IJETST- Vol.||02||Issue||03||Pages 1997-2001||March||ISSN 2348-9480 2015

Open access Journal International Journal of Emerging Trends in Science and Technology

Relevant Feature Selection from High-Dimensional Data Using MST Based Clustering

Authors

Yaswanth Kumar Alapati[1], K. Sindhu[2], S. Suneel[3]

[1]Assistant Professor, Dept. of Information Technology, R.V.R. & J.C. College of Engineering, Guntur, A.P
Email: *alapatimail@gmail.com*

[2]Assistant Professor, Department of CSE, R.V.R. & J.C. College of Engineering, Guntur, A.P
Email: *sindhu926@yahoo.in*

[3]Assistant Professor, Department of CSE, PRRM College of Engineering, Shahabad, Telangana
Email: *sajja.suneel@gmail.com*

Abstract

Feature selection is the process of identifying a subset of the most useful features that produces compatible results as the original entire set of features. Features provide the information about the data set. In High-dimensional data representation each sample is described by many features. The data sets are typically not task-specific, many features are irrelevant or redundant and should be pruned out or filtered for the purpose of classifying target objects. Given a set of features the feature selection problem is to find a subset of features that "maximizes the learner's ability to classify patterns". A graph theoretic clustering algorithm based on boruvka's algorithm is implemented and experimentally evaluated in this paper. The proposed algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. All the representative features from different clusters form the final feature subset. After finding feature subset accuracy of a classifier, time required for classification and proportion of features selected can be calculated.

Keywords: *Boruvka's Algorithm, Graph theoretic clustering, Filter Method, Wrapper Method, Embedded Approach*

1. Introduction

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. Feature selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving result comprehensibility. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).In High dimensional data each sample is defined by more number of measurements. High-Dimensional data slow down the mining process and reduce the accuracy. The problems with High-Dimensional data are curse of dimensionality and relative shortage of instances.

The objective of Feature Subset Selection is identifying a subset of the most useful features that produces compatible results as the original entire set of features. The best subset contains the least number of dimensions (features) that most contribute to accuracy so that the problem of curse of dimensionality can be avoided. The main idea of feature subset selection is to choose a subset of

features by eliminating features with little or no predictive information.

In this paper a graph theoretic clustering algorithm based on boruvka's algorithm is used because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice.

The rest of the paper is organized as follows: in section 2, we describe the related work. In section 3, we present the new feature subset selection algorithm based on boruvka's algorithm. In section 4, we report the experimental setup. In section 5, we report the experimental results. Finally, in section 6, we summarize the present study and draw some conclusion.

2. Related Work

Feature selection is the process of removing both irrelevant and redundant features. The feature subset selection algorithms are divided into four categories: Embedded Approach, Filter Approach, Wrapper Approach and Hybrid Approach. In embedded feature selection, feature selection is a part of the induction algorithm itself. An induction algorithm is a learning algorithm used to capture the concept behind the examples. Embedded algorithms [6] require lots of data to effectively select an appropriate subset. In filter [1] techniques subset selection is a process that is applied prior to induction. The selected subset serves as an input to the induction algorithm. Feature relevance score is calculated, and low-scoring features are removed.

In wrapper approach [7] a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering the filter approach tailored to a specific classification algorithm. Relief algorithm is an example of filter approach for attribute selection. Initially feature subset selection algorithms are proposed to find the relevant features. Relief [2] is an algorithm designed to find the relevant features.

Relief algorithm assigns weight to each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. Relief uses a statistical method to select the relevant features. It is a feature weight-based algorithm inspired by instance-based learning algorithms. Relief algorithm assigns relevance score to each feature. The formula used for finding the relevance score is

$$W = W - \text{diff}(X, \text{NearHit})^2 + \text{diff}(X, \text{NearMiss})^2$$

Relief Algorithm does not work with redundant features. Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted.

Along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms. Correlation-based Feature Selection (CFS) is an algorithm which avoids both irrelevant and redundant features. CFS [3] quickly identifies and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features. CFS is a fully automatic algorithm it does not require the user to specify any thresholds or the number of features to be selected. When features depend strongly on others given the class, CFS can fail to select all the relevant features.

3. Proposed Method

The proposed algorithm has two components i) Irrelevant Feature Removal ii) Redundant Feature Elimination. The proposed technique involves the following modules 1) Discretization 2) Irrelevant Feature Removal 3) Construction of Minimum Spanning Tree using boruvka's algorithm 4)Minimum Spanning Tree Clustering and Cluster Representative Identification.

3.1 Discretization

If the attributes in the dataset are continuous valued attributes then apply the MDL Discretization [4] technique.

