International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Fast for Feature Subset Selection Over Dataset

Jesna Jose[1], Reeba R[2]

Department of Computer Science & Engineering, Sree Buddha College of Engineering, Alappuzha, Kerala, India

Abstract: ***Feature selection is the process of identifying the most suitable features that is compatible with the target set features and thereby reducing feature space to an optimal minimum. The feature selection algorithm can be evaluated on the basis of two criteria: efficiency and effectiveness. Efficiency is measured on the basis of time required to find the feature set and effectiveness measures the quality of the feature. In fact feature selection, as a preprocessing step which is effective in reducing dimensionality, removing irrelevant data, removing redundant data etc. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. In this paper various feature selection methods are depicted and proposes a new clustering based feature subset selection algorithm for feature selection.***

Keywords: Feature clustering, feature subset selection

1. Introduction

Feature selection has been an active and fruitful field of research and development since 1970's and it is proven to be effective in removing irrelevant and redundant features, increasing efficiency in learning tasks etc. Let G be some subset of F and fG be the value vector of G. In general, the goal of feature selection can be formalized as selecting a minimum subset G such that **P**(C | G= fG) is equal or as close as possible to **P**(C | F = f), where **P**(C | G = fG) is the probability distribution of different classes given the feature values in G and **P**(C | F = f) is the original distribution given the feature values in F (Koller and Sahami, 1996).

Different feature selection methods can be broadly categorized into the wrapper model (Kohavi and John, 1997; Kim et al., 2000) and the filter model (Liu and Setiono, 1996; Liu et al., 2002b; Hall, 2000; Yu and Liu, 2003). The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. These methods are computationally expensive for data with a large number of features (Kohavi and John, 1997). The filter model separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm. It relies on various measures of the general characteristics of the training data such as distance, information, dependency, and consistency (Liu and Motoda, 1998).

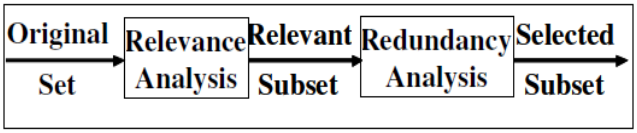
With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira et al. [8], Baker et al. [5], and Dhillon et al. [6] employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance [7]. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graph theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, 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. Based on the MST method, this paper proposes a FAST clustering-bAsed feature Selection algoriThm (FAST).

2. Related Works

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because: (i) irrelevant features do not contribute to the predictive accuracy, and (ii) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features, yet some of others can eliminate the irrelevant while taking care of the redundant features.

Lei Yu and et al[1] proposed a correlation based filter solution for feature extraction. This paper, proposes a novel concept of predominant correlation, introduce an efficient way of analyzing feature redundancy, and design a fast correlation based filter approach. A new feature selection algorithm FCBF is implemented and evaluated through extensive experiments. The algorithm consist of two parts, Select relevant features and arrange them in descending order according to the correlation value is the first step. Then remove redundant features and only keeps predominant ones. For predominant feature selection, Take the first element Fp as the predominant feature. Then take the next element Fq. If Fp happens to be redundant peer of Fq, remove Fq After one round of filtering based on Fp, take the remaining features next to Fp as the new reference and repeat. The algorithms stop until there is no more feature to be removed. The disadvantage was it does not work with high dimensional data.

Efficient Feature Selection via Analysis of Relevance and Redundancy 2] proposed in 2004 classified features on the basis of relevance and redundancy. The framework is shown in figure.



It classified the features on relevance basis as whether is of category strong relevance, weak relevant or irrelevant. Strong relevance of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all.

Graph based clustering method [3] summarized graph based clustering into five parts. The five-part story describes the general methodology of graph-based clustering: (1) Hypothesis: A graph can be partitioned into densely connected subgraphs that are sparsely connected to each other.(2) Modelling: It deals with the problem of transforming data into a graph or modelling the real application as a graph. (3) Measure: A quality measure is an objective function that rates the quality of a clustering.(4) Algorithm :An algorithm is to exactly or approximately optimize the quality measure.(5) Evaluation :Various metrics can be used to evaluate the performance of clustering.

Another clustering method was minimum spanning tree clustering tree method [4]. Here the methodology was Generate an minimum spanning tree (MST) from spanning tree (ST). Then generating clusters using MST. The steps includes 1) Calculate mean (M) , SD of the edge weight in MST. 2) Calculate a threshold I = M+SD. 3) Remove edge having lower weight than I.4). This gives disjoint sub trees (clusters). The contributions of this paper was the concept of MST clustering. Also it takes less time and it doesn't need number of clusters as a prior parameter. But it was got a great disadvantage that it can be used only for numerical data.

