IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. I (Sep – Oct. 2014), PP 108-111
www.iosrjournals.org

Feature subset selection for high dimensional data with domain analysis using Semantic Mining

Abdul Majeed K. M, Pallavi K.N, Tanvir Habib Sardar
[1,3]*Dept of ISE, PACE Mangalore,* [2]*Dept of CSE, NMAMIT Nitte*

Abstract: *Feature subset selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Current existing algorithms for feature sub set selection works only based on conducting statistical test like Pearson test or symmetric uncertainty test to find the correlation between the features and apply threshold to filter redundant and irrelevant features. FAST proposed by Qinbao Song [9] uses symmetric uncertainty test for feature subset selection. In this work we extend the FAST algorithm by applying the domain analysis using semantic Mining to improve the relevance of the feature subset selection.*

Index Terms: *Feature subset selection, filter method, feature clustering, graph-based clustering*

I. Introduction

Feature selection is typically a search problem for finding an optimal or suboptimal subset of m features out of original M features. Feature selection is important in many pattern recognition problems for excluding irrelevant and redundant features. It allows reducing system complexity and processing time and often improves the recognition accuracy. 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 decision 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 selected subsets, 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 over fit 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. Thus, we will focus on the filter method in this paper.

II. Literature Survey

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.

Traditionally, feature subset selection research has focused on searching for relevant features. A well known example is Relief [1], which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted [2]. Relief-F [3] extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class

problems, but still cannot identify redundant features.FCBF [4] is a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis.

CMIM [5] iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked

Butterworth et al. [6] proposed to cluster features using a special metric of Barthelemy-Montjardet distance, and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on Barthelemy-Montjardet distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Further more, even compared with other feature selection methods, the obtained accuracy is lower

Hierarchical clustering also has been used to select features on spectral data. Van Dijk and Van Hullefor [7] proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. [8] presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Their feature clustering method is similar to that of Van Dijk and Van Hullefor [7] except that the former forces every cluster to contain consecutive features only. Both methods employed agglomerative hierarchical clustering to remove redundant features.

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. 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 . 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. Qinbao Song proposed FAST [9] which is a minimum spanning tree based graph theoretic approach. The FAST 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 the final subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features.

III. Proposed Solution

3.1 Problem Statement

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines from the data sets. Feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible from the dataset.

3.2 Existing Solution

FAST is the existing solution shown in Fig 3.1

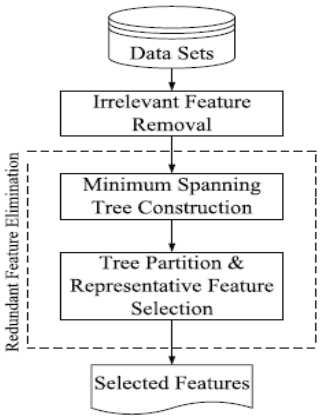

Fig 3.1 Steps for FAST algorihm

The FAST algorithm works in three steps.

- In the first step, the irrelevant features are removed
- In the second step, it construct the minimum spanning tree (MST) from a weighted complete graph
- In the third step, it will do the partitioning of the MST into a forest with each tree representing a cluster; and.the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features

3.3 Proposed Solution

The existing solution "FAST "has one problem in the step 1 – irrelevant feature removal. The existing solution for step 1 is based only on entropy [deviation in the values across the dataset]. But entropy measure is not always accurate in identifying the irrelevant feature removal. The feature relevancy to the clustering depends on the domain area of classification. We propose to extract the semantic relation between the domain area of application & the features. Based on the domain analysis, the irrelevant features will be identified & removed in the Step 1. This will result in highly relevant features subsets. Semantic analysis will be drawn based on web mining the domain area concept. We will propose a algorithm to mine the web & provide the relation of any features to the domain area concept.

