assigns a weight to each word.

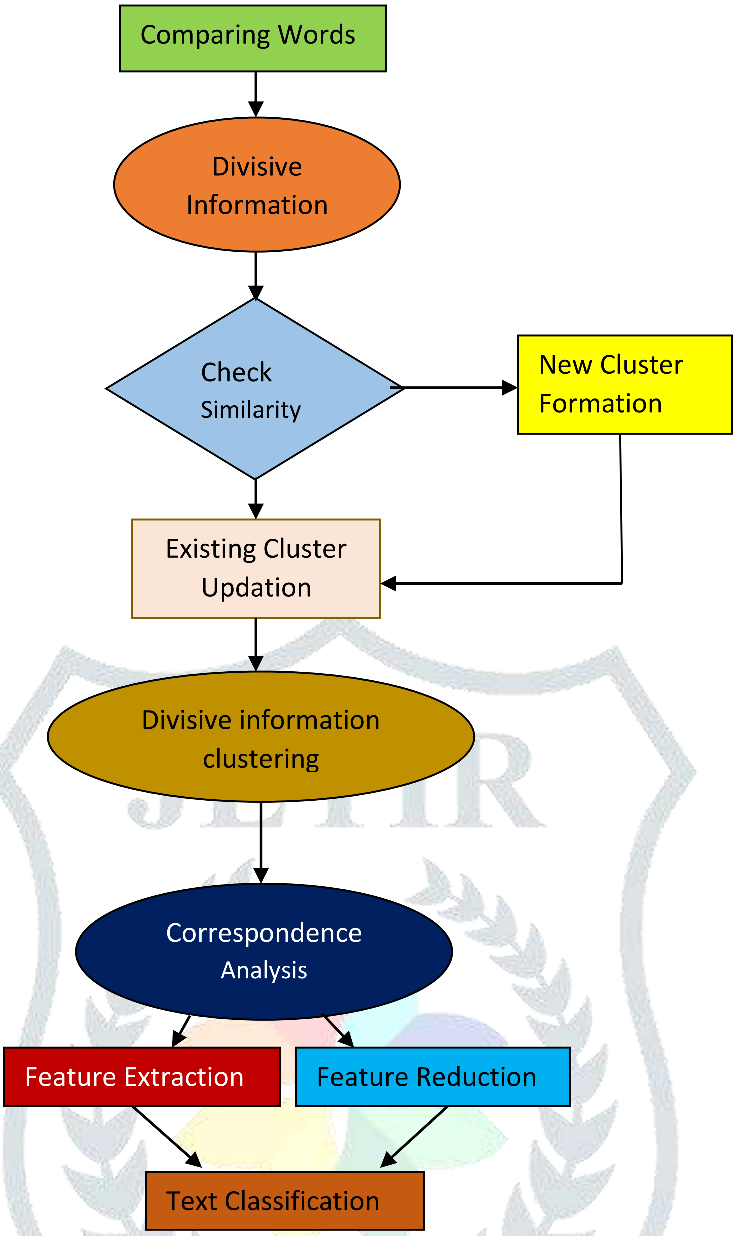
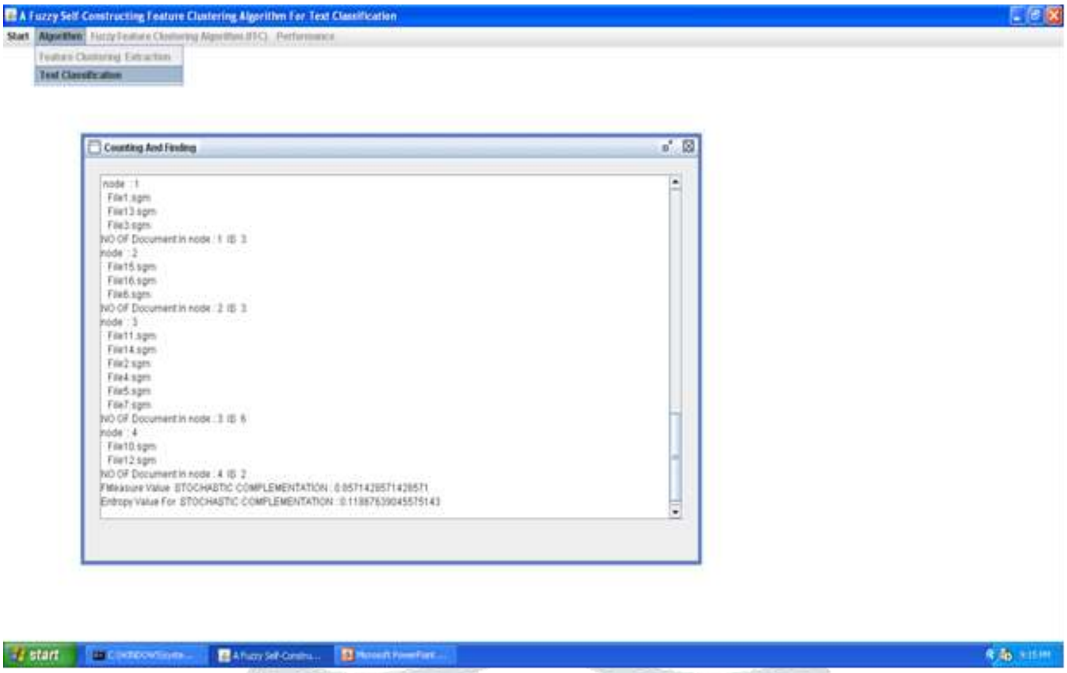


