International Journal of Science Engineering and Advance Technology

Imminent Rift Assortment Algorithm for Elevated Facet Data Using Wanton Collecting

Matta Rambabu#1,Mayyappa Chakravarthi#2
#1 Student of M.Tech Department of Computer Science Engineering G.V.R & S college of Engineering & Technology, GUNTUR
#2 Department of Computer Science Engineering G.V.R&S college of Engineering & Technology, GUNTUR.

Abstract

Feature subset clustering is a powerful technique to reduce the dimensionality of feature vectors for text classification. In this paper, we propose a similarity-based self-constructing algorithm for feature clustering with the help of K-Means strategy. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words that are similar to each other are grouped into the same cluster, and make a head to each cluster data sets.
By the FAST algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experimental results show that our FAST algorithm implementation can run faster and obtain better-extracted features than other methods.

Keywords- Feature subset selection, filter method, feature clustering, graph-based clustering

Introduction

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories.

Traditional machine learning algorithms like decisión trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selectedsubsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large.

The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods, by using a filter method to reduce search space that will be considered by the subsequent wrapper.

They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

The wrapper methods are computationally expensive and tend to overfit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms.

Pereira Baker and Dhillon employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications.

System Architecture

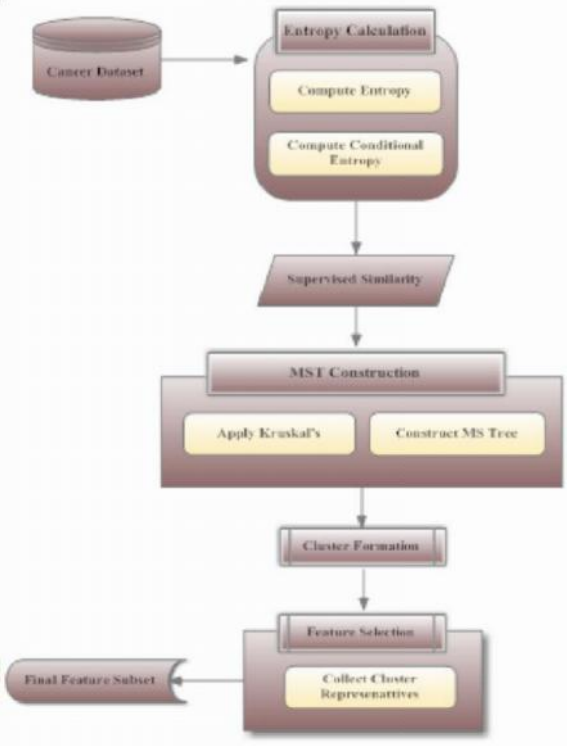


Fig 1Architecture of Proposed Method

II. CLUSTERING

Clustering and segmentation are the processes of creating a partition so that all the members of each set of the partition are similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Objects are often decomposed into an exhaustive and/or mutually exclusive set of clusters. Clustering according to similarity is a very powerful technique, the key to it being to translate some intuitive measure of similarity into a quantitative measure. When learning is unsupervised then the system has to discover its own classes i.e. the system clusters the data in the database. The system has to discover subsets of related objects in the training set and then it has to find descriptions that describe each of these subsets. There are a number of approaches for forming clusters. One approach is to form rules which dictate membership in the same group based on the level of similarity between members. Another approach is to build set functions that measure some property of partitions as functions of some parameter of the partition.

III. FEATURE SELECTION

It is widely recognized that a large number of features can adversely affect the performance of inductive learning algorithms, and clustering is not an exception. However, while there exists a large body of literature devoted to this problem for supervised learning task, feature selection for clustering has been rarely addressed. The problem appears to be a difficult one given that it inherits all the uncertainties that surround this type of inductive learning. Particularly, that there is not a single performance measure widely accepted for this task and the lack of supervision available.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). The archetypal case is the use of feature selection in analyzing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples. Feature selection techniques provide three main benefits when constructing predictive models

- Improved model interpretability,
- Shorter training times,

Enhanced generalization by reducing over fitting.

