Feature Subset Selection for High Dimensional Data

Pavan Mallya P
Department of Information Science and Engineering
Sri Jayachamarajendra College Of Engineering
Mysore, India

Professor Roopa C. K
Department of Information Science and Engineering
Sri Jayachamarajendra College Of Engineering
Mysore, India

Abstract— **This paper considers feature selection for data classification in the presence of a huge number of irrelevant and redundant features. We propose a new feature selection algorithm that addresses several major issues with prior work, including problems with algorithm implementation, computational complexity, and solution accuracy. The key idea is to decompose an arbitrarily complex nonlinear problem into a set of locally linear ones through local learning, and then learn feature relevance globally within the large margin framework. It is capable of processing many thousands of features within minutes on a personal computer while maintaining a very high accuracy that is nearly insensitive to a growing number of irrelevant features.**

Keywords— Feature selection; machine learning; symmetric uncertainty; irrelevance; redundancy; conditional mutual information

I. INTRODUCTION

There is vast amount of data in the present day world and hence the storage, analysis and maintenance of this data becomes a tedious job. So our machine learning [1] algorithm also has to keep pace with this growth so as to enhance the ability to understand and make use of the existing data. Machine learning provides tools by which large quantities of data can be automatically analyzed. Fundamental approach to machine learning is feature selection. Feature selection [3, 4] is done by identifying the most salient features for learning. This approach focuses on developing a learning algorithm to obtain those aspects of the data which is most useful for analysis and can be further used for future prediction. A feature is considered as good and thus will be selected only if it is found to be of significant relevance and the relevance value is greater than a given threshold value. A typical application of machine learning algorithm requires two examples: training examples and test examples. Training examples are used to produce the concept descriptions and test examples are needed to evaluate the accuracy. When testing, the class labels are not presented to the algorithm. The algorithm takes test example as input and produces a class label as output.

II. PROBLEM FORMULATION

Feature selection is most frequently used as a preprocessing step to machine learning. It is basically the process of choosing subsets from original features so that the feature space is optimally reduced, which follows a certain evaluation criterion and these selected features can be used in machine learning. This method has been proven effective in removing both irrelevant and redundant features and hence increasing the efficiency in learning tasks, improving performance for predictive accuracy and enhancing comprehensibility of observed results. Nowadays, the amount of information is very high. This increase has started to pose serious threat to machine learning in terms of scalability and accuracy. For example, high dimensional data i.e. data to a level of hundreds of thousands, may contain high amount of irrelevant and redundant information leading to degradation of performance of learning algorithms. This results in the need to bring methodology of feature subset selection for dimensionality reduction in this era. Relevant features have values which are dependent on the values of any other features and values of the corresponding class, and provide further information about the given class. Whereas, redundant features, are those whose values are dependent on values of other features which are irrespective of the class and provides no extra information about the given class.

III. FEATURE SUBSET SELECTION ALGORITHMS

Feature subset selection algorithms can be grouped into two broad categories: Wrapper and Filter [2]. Wrapper method evaluation function uses error rate of classification algorithm to evaluate a feature subset, while filter method evaluation function is independent of the classification algorithm. Filter does not consider a classifier (formulates a hypothesis) as it is not dedicated to any specific type of classification method. On the contrary, wrapper relies on the performance of one particular type of classifier which is used to evaluate the quality of a subset of features. Accuracy of wrapper is comparatively high; however, the generality of the observed result is limited and its computational complexity is high. In comparison, filter is more general and its computational complexity is low. Due to the fact that wrapper is computationally expensive, filter is usually a very good choice when number of features is large. Hence we focus on filter method in our paper.

Both irrelevant and redundant features affect speed and accuracy of machine learning algorithms and thus both should be eliminated. Therefore, pure relevance-based feature weighting algorithms do not entirely meet the need of feature selection. There is also a need for removal of redundant features as well in the context of feature selection for high dimensional data.

IV. CORRELATION BASED MEASURES

Here, we discuss how to evaluate the importance of features for classification. In general, a feature is good if it is relevant to the class but is not redundant compared to any other relevant features. If correlation between a feature and class is high enough to make it relevant to the class and correlation between feature and any other relevant features does not reach a level where it can be predicted by any other relevant features. Then, it will be regarded as a good feature for task of classification. The problem of feature selection digs down to find suitable measure of correlation among features and uses a procedure to select features based on the calculated measure.

In Information Theory, we consider entropy as the degree of uncertainty in a random variable. For example H(X) is Entropy of random variable X. Information gain is the amount by which the entropy of Y decreases. It reflects additional information about Y provided by X. It is given by

$$IG(X|Y) = H(X) - H(X|Y) \quad \text{eqn.1}$$

If IG(X|Y) > IG(Z|Y), then Y is considered to be highly correlated with feature X than feature Z.