Fig. 1 Data Flow Diagram

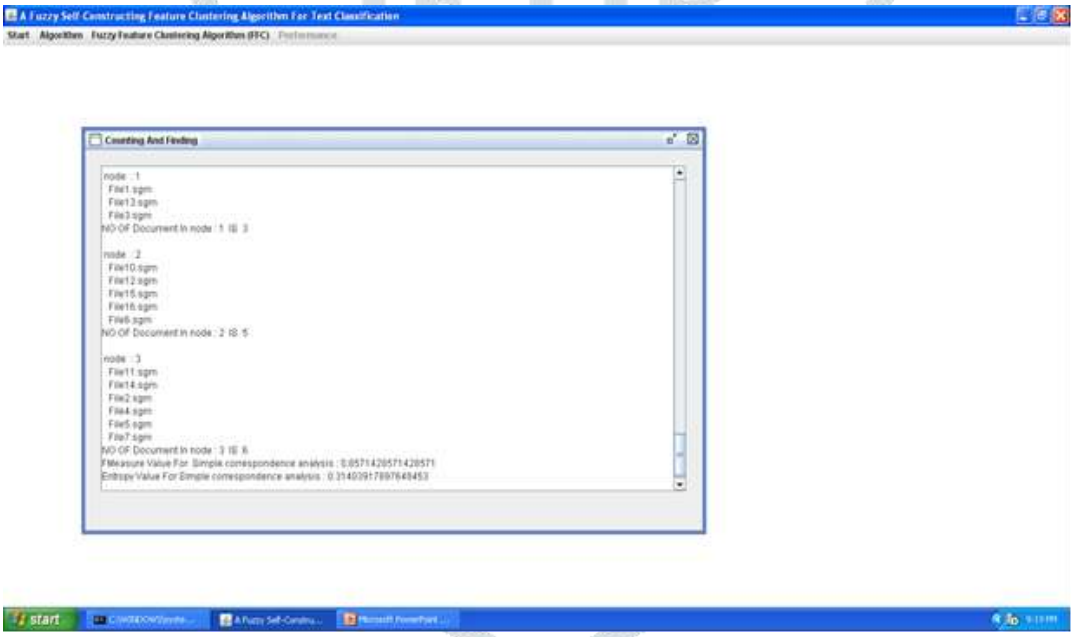
4. Result and discussion

The system's processing quality is determined by the quality of the input. The way data enters the system process is described by input specifications. Input design elements may either assure the system's and procedure's dependability by using accurate data, or they can lead to the creation of erroneous data. The user's ability to engage with the system is also determined by the input design. The Data Report is an ActiveX Designer, which means it is a customised ActiveX object that connects into the Visual Basic ecosystem. The clustered content sets are further grouped depending on the membership function, i.e. every clustering seems to have its own membership function.