3. Problem Statement

To reduce the feature space when features are presented as hundreds and thousands of data and to obtain features that matches with the given target concept by eliminating irrelevancy and redundancy in feature set. For that an algorithm known as Fast clustering-bAsed feature Selection algoriThm (FAST) is used. The algorithm works in two steps. In the first step, clustering of the available data set is made. After this step redundant datas will be available in a single cluster. In the second step, most representative features compatible with the target concept will be taken out from those clusters. The main objectives are redundant data removal and irrelevant data elimination to optimize speed.

4. System Design

Irrelevant data and duplicate data affect the accuracy of the system. So a good feature selection algorithm should be able to remove all those irregularities. The framework of the proposed algorithm is shown in the figure.

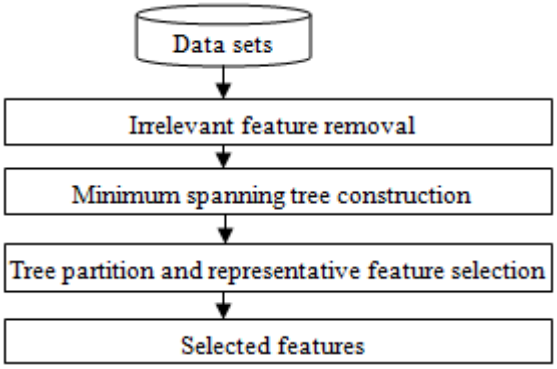


Figure 1: Design of feature subset selection algorithm

Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework (shown in Fig.1) which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm [9], it involves (i) the construction of the minimum spanning tree (MST) from a weighted complete graph; (ii) the partitioning of the MST into a forest with each tree representing a cluster; and (iii) the selection of representative features from the clusters. The methodology used here is a five step methodology. The steps are data set management, finding symmetric uncertainty, irrelevant feature removal, minimum spanning tree construction and its clustering and the final step is target matching.

5. Algorithm Implementation

Feature selection strategies aim at selecting an informative subset of features out of the complete set. Feature selection methods choose features from the original set based on some criteria like information gain, correlation and mutual information that are used to filter out unimportant or redundant features. The feature selection is done on the basis of strength of each character in the feature set.

One of the most informative measures for feature selection is mutual information (MI). In terms of MI, the optimal FE creates new features that jointly have the largest dependency on the target class. However, obtaining an accurate estimate of a high-dimensional MI as well as optimizing with respect to it is not always easy, especially when only small training sets are available. Feature selection is mainly used for high dimensional data attribute reduction for increasing computational efficiency. One of the application is that when thousands of sensors providing data at the same time, to know which sensor data is most important. Survey results will be having thousands of attributes. To select one or two

with most prominent attributes, feature subset selection is important.

To decide whether a feature is relevant, here the entropy will be determined. Entropy will give the data contained in the feature i.e. the strength of each variable or character in the text. Entropy between two words provides the relevance between the two terms. Also we are fixing a threshold to evaluate the selected features. Number of selected attributes will be based on the threshold value. Selected features will be plotted as a graph. Minimum spanning tree algorithm is applied to form clusters in which each cluster will have related feature values. Target text will be input by the user for searching. Symmetric uncertainty as well as entropy values of the target text will be calculated and it will be checked with entropy of data elements in each feature column for matching. Significant features will be selected and clustered. The algorithm used here for feature selection is Fast clustering-bAsed feature Selection algoriThm (FAST). The proposed FAST algorithm logically consists of three steps:

- Removing irrelevant features.
- Constructing a minimum spanning tree from relative ones.
- Partitioning the minimum spanning tree and selecting representative features.

The algorithm is given below.