3.2 Irrelevant Feature Removal

For a data set D with m features F = { $F_1, F_2, \ldots, F_m$ } and class C, compute the T-Relevance SU(F_i,C) value for each feature F_i ($1 \le i \le m$). The symmetric uncertainty is defined as

$$SU(X, Y) = \frac{2 \times Gain\left(X/Y\right)}{H(X) + H(Y)}$$

Where,

- H(X) is the entropy of a discrete random variable X. suppose p(x) is the prior probabilities for all values of X, H(X) is defined by

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Gain(X/Y) is the amount by which the entropy of Y decreases.

$$Gain(X/Y) = H(X) - H(X/Y)$$
$$= H(Y) - H(Y/X)$$

$$H(X/Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x/y) \log_2 p(x/y)$$

The relevance between the feature $F_i \in F$ and the target concept C is referred to as the T-Relevance of F_i and C, and denoted by SU(F_i,C).If SU(F_i,C) is greater than a predefined threshold θ, then F_i is a strong T-Relevance feature. The features whose T-Relevance is less than the predefined threshold are considered as irrelevant features and are removed.

3.3 Construction of Minimum Spanning Tree

After finding relevant features, a completely connected weighted graph G=(V,E) is constructed from the relevant features by considering relevant features as vertices and the F-correlation value as the weight of the edge. The correlation between any two features F_i and F_j is called F-Correlation of F_i and F_j and denoted by SU (F_i, Fj).The complete graph G represents the correlations among the target relevant features. From the graph G a minimum spanning tree is constructed using boruvka's algorithm

3.3.1Boruvka's Algorithm

The Boruvka's algorithm [5] is based on merging of disjoint components. At the beginning of the procedure each vertex is considered as a separate component. In each step the algorithm connects every component with some other using strictly the cheapest outgoing edge of the given component. Using Boruvka's algorithm the minimum spanning tree can be constructed in parallel because the choice of the cheapest outgoing edge for each component is completely independent of the choices made by other components.

3.4 Minimum Spanning Tree Clustering and Representative Identification

After constructing the minimum spanning tree the tree is partitioned into clusters by removing the inconsistent edges. The features in each cluster are redundant so for each cluster select the representative feature. Example for partitioning is shown in Figure1

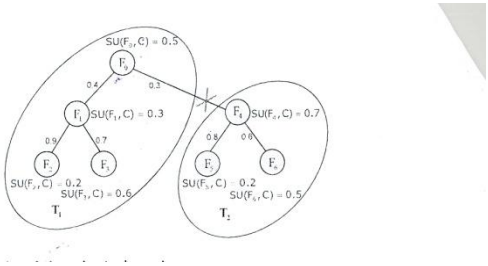

Figure 1: Example of the Clustering Step

3.4.1 Algorithm for Inconsistent Edge Removal and Cluster Representative Identification

Input: Minimum spanning tree

Output: Selected feature subset

Steps:

1. Partition the Minimum Spanning Tree into clusters by Removing the edges $E = \{(F_i', F_j') \mid (F_i', F_j' \in F' \wedge i, j \in [1,k] \wedge i \neq j)\}$, whose weights are smaller than both of the T-Relevance $SU(F_i', C)$ and $SU(F_j', C)$

