Efficient Feature Selection By Reducing Redundant Features From High Dimensional Data Using Clustering

Nandini N,PG Student
Department of Computer Science
AMC Engineering College
Bangalore, India
nandini.bms07@gmail.com

Doddegowda B J, Assoc. Prof.
Department of Computer Science
AMC Engineering College
Bangalore, India
bjdgowda10@gmail.com

Abstract-

Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, in-creasing learning accuracy, and improving result comprehensibility. However, the re-cent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. Based on this criteria Redundancy feature reduction algorithm is proposed. The RFR (Redundancy Feature Reduction) algorithms 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. Features in different clusters are relatively independent; the clustering-based strategy of RFR has a high probability of producing a subset of useful and independent features. The real-world high-dimensional image, microarray, and text data are given as input to the algorithm. To ensure efficiency we adopt the efficient minimum-spanning tree clustering method. Finally this algorithm will be able produce good set of feature subset from high dimensional data sets using clustering method.

Keywords: — RFR; Feature subset selection; correlation; graph-theoretic, cluster analysis.

I INTRODUCTION

The performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. Thus, selecting relevant features for the construction of classifiers has received a great deal of attention. 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 sub-sets, 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.

In 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 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 based clustering algorithms which is efficient.

Based on the MST method, a Redundancy Feature Reduction for High-dimensional Data using Clustering algorithm is proposed. The RFR 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 RFR has a high probability of producing a subset of useful and independent features.

II PRIOR WORK

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because 1) irrelevant features do not contribute to the predictive accuracy, and 2) 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. Our proposed RFR algorithm falls into the second group.

Feature subset selection research has focused on searching for relevant features. An example is Relief, which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

With irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well [5] [4] [3]. CFS [2] FCBF [6] and CMIM [1] are examples that take into consideration the redundant features. CFS [2] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, still uncorrelated with each other. FCBF ([6] [7]) is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. CMIM [22] iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from these algorithms, proposed RFR algorithm employs the clustering-based method to choose features.

III PROPOSED WORK

Feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Good feature subsets contain features highly correlated with the class, but uncorrelated with each other. Based on this, a novel algorithm is developed, which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. This is achieve this through a new feature selection framework 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 latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

The main objective is to improve efficiency and effectiveness of the feature selection process ,improve the classification performance and also to remove the redundant features.

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit tedious. In the proposed RFR algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

Definition 1 (T-Relevance). The relevance between the feature $F_i \varepsilon F$ 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 predetermined threshold θ then F_i is a strong T-Relevance feature.

Definition 2 (F-Correlation). The correlation between any pair of features F_i and F_j, (F_i, F_j ε F ^ i ≠ j) is called the F-Correlation of F_i and F_j, and denoted by SU(F_i, F_j).

Definition 3 (F-Redundancy). Let S = {F1, F2 . . . F_i . . . $F_{k<|F|}$} be a cluster of features. If for every $F_j \varepsilon$ S, SU (F_j, C) ≥ SU (F_i, C) ^ SU (F_i, F_j) > SU (F_i, C) is always corrected for each $F_i \varepsilon$ S (i ≠ j), then F_i are redundant features with respect to the given F_j (i.e., each F_i is a F-Redundancy).

Definition 4 (R-Feature). A feature $F_i \varepsilon$ S = {F1, F2 . . . F_k} (k< |F|) is a representative feature of the cluster S (i.e., F_i is a R-Feature) if and only if, $F_i = \text{argmax}_{Fj \varepsilon S}$ SU (F_j, C).

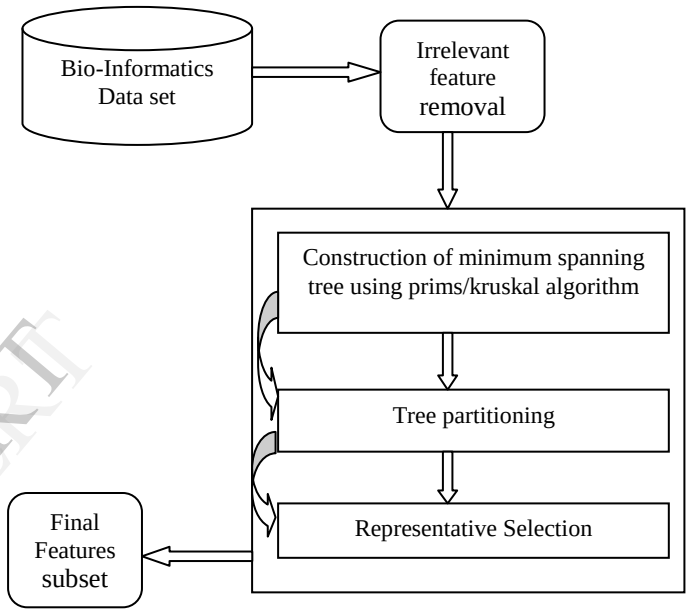


Fig-1 Detailed description of proposed work

IV METHODOLOGY

1. Load Data and Classify

Load the data into the process. The data has to be preprocessed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format which is the standard format for WEKA toolkit. From the arff format, only the attributes and the values are extracted and stored into the database. By considering the last column of the dataset as the class attribute and select the distinct class labels from that and classify the entire dataset with respect to class labels.

2. Information Gain Computation

Relevant features have strong correlation with target concept and necessary for a best subset, but redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation. To find the relevance of each attribute with the class label, Information gain is computed in this module. This is also said to be Mutual Information measure.

Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes. The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification.