A Fuzzy Self Constructing Feature Clustering Algorithm For Text Classification
Start
Algorithm
Fuzzy Feature Clustering Algorithm (FFC)
Performance
Feature Clustering Extraction
Text Classification
Counting And Finding
start

```
node : 1
 File1.sgm
 File13.sgm
 File3.sgm
NO OF Document in node : 1 IS 3
node : 2
 File15.sgm
 File16.sgm
 File6.sgm
NO OF Document in node : 2 IS 3
node : 3
 File11.sgm
 File14.sgm
 File2.sgm
 File4.sgm
 File5.sgm
 File7.sgm
NO OF Document in node : 3 IS 6
node : 4
 File10.sgm
 File12.sgm
NO OF Document in node : 4 IS 2
FMeasure Value STOCHASTIC COMPLEMENTATION : 0.8571428571428571
Entropy Value For STOCHASTIC COMPLEMENTATION : 0.11887639045575143
```

Fig 2. Text Classification


A Fuzzy Self Constructing Feature Clustering Algorithm For Text Classification
Start
Algorithm
Fuzzy Feature Clustering Algorithm (FFC)
Performance
Counting And Finding
start

```
node : 1
 File1.sgm
 File13.sgm
 File3.sgm
NO OF Document in node : 1 IS 3

node : 2
 File10.sgm
 File12.sgm
 File15.sgm
 File16.sgm
 File6.sgm
NO OF Document in node : 2 IS 5

node : 3
 File11.sgm
 File14.sgm
 File2.sgm
 File4.sgm
 File5.sgm
 File7.sgm
NO OF Document in node : 3 IS 6
FMeasure Value For Simple correspondence analysis : 0.8571428571428571
Entropy Value For Simple correspondence analysis : 0.31403917897648453
```

Fig 3. Analysis of Text

5. Conclusion

When matching a word to a cluster, the mean and variance of the cluster are taken into consideration. A weighted mixture of the cluster's words is the retrieved feature that corresponds to a cluster. This approach produces membership functions that closely match and properly represent the real distribution of the training data. Therefore, the user is not required to determine the optimal number of extracted features in ahead, removing the need for trial-and-error to find the appropriate number of extracted features. It can extract characteristics quicker and more correctly than earlier algorithms, according to tests on three real-world data sets.

References

1. 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.
2. 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.
3. Oja.E(1983), "Subspace Methods of Pattern Recognition" Research Studies Press.
4. Park.H, Jeon.M, and Rosen.J(2003), "Lower Dimensional Representation of Text Data Based on Centroids and Least Squares," BIT Numerical Math, vol. 43, pp. 427-448.
5. Ricardo.B.Y and Berthier.R(1999), Modern Information Retrieval. Addison Wesley Longman.
6. Roweis.S.T, and Saul.L.K(2000), "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, pp. 2323-2326.
7. Sebastiani.F(2002), "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47.
8. Slonim.N and Tishby.N(2001), "The Power of Word Clusters for Text Classification," Proc. 23rd European Colloquium on Information Retrieval Research (ECIR).
9. Tenenbaum.J.B, De Silva.V, and Langford.J.C(2000), "A GlobalGeometric Framework for Nonlinear Dimensionality Reduction,"Science, vol. 290, pp. 2319-2323.
10. Yan.J, Zhang.B, Liu.N, Yan.S, Cheng.Q, Fan.W, Yang.Q, Xi.W, and Chen.Z,(2006) "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 3, pp. 320-333.