For irrelevant feature removal we use domain analysis with semantic mining. For identifying the redundant features & removing it we use MST based clustering technique

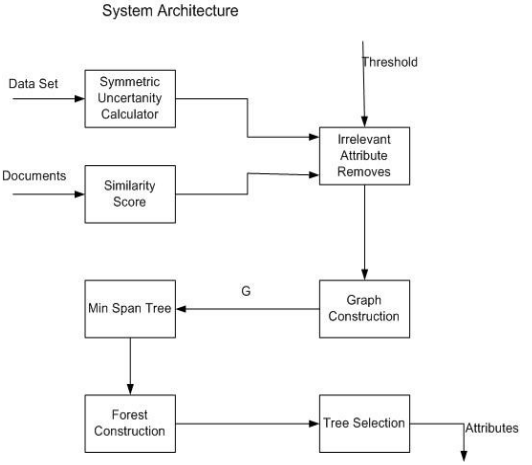

Fig 3.2 architecture for the proposed algorithm (semantic mining)

3.3.1Removing irrelevant Features

In our proposed algorithm The feature relevancy to the clustering depends on the domain area of classification. We propose to extract the semantic relation between the domain area of application & the features. Based on the domain analysis, the irrelevant features will be identified & removed in the Step 1. This will result in highly relevant features subsets.

The following flowchart explains the steps for the proposed algorithm. The user needs to upload the documents for mining and the attribute list for feature selection. Then the documents will be mined one by one for identifying and selecting the terms in the documents. A term is a selected word or item from the documents after mining. It also calculates the frequency of each term in the documents and if the frequencies are greater than 2 then it will be stored in a vector. Then the each attribute will be compared with the term one by one. If the attribute is matching with any of the term in the vector with a given threshold that attribute will be selected as feature. This process will be repeated until the entire attribute in the attribute list will be checked and compared with the list of terms.

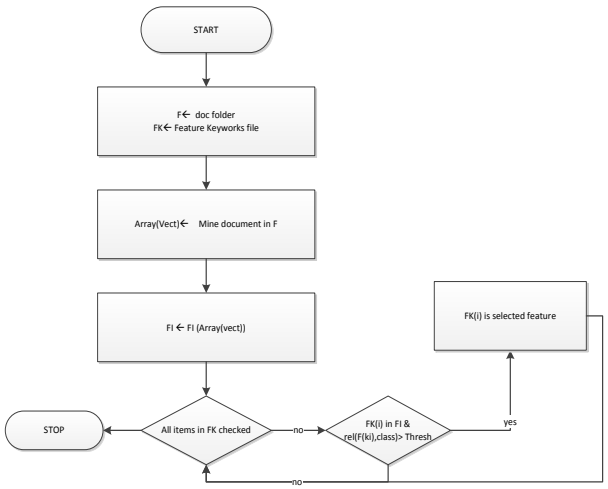

Fig 3.3 flowchart for removing irrelevant features

3.3.2Removing redundant Features

Once the irrelevant features are removed, we start to find the redundant features and remove those features. In our algorithm, 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.

IV. Conclusion and Future Work

In this paper, we have detailed our solution for feature subset selection. Our mechanism effectively removes the irrelevant & redundant features from the feature set. Our proposed solution is different from the FAST algorithm [9] in following way.

FAST uses the entropy measure for relevancy. But in our solution we have proposed domain analysis with semantic mining.

As the future work, we have planned to compare the performance of the proposed algorithm with that of the FAST algorithm on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record.

.

References

[1]. Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.

[2]. Koller D. and Sahami M., Toward optimal feature selection, In Proceedings of International Conference on Machine Learning, pp 284-292, 1996.

[3]. Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.

[4]. 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.

[5]. Fleuret F., Fast binary feature selection with conditional mutual Information, Journal of Machine Learning Research, 5, pp 1531-1555, 2004.

[6]. Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.

[7]. Van Dijk G. and Van Hulle M.M., Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, International Conference on Artificial Neural Networks, 2006

[8]. Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162, 2007.

[9]. A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data - QinBao Song in "IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1 YEAR 2013"