Symmetry is generally a desired property used to measure correlations between features. However, information gain is biased towards features with more values. In addition, all values have to be normalized before they can be used for comparison between features. Therefore, symmetrical uncertainty can be defined as follows:

$$SU(X, Y) = 2\,[\,IG(X|Y) / (\,H(X) + H(Y)\,)\,]$$

It can be used to compensate for information gain's bias towards different features with more values. It normalizes values to a range [0, 1], where 1 indicates that knowing one variable completely predicts value of the other and 0 indicates that both variables are completely independent. It ensures that SU (X, Y) = SU (Y, X).

Development of a procedure for selecting good and relevant features for classification based on correlation based analysis of features involve two aspects:

1)Decision regarding the relevance of a feature to the class.

2)Decision whether a relevant feature is redundant or not with respect to any other relevant feature.

V. PROPOSED METHODOLOGY

A. Mutual Information:

It evaluates Mutual Information between individual features and target class labels. Then only those features are selected which have maximum correlation with target class labels and are less correlated with other relevant features. Greater the value of mutual information between feature and target class, lower is the probability of error in feature subset selection.

Mutual Information between X and Y is defined as

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

where, I(X; Y) means entropy of X and Y, H(X) is entropy of X, H(Y) is entropy of Y, H(X|Y) is entropy of X when Y is known, H(Y|X) is entropy of Y when X is known and H(X, Y) is entropy between X and Y. In information theory, '|' means '**given**', '**;**' means '**between**' and '**,**' means '**and**'.

B. Conditional Mutual Information:

Mutual information considers the correlation between two features without taking the target class into consideration. This is overcome using Conditional Mutual Information. Conditional Mutual Information provides an extension to Mutual Information. It measures correlation between two independent features, when the target class is known. It is used to evaluate inter-feature correlation within a selected subset. This helps reduce redundancy [6] among the selected features.

Conditional Mutual Information between a target X and independent variables Y and Z is given below:

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= H(X, Z) - H(Z) - H(X, Y, Z) + H(Y, Z) \end{aligned}$$

Y is considered as a good feature if and only if I (X; Y|Z) is very large for every Z already picked. This means that Y is good only if it carries more information about X, and this information should not be already given by Z.

Proposed algorithm takes dataset and threshold value as input. Relevance of each feature is calculated and a feature is selected only if the relevance of feature and target class is more than the input threshold value. A connected graph is now created using features as nodes and relevance between features as weight between nodes. Minimum spanning tree is now created to eliminate edges with maximum correlation. Now conditional mutual information is calculated between two features and the target class. If a feature is found to be redundant, the feature is eliminated. Finally we get a subset of features that are relevant and non-redundant.

VI. REQUIREMENTS

A. Functional Requirements:

1)Select most relevant features among entire dataset.

2)Do not select redundant features for creating subset.

3)Maximize correlation of one feature with target class.

4)Minimize correlation of one feature with another.

5)Selection of most appropriate representative feature for each class from the selected feature subsets.

B. Non Functional Requirements:

1) *Improve efficiency of feature subset selection process.*
2) *Enhance effectiveness of the selected features.*
3) *Improve performance for predictive accuracy.*
4) *Enhance comprehensibility of observed results.*
5) *The system has to be user friendly and reliable.*

VII. CONCLUSION

This paper provides a comprehensive overview of various aspects of feature selection. They categorize the large body of feature selection algorithms and guide the selection of algorithms for intelligent feature selection. The categorizing framework is developed from an algorithm designer's viewpoint that focuses on the technical details about the general procedures of feature selection process. A new feature selection algorithm can be incorporated into the framework according to the three dimensions. The ultimate goal for intelligent feature selection is to create an integrated system that will automatically recommend the most suitable algorithm(s) to the user while hiding all technical details irrelevant to an application.

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. In this study, we apply graph-theoretic clustering methods to features. In particular, we adopt the minimum spanning tree 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 minimum spanning tree method, we propose a clustering based feature selection algorithm. Features in different clusters are relatively independent, the clustering based strategy has a high probability of producing subset of useful and independent features.

REFERENCES

[1] T. M. Mitchell, "Generalization as Search," *Artificial Intelligence, vol. 18, no. 2, pp*. 203-226, 1982

[2] Dash, M., Choi, K., Scheuermann, P., Liu, H.: "*Feature selection for clustering-a filter solution*". In: *ICDM, pp*. 115–122 (2002).

[3] Dy, J., Brodley, C.: "*Feature selection for unsupervised learning*". *JMLR 5*, 845–889 (2004).

[4] He, X., Cai, D., Niyogi, P.: "*Laplacian score for feature selection*". In: *NIPS, pp*. 507–514 (2005).

[5] Mitra, P., Murthy, C.A., Pal, S.K.: "*Unsupervised feature selection using feature similarity*". *PAMI 24(3),* 301–312 (2002).

[6] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *J. Machine Learning Research, vol. 10, no. 5, pp*. 1205-1224, 2004.

[7] Ng, A., Jordan, M., Weiss, Y.: "*On spectral clustering: analysis and an algorithm*". In: *NIPS* (2002).