3. T-Relevance

The relevance between the feature F_i € F and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by SU (F_i, C). If SU (F_i, C) is greater than a predetermined threshold, then it is implied that Fi is a strong T-Relevance feature. After finding the relevance value, the redundant attributes will be removed with respect to the threshold value.

4. F-Correlation

The correlation between any pair of features F_i and F_j (F_i, F_j € ^ F ^ i ≠ j) is called the F-Correlation of F_i and F_j, and denoted by SU (F_i, F_j). The equation symmetric uncertainty which is used for finding the relevance between the attribute and the class is again applied to find the similarity between two attributes with respect to each label.

5. MST Construction

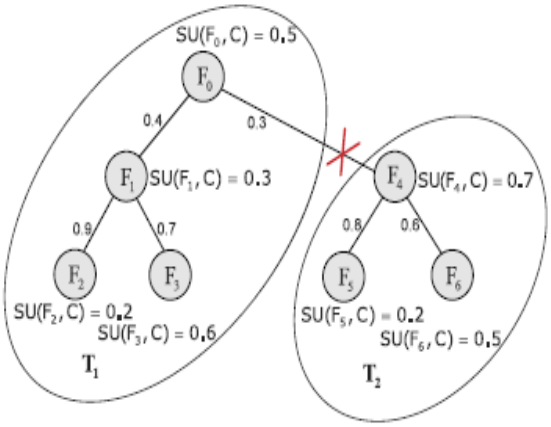
With the F-Correlation value computed, the Minimum Spanning tree is constructed. Tree is constructed using, Kruskal algorithm, which form MST effectively. Kruskal algorithm is a greedy algorithm in graph theory that finds a minimum spanning tree for a connected weighted graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a minimum spanning tree for each connected component).

Description:

1) Create a forest F (a set of trees), where each vertex in the graph is a separate tree.
2) Create a set S containing all the edges in the graph
3) While S is nonempty and F is not yet spanning
    a. remove an edge with minimum weight from S
    b. if that edge connects two different trees, then add it to the forest, combining two trees into a single tree
    c. Otherwise discard that edge.

At the termination of the algorithm, the forest forms a minimum spanning forest of the graph. If the graph is connected, the forest has a single component and forms a minimum spanning tree.

The sample tree is as shown in the figure,



In this tree, the vertices represent the relevance value and edges represent the F-Correlation value. The complete graph G reflects the correlations among all the target-relevant features. Unfortunately, graph G has k vertices and k (k-1)/2 edges. For high-dimensional data, it is heavily dense and the edges with different weights are strongly interwoven. Moreover, the decomposition of complete graph is NP-hard.

So for graph G, an MST is built, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well known Kruskal algorithm. The weight of edge (F_i`, Fj`) is F-Correlation SU (F_i`,$F_j$`).

6. Cluster Formation

After building the MST, in the third step, we first remove the edges whose weights are smaller than both of the T-Relevance SU (F_i`, C) and SU ($F_j$`, C), from the MST. After removing all the unnecessary edges, a forest is obtained. Each tree T_j € Forest represents a cluster that is denoted as V (T_j), which is the vertex set of T_j as well. As illustrated above, the features in each cluster are redundant, so for each cluster V (T_j) we choose a representative feature F_j, where T- Relevance SU (F_j R, C) is the greatest.

Table 1: Computation of Relevant and Redundancy Features

Feature	Formula
Symmetric Uncertainty	SU(X,Y)= 2*Gain(X|Y)/H(X)+H(Y)
Entropy	$H(X)= -\sum_{x\in X} p(x)\log_2 p(x)$
Conditional Entropy	$H(X\|Y)= -\sum_{y\in Y} p(y)\sum_{x\in X} p(x\|y)\log_2 p(x\|y)$
Gain	Gain(X|Y)=H(X)-H(Y) =H(Y)-H(Y|X)

V CONCLUSION.

In this paper, a novel clustering-based feature subset selection algorithm for high dimensional data is proposed. 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 algorithm, cluster method is used, in which each cluster consists of features and each cluster is treated as a single feature and thus dimensionality is drastically reduced.

REFERENCES

[1] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, vol. 5, pp. 1531-1555, 2004.

[2] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.

[3] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

[4] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324, 1997

[5] D. Koller and M. Sahami, "Toward Optimal Feature Selection,"Proc. Int'l Conf. Machine Learning, pp. 284-292, 1996.

[6] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[7] 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.

[8] Z. Zhao and H. Liu, "Searching for Interacting Features in Subset Selection," J. Intelligent Data Analysis, vol. 13, no. 2, pp. 207-228, 2009

[9] Z. Zhao and H. Liu, "Searching for Interacting Features," Proc.20th Int'l Joint Conf. Artificial Intelligence, 2007.

[10] C. Sha, X. Qiu, and A. Zhou, "Feature Selection Based on a New Dependency Measure," Proc. Fifth Int'l Conf. Fuzzy Systems and Knowledge Discovery, vol. 1, pp. 266-270, 2008.

[11] J. Biesiada and W. Duch, "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[12] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

[13] S. Chikhi and S. Benhammada, "Relief MSS: A Variation on a Feature Ranking Relieff Algorithm," Int'l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.

[14] Nokia (2005) Nokia NFC & RFID SDK 1.0 Programmer's Guide NFC. V 1.0,

[15] NFC Forum (2008) Generic control record type definition, NFCForum-TS-GenericControlRTD_1.0, 2008.3.7

[16] Han J, Lee H, Park K-R (2009) Remote-controllable and energy saving room architecture based on ZigBee communication. IEEE Trans Consumer Electron 55(1):264 268.

a.