KModes Clustering for Data Categorization

SARITHA BANOTH,
Assistant Professor,
Department of Computer Science and Engineering,

RAMAKRISHNA PASULA,
Assistant Professor,
Department of Electronics and Communications Engineering,

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

Abstract

The feature that corresponds to a cluster is extracted using a weighted combination of the cluster's words. This method yields membership functions that closely resemble and accurately represent the training data's true distribution. Furthermore, the user is not needed to indicate the number of extracted features in advance, which eliminates the requirement for trial-and-error to determine the correct number of extracted features. Experiments demonstrate that it extracts properties faster and more accurately than previous techniques. Feature clustering is a quicker classification approach that reduces the dimensionality of recovered information. For feature clustering, a KModes Clustering is proposed. A similarity test is used to arrange the words in a content set's training samples into clusters. Words having similar meanings are grouped together to form a cluster. A membership function with a statistical mean and deviation distinguishes each cluster. The algorithm produces the required number of clusters when part of the phrases are input. Each cluster has a single selected features.

1. Introduction

For data instances, classification is being used to recognise collective identity. It works in a similar way to clustering in that it divides client records into various "classes." Each record in the set of data used to generate the classifier must contain a value for the attribute used to define classes. Since each dataset includes a value for an attribute that is used to define classes, and end users choose which attribute to utilise. The classifier's job is to determine how entries should be classed, not to examine data to find segments.

Semantic similarity, also known as semantic relatedness, is a paradigm in which a group of reports or words within a phrase list are allocated a metric based on the closeness of their meaning / semantic content. This can be accomplished, for example, by establishing a topological similarity, utilising taxonomies to determine a distance among words, or employing statistical methods, such as a latent space, to connect words and textual circumstances from a relevant text corpus (co-occurrence).

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. When all of the words have been entered, the system automatically creates the required number of clusters. Each cluster has a single extracted feature. The extracted feature that corresponds to a cluster is a weighted mixture of the cluster's words.

Even though the variations among these magnitudes are minimal, a word is precisely allocated to a subset, i.e. hard-clustering, based on the similarity magnitudes between the word and the existing subsets. In addition, while computing similarity with regard to a cluster, the mean and variance of the cluster are ignored. Furthermore, these solutions need the user to specify the amount of new elements in advance.

2. Literature survey

Despite the fact that the intellectual ability and computation cost of learning in support vector machines are distinct of the dimension of the subspace, minimising computation cost is a critical topic in practical uses of text categorization. Unique features extraction methods are used to drastically reduce the number of variables of the vector space.

To solve the classification tasks where a content may belongs to many classes, there are selection parameters for the centric-based classification technique and support vector classifiers. The extensive experimental findings demonstrate that using numerous dimension reduction approaches created specifically for clustered data, improved training and testing efficiency may be achieved without reducing text classification prediction accuracy.

It employs unique dimension reduction approaches to drastically reduce the dimension of the document vectors. The extensive experimental findings demonstrate that, even when the dimension of the input space is greatly decreased, greater efficiency for both training and testing may be attained using numerous dimension reduction approaches built specifically for clustered data.

As a result, feature reduction is frequently used to improve the classification's efficiency and efficacy. It is suggested that relevant characteristics be selected using a set of linear filtering measures that are simpler than the standard measures used for this purpose. It does trials on two distinct corpora and discovers that the suggested metrics outperform the existing ones.

The paper examines studies in machine learning on approaches for dealing with data sets that contain a substantial quantity of irrelevant information. It focuses on two main issues: the challenge of picking relevant attributes and the problem of finding relevant instances. It describes the achievements achieved on these subjects in both empirical and theoretical work in machine learning, and it presents a broad framework for comparing different techniques.

3. Methodology

The disadvantages of the present system can be eliminated by utilising the suggested system. The current system's primary goal is to create a user-friendly interface. The suggested method now computerises all of the details that were previously kept manually. Once the information is entered into the computer, there is no need for different people to deal with different portions. Only one person is required to keep the entire record. Security can also be provided based on the needs of the users.

Finding an adequate T so that k is less than m achieves the aim of feature reduction. Word patterns have been clustered, and words in the feature vector W have been clustered as well. We have k retrieved features 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.