Inputs : D(F1, F2….Fm,C) - the given set
θ : the T-Relevance threshold.
Output S : selected feature subset

// Part 1 : Irrelevant Feature Removal
1 for i = 1 to m do
2 T- Relevance = SU(Fi,C)
3 if T - Relevance > θ then
4 S = S ∪ {Fi};
// Part 2 : Minimum Spanning Tree Construction
5 G = NULL; //G is a complete graph
6 for each pair of features $\{\{F'i, F'j\} \subset S$ do
7 F-Correlation = SU ($F'i$, $F'j$)
8 Add F' i and/*or* $F'j$ to G with F-Correlation as the weight of the corresponding edge;
9 minSpanTree = Prim (G); //Using Prim Algorithm to Generate minimum spanning tree

// Part 3: Tree Partition and Representative Feature Selection

10 Forest = minSpanT ree
11 for each edge $Eij \in$ Forest do
12 if SU($F'i, F'j$) < SU($F'i, C$) ∧ SU($F'i, F'j$) < SU($F'j, C$) then
13 Forest = Forest – Eij

14 S = ϕ
15 for each tree $Ti \in$ Forest do
16 FR = feature with maximum value of symmetric uncertainity SU(F' k, C)
17 S = S ∪ {FR};
18 return S

The working of the algorithm is described below. The first step is the data set management. In this data set about the corresponding application can be downloaded and kept. For example, in a car data set all features about the car can be downloaded from the internet. Some of the features in the data set are buying, maintenance, doors, persons, acceptability etc. Each of these features has sub attributes. For the feature buying, sub attributes are buying rate like high, low, medium etc and likewise for each of feature in the feature set.

Symmetric uncertainty can be found out in the second step. Symmetric Uncertainty (SU) is defined as the measure of correlation between either two features or a feature and the target concept. Correlation is a well known similarity measure between two random variables. If two random variables are linearly dependent, then their correlation coefficient is close to 1. If the variables are uncorrelated the correlation coefficient is 0. The correlation coefficient is invariant to scaling and translation. Hence two features with different variances may have same value of this measure. Symmetric uncertainty can be defined as

$$SU(X,Y) = \frac{2*Gain(X|Y)}{H(X)+H(Y)}$$

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. It reflects the additional information about Y provided by X and is called the information gain which is given by

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

Where H (X|Y) is the conditional entropy which quantifies the remaining entropy (i.e. uncertainty) of a random variable X given that the value of another random variable Y is known. Suppose p(x) is the prior probabilities for all values of X and p(x|y) is the posterior probabilities of X given the values of Y , H(X|Y) is defined by

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

Irrelevant feature removal is the next step which removes those features that does not contribute anything to the predictive accuracy. The relevance between a feature and the target concept is referred to as T-relevance. Here T-relevance of each feature is finding out and compares it with the threshold value we already calculated. If it is greater than the threshold that feature is relevant.

Next step is minimum spanning tree creation and clustering. For this correlation value between every pair of features is

calculated and drawn the graph. The vertices of the graph are features and edge weight is the correlation value between feature and target concept. The vertex value will be the correlation between that vertex feature and the target feature and edge value represents the correlation value between the vertices joining the edge. Clustering of the feature set will be carried out in the next step. Clusters are created by removing those edges having edge weight less than both its vertices value. Thereby distinct clusters will be obtained and each cluster contains similar features. Then the Target matching which means most representative features that is strongly related to target classes is selected from each cluster to form subset of features.

6. Conclusion

Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. Existing feature selection methods mainly focus on finding relevant features.

FAST is a novel clustering-based feature subset selection algorithm for high dimensional data. Clustering is the process of grouping the data objects into classes or clusters so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. The algorithm involves (i) removing irrelevant features as irrelevant features have no or weak correlation with target concept, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features, i.e. redundant data elimination . In the proposed algorithm, each cluster consists of set of features. Redundant features are assembled in a cluster and most representative feature can be taken out of the cluster. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

7. Future Work

As future work, to find correlation value between two features or a feature and the target concept, this work can be extended by making use of Pearson co-efficient

References

[1] Yu L. and Liu H., “Feature selection for high-dimensional data: a fast correlation-based filter solution”, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.
[2] Yu L. and Liu H.,” Efficient feature selection via analysis of relevance and redundancy”, Journal of Machine Learning Research, 10(5), pp 1205-1224, 2004.
[3] Zheng Chen and Heng Ji, “Graph-Based Clustering for Computational Linguistics: A Survey”, Proceedings of the 2010 Workshop on Graph-based Methods for NLP, ACL 2010.
[4] Bhaskar Adepu and Kiran Kumar Bejjanki., “A Novel Approach for Minimum Spanning Tree based Clustering Algorithm”.
[5] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.
[6] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3, pp 1265-1287, 2003.
[7] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and Their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517, 1992.
[8] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.
[9] Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast clustering based feature subset selection algorithm for high dimensional data, In proceedings of the IEEE Transactions n Knowledge and data engineering, 2013