2. For each cluster select a representative feature F_R^j whose T-Relevance $SU(F_R^j, C)$ is the greatest

   The framework of the proposed method is shown in Figure 2

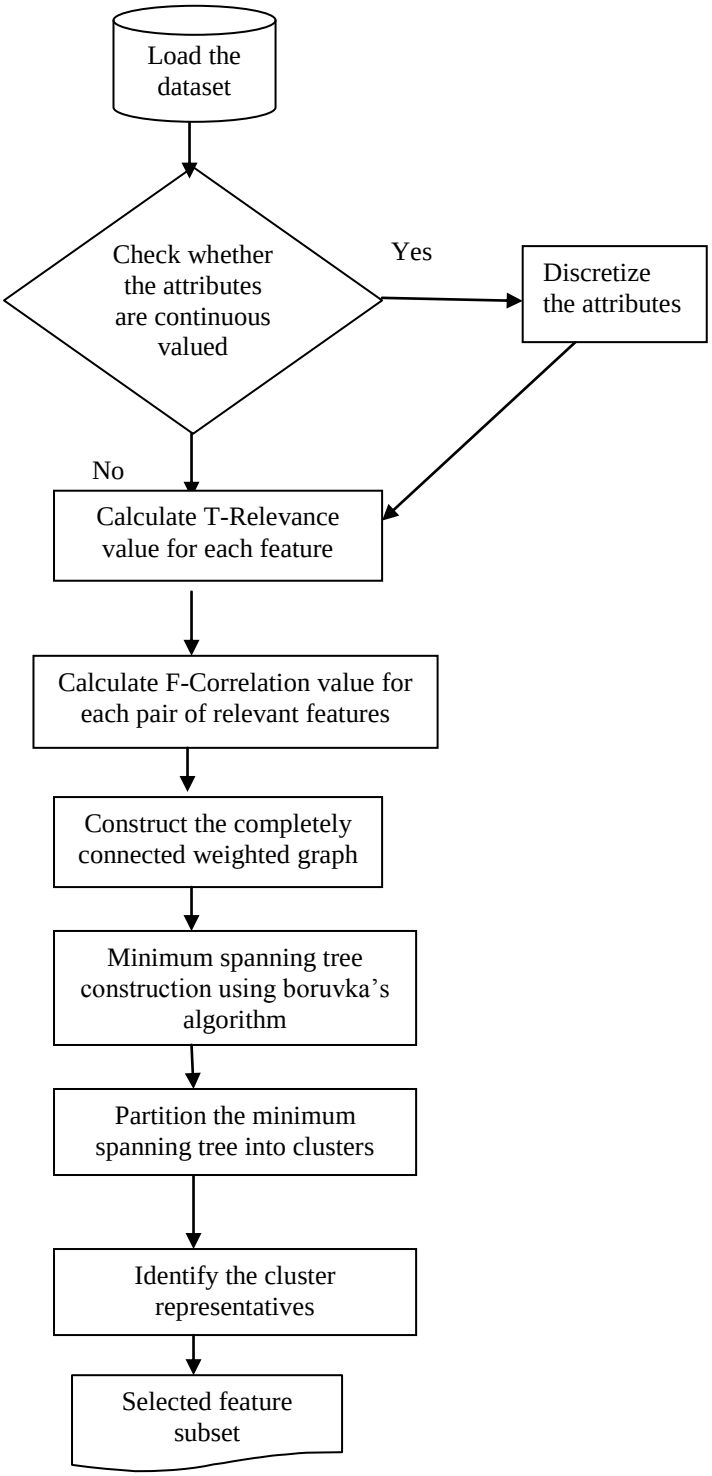

Figure 2: Framework of the proposed feature subset selection algorithm

4. Experimental Setup

A number of experiments on benchmark data sets [8] have been conducted to verify the strength of the proposed approach. A summary of the data sets is presented in Table 1. We have used 10-fold cross validation for reporting the classification results for all the data sets.

Table 1: Data Sets used

Data Set ID	*Data Set Name*	*Number of Features*	*Number of Instances*	*Number of Classes*	*Domain*
1	chess	37	3196	2	Text
2	mfeat-fourier	77	2000	10	Image, Face
3	coil2000	86	9822	2	Text
4	elephant	232	1391	2	Microarray, Bio

All the experiments were conducted on MATLAB R2010b.

5. Experimental Results

To evaluate the performance and effectiveness of proposed algorithm, three different metrics are used i) proportion of selected features ii) Classification Accuracy and iii) Runtime

5.1 Proportion of selected features

The proportion of selected features is the ratio of the number of features selected by a feature selection algorithm to the original number of features of a data set. The proportion of selected features is given in Table 2

Table 2: Proportion of Features Selected

Data Set	*Proportion of Selected Features*
chess	16.22
mfeat-fourier	16.88
coil2000	4.65
elephant	0.86

5.2 Classification Accuracy

Accuracy is defined as the ratio of the number of instances for which the outcome is correct to the total number of tests made. Here the Naive Bayes classifier is used. Accuracy of Naive Bayes classifier is given in Table 3

Table 3: Classification Accuracy

Data Set	*Accuracy of Naive Bayes classifier with original data set (%)*	*Accuracy of Naive Bayes classifier with reduced data set(%)*
chess	87.89	92.92
mfeat-fourier	76.25	79.30
coil2000	78.71	93.90
elephant	82.34	99.14

5.3 Runtime

Table 4: Runtime

Data Set	*Time Taken to Build Classifier with Original Data Set(in ms)*	*Time Taken to Build Classifier with Reduced Data Set (in ms)*
chess	100	80
mfeat-fourier	772	91
coil2000	896	151
elephant	1201	171

6. Conclusions

A novel clustering-based feature subset selection algorithm for high-dimensional data is presented. We have calculated the accuracy of a classifier, time required for classification and proportion of features selected can be calculated.

References

1. H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection: A Filter Solution," Proc. 13th Int'l Conf. Machine Learning, pp. 319-327, 1996
2. K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," Proc. 10th Nat'l Conf. Artificial Intelligence, pp. 129-134, 1992
3. M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.
4. M.F. Usama and B. Keki, "Irani: Multi-Interval Discretization of Continuousvalued Attributes for Classification Learning," Proc. 13th Int'l Joint Conf. Artificial Intelligence, pp. 1022-1027, 1993
5. Jinna Lei, Three minimum spanning tree algorithms, University of California, Berkeley, May 2010
6. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol 3, pp. 1157-1182, 2003.
7. M. Dash and H. Liu, "Feature Selection for Classification,"Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.
8. The data sets can be downloaded at: http://archive.ics.uci.edu/ml/,http://tunedit.org/repo/Data/Text-wc,http://featureselection.asu.edu/datasets.php, http://www.lsi.us.es/aguilar/datasets.