Each word is permitted to contribute to all new extracted features in the soft-weighting strategy, with the degrees varying based on the membership function values. The mixed-weighting strategy combines the hard- and soft-weighting approaches. By setting 1 or 0 to 1, the merging can be "hard" or "soft." The number of clusters is reduced when the similarity criterion is low, and each cluster covers more training patterns. When the similarity criterion is high, however, the number of clusters grows, and each cluster covers fewer training patterns.

Each cluster is defined by a membership function with a statistical mean and deviation. If a phrase does not fit into any of the existing clusters, a new one is created. The fuzzy similarity-based self-constructing parameter clustering technique is an incremental feature grouping mechanism for text classification that uses fuzzy similarity to decrease the amount of features required. The phrases in a document collection's feature space are expressed as distributions and assessed one by one.

Information gain is a state-of-the-art feature selection strategy, where IOC stands for incremental feature extraction, DC for feature clustering, and FFC for KModes Clustering. Text categorization may be done as follows given a set D of training documents: It uses our clustering technique and specifies the similarity threshold. Assume that the words in the feature vector W are clustered into k clusters. Then, using T as a weighting matrix, change D to D′. A classifier based on support vector machines (SVM) is created using D′ as training data. To accommodate for misclassifications, slack variables I are added.

Five Objectives guiding the design of the input focus on:

- Effectiveness
- Accuracy
- Easy to use
- Consistency

- Attractiveness

The words that were picked from the database are grouped according to their semantic relatedness.

When the theoretical design is refined into an actual system, system implementation is a crucial step of the project. The system should be implemented after thorough testing and validation. All of the operations involved in converting an existing system to a new one are referred to as "system implementation." It's possible that the new system will be completely different from the previous one. It may be necessary to replace an existing manual or automated system.

4. Result and discussion

To store the features, the system employs a SQL server as a back end. In the database, tables are formed. The SQL server back end is where the database tables are kept. Jdbc odbc drivers are used to connect to the database. The system is written in Java and has a SQL database as its backend. In order to build the system, the Java development kit 1.7 is utilised. The JDK is installed, and the environment variables' path settings are specified. To begin the programme execution, utilise the Java main menu command. The user has the ability to choose from a menu of settings and run the software.

A large number of test cases are used to assess feature clustering. Test cases serve as a check list for the testing procedure. The test cases identify the facts that are taken into account throughout the testing procedure. During the testing process, this test case is employed. The designer creates the test cases using information from the database. All testing is done using the test cases as a guide. Some of the system's key test cases are feature clustering extraction and text categorization.