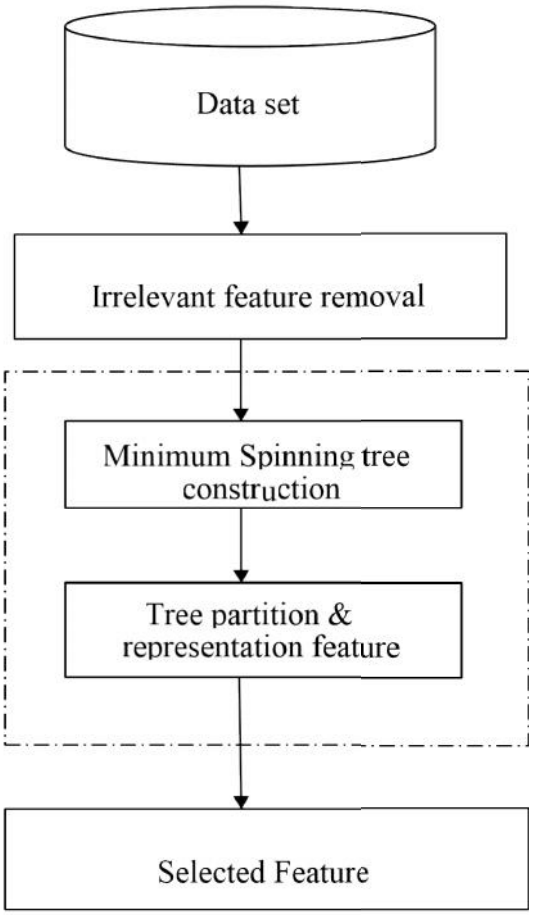
Related Works

The process of identifying and removing the irrelevant and redundant features is possible in feature is possible in feature subset selection. Due to 1) irrelevant features do not participate to the expected accuracy and 2) redundant features getting information which is already present.

Many feature subset selection algorithm can effectively removes irrelevant features but does not handle on redundant features. But our proposed FAST algorithm can remove irrelevant features by taking care of the redundant features.

In earlier days, feature subset selection has concentrate on finding for relevant features. Relief is a good example for it. But Relief is ineffective at finding redundant features. Later, Relief is extended into Relief-F to deal with noisy and incomplete data sets but it still cannot identify redundant features. CFS, FCBF, and CMIM are examples considering redundant features. FCBF is a fast filtering method that finds relevant features as well as redundancy among it. Differing from these algorithms, our proposed FAST algorithm uses the clustering-based method to choose features. It uses MST method to cluster features.

Feature Subset Selection Algorithm redundant information as possible. Because irrelevant and redundant features severely affect the accuracy of the learning machines. So we develop a novel algorithm to deal with both irrelevant and redundant features. Finally, it will obtain a good feature subset.



CONCLUSION AND FUTURE WORK

In this Project present a FAST clusteringbased feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed reduced. The text data from the four different aspects of the proportion of selected features, run time, classification accuracy of a given classifier.Clusteringbased feature subset selection algorithm for high dimensional data. For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space. In feature we are going to classify the high dimensional data.

REFERENCES

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 3845,1992.

[2] Bell D.A. and Wang, H., A formalism for relevance and its application infeature subset selection, Machine Learning, 41(2), pp 175195, 2000.

[3] Biesiada J. and Duch W., Features election for highdimensionaldatałaPearsonredundancy based filter, Advances inSoftComputing, 45, pp 242C249,2008.

[4] Dash M., Liu H. and Motoda H., Consistency based feature Selection, InProceedings of the Fourth Pacific Asia Conference on Knowledge Discoveryand Data Mining, pp 98-109, 2000.

[5] Das S., Filters, wrappers and a boostingbased hybrid for feature Selection,InProceedings of the Eighteenth International Conference on MachineLearning, pp 74-81, 2001.

[6] Dash M. and Liu H., Consistency-based search in feature selection. Artificial Intelligence, 151(1-2), pp 155-176, 2003.

[7] Demsar J., Statistical comparison of classifiers over multiple data sets, J.Mach. Learn. Res., 7, pp 1-30, 2006.

[8] Fleuret F., Fast binary feature selection with conditional mutual

Information,Journal of Machine Learning Research, 5, pp 15311555, 2004.

[9] Forman G., An extensive empirical study of feature selection metrics fortext classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.

[10] Garcia S and Herrera F., An extension on "Statistical Comparisons ofClassifiers over Multiple Data Sets" for all pairwise comparisons, J. Mach.Learn. Res., 9, pp 2677-2694, 2008.

[11] Guyon I. and Elisseeff A., An introduction to variable and feature selection,Journal of Machine Learning Research, 3, pp 11571182, 2003.

[12] Hall M.A., Correlation-Based Feature Selection for Discrete and NumericClass Machine Learning, In Proceedings of 17th International Conferenceon Machine Learning, pp 359-366, 2000.

AUTHORS PROFILE:

MATTA RAMBABU is a student of Computer Science Engineering from G.V.R&S college of Engineering & Technology Presently pursuing M.Tech (CSE) from this college. He received MCA from ANU in the year of 2011

M AYYAPPA CHAKRAVARTHI is a Associate Professor at G.V.R&S college of Engineering & Technology, Guntur. He received M.Tech in Computer science engineering from ANU. He gained 10 years of Experience on Teaching. He is a good Researcher in Compiler Design. He attended Various National and International Workshops and Conferences.