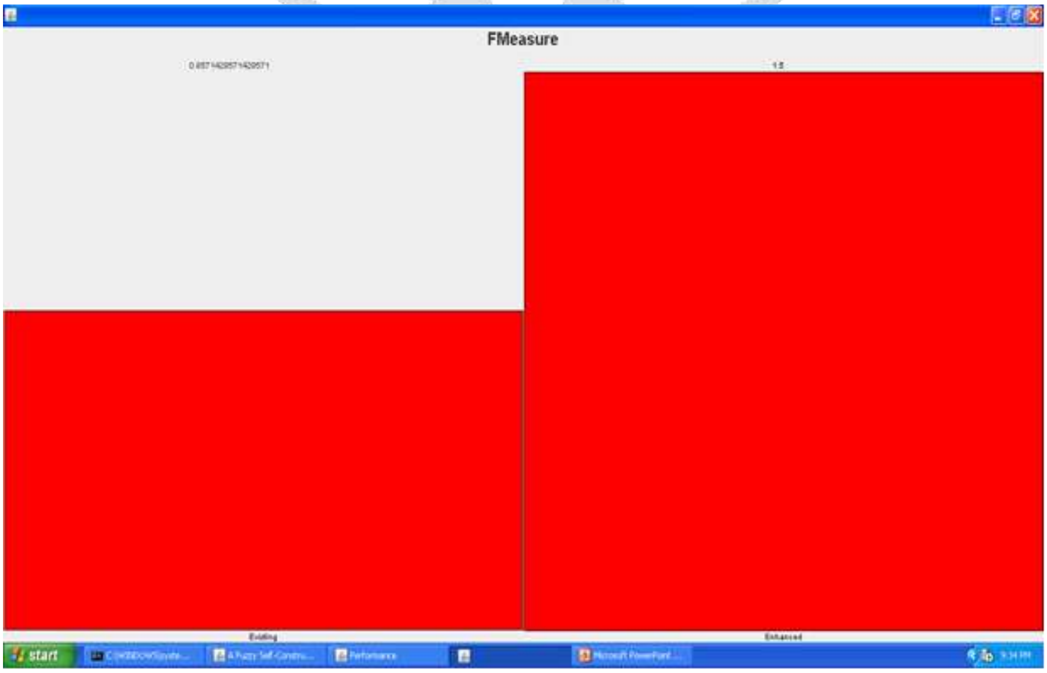

Fig 1. F-Measure Graph Model

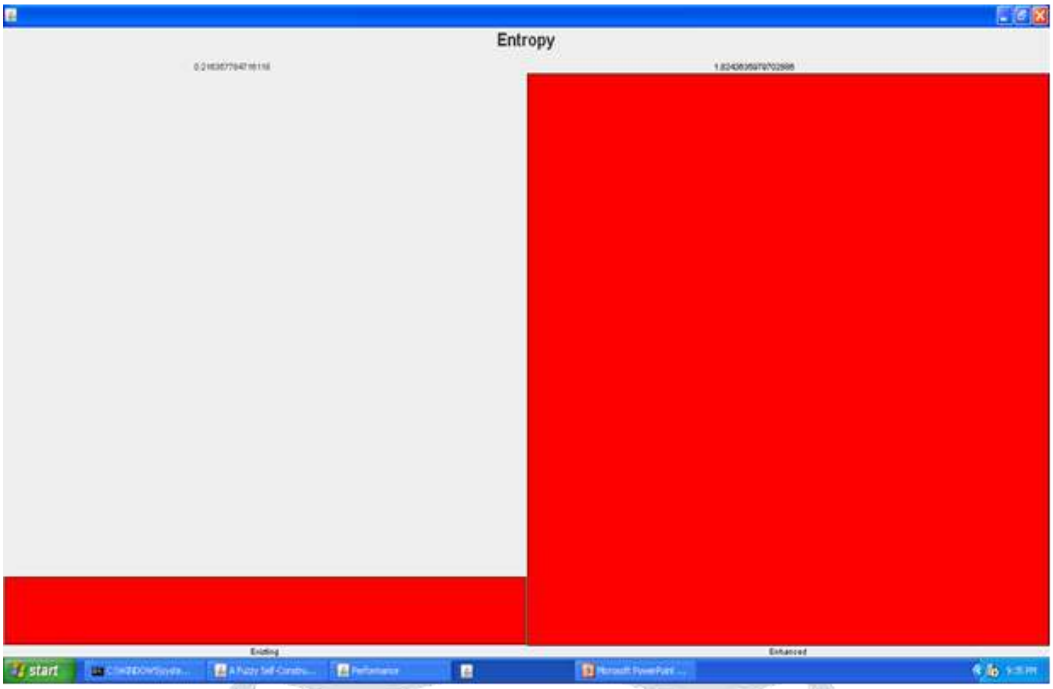

Fig 2. Entropy Graph Model

5. Conclusion

The quality of the input determines the system's processing quality. Input requirements explain how data enters the system process. Input design features may either ensure the system and procedure's dependability by employing reliable data, or they can cause erroneous data to be created. The input design also influences the user's ability to interact with the system. The Data Report is an ActiveX Designer, which means it's a custom ActiveX object that works with Visual Basic. Every clustering appears to have its own membership function, therefore the clustered content sets are further classified based on the membership function. 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. It can extract characteristics quicker and more correctly than earlier algorithms, according to tests on three real-world data sets.

References

1. 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.
2. Kohavi.R and John.G(1997), "Wrappers for Feature Subset Selection," Aritficial Intelligence, vol. 97, no. 1-2, pp. 273-324.
3. Lewis.D.D(1992), "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217.

4. 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.
5. 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.
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.