Clustering Based Feature Subset Selection for High Dimensional data

Lalith Prasad K A[1] Prof. Prakasha S[2]
[1]M.Tech Scholar [2]Assistant Professor
[2]Department of CNE
[1,2]RNS Institute of Technology Bangalore, Karnataka, India

Abstract— Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis. Feature 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. 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. Features in different clusters are relatively independent; the clustering-based strategy of this algorithm has a high probability of producing a subset of useful and independent features. To ensure the efficiency of this algorithm, we adopt the efficient minimum-spanning tree clustering method. Features in different clusters are relatively independent; the clustering-based strategy of this algorithm has a high probability of producing a subset of useful and independent features. The efficiency and the effectiveness of the algorithm can be evaluated by the runtime efficiency, classification accuracy and the subset of features selected. Extensive experiments are performed on this algorithm by comparing it with known classifiers like tree-based C4.5. The results on the high dimensional image and microarray data, demonstrate that the algorithm not only produces smaller subsets of features but also improves the performance of the C4.5 classifier.

I. INTRODUCTION

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many *redundant* or *irrelevant* features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques provide three main benefits when constructing predictive models:

- Improved model interpretability,
- Shorter training times,
- Enhanced generalization by reducing over fitting.

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. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods.

Wrapper model use the predictive accuracy of predetermined algorithm to determine the goodness of selected subset features and the accuracy of the learning algorithm is usually high. However the generality of the selected features is limited and computational complexity is large. Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Common measures include the mutual information, the point wise mutual information, Pearson product-moment correlation coefficient, inter/intra class distance or the scores of significance tests for each class/feature combinations. Filters are usually less computationally intensive than wrappers, but they produce a feature set which is not tuned to a specific type of predictive model. Many filters provide a feature ranking rather than an explicit best feature subset, and the cutoff point in the ranking is chosen via cross-validation. Filter methods have also been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems.

Embedded methods are a catch-all group of techniques which perform feature selection as part of the model construction process. The exemplar of this approach is the LASSO method for constructing a linear model, which penalizes the regression coefficients, shrinking many of them to zero. Any features which have non-zero regression

coefficients are 'selected' by the LASSO algorithm. One other popular approach is the Recursive Feature Elimination algorithm, commonly used with Support Vector Machines to repeatedly construct a model and remove features with low weights. These approaches tend to be between filters and wrappers in terms of computational complexity.

In this paper, graph-theoretic clustering methods are applied to the features. In particular, *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. 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. 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 feature Selection algorithm is designed. 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 the algorithm has a high probability of producing a subset of useful and independent features.

II. Related Work

It can be inferred that there are two broad categories of feature selection filter method and wrapper method. An algorithm FWHM is used. FWHM is divided into two different phases which performs some processing on the features selected. The Filter-Wrapper Hybrid Method (FWHM) improves the conventional simple two-phase feature selection algorithm. It integrates the ranking results of multi-filter methods so as to avoiding the uncertainty of using single criterion method. The improvements of the random search strategies in the wrapper phase link the filter and wrapper phase closely. Hence, the rationality and scientificity of the algorithm is guaranteed [1].

Feature selection in the "wrapper" model, typically involves an NP-hard optimization problem that is approximated by heuristic search for a "good" feature subset. First considering the idealization where this optimization is performed exactly, author gives a rigorous bound for generalization error under feature selection. The search heuristics typically used are then immediately seen as trying to achieve the error given in bounds, and succeeding to the extent that they succeed in solving the optimization. The bound suggests that, in the presence of many "irrelevant" features, the main source of error in wrapper model feature selection is from "over fitting" hold-out or cross-validation data.

A new algorithm under the idealization of performing search exactly has sample complexity (and error) that grows logarithmically in the number of "irrelevant" features - which means it can tolerate having a number of "irrelevant" features exponential in the number of training examples - and search heuristics are again seen to be directly trying to reach this bound. This paper discusses ramifications that sample complexity logarithmic in the number of irrelevant features might have for feature design in actual applications of learning [2].

Distributional Clustering can be used for document classification. This approach clusters words into groups based on the distribution of class labels associated with each word. Thus, unlike some other unsupervised dimensionality-reduction techniques, such as Latent Semantic Indexing, author is able to compress the feature space much more aggressively, while still maintaining high document classification accuracy. Experimental results obtained on three real-world data sets show that the method can reduce the feature dimensionality by three orders of magnitude and lose only 2% accuracy significantly better than Latent Semantic Indexing, class-based clustering, feature selection by mutual information, or Markov blanket-based feature selection. It also shows that less aggressive clustering sometimes results in improved classification accuracy over classification without clustering [3].

High dimensionality of text can be a deterrent in applying complex learners such as Support Vector Machines to the task of text classification. Feature clustering is a powerful alternative to feature selection for reducing the dimensionality of text data. Information-theoretic divisive algorithm can be used for feature/word clustering and apply it to text classification. Existing techniques for such "distributional clustering" of words are agglomerative in nature and result in (i) sub-optimal word clusters and (ii) high computational cost. In order to explicitly capture the optimality of word clusters in an information theoretic framework, method first derives a global criterion for feature clustering. A divisive algorithm that monotonically decreases this objective function value. The algorithm minimizes the "within-cluster Jensen-Shannon divergence" while simultaneously maximizing the "between-cluster Jensen-Shannon divergence". In comparison to the previously proposed agglomerative strategies the divisive algorithm is much faster and achieves comparable or higher classification accuracies. It is observed that feature clustering is an effective technique for building smaller class models in hierarchical classification [4].

III. Proposed Work

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other".

A novel algorithm is developed 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 4.2) 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.

Based on the MST method, we propose an algorithm. This 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 this algorithm has a high probability of producing a subset of useful and independent features.

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 the proposed 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. Relevant features have strong correlation with target concept so are always necessary for a best subset, while 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.

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 by a number of researchers. Therefore, we choose symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept.

Symmetric uncertainty is defined as follows:

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

Where,

1) H(X) is the entropy of discrete random variable X. Suppose $p(x)$ is the prior possibilities for all the values of X , H(X) is defined by
$$H(X) = -\sum_{x \varepsilon X} p(x) log\ p(x)$$
2) *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 about Y provided by X and is called the information gain, which is given by
$$Gain\ (X|Y) = H(X) - H\ (X|Y)$$
$$= H(Y) - H\ (Y|X)$$

Information gain is a symmetrical measure. That is the amount of information gained about X after observing Y is equal to the amount of information gained about Y after observing X. This ensures that the order of two variables (e.g.,(X, Y) or(Y,X)) will not affect the value of the measure. Symmetric uncertainty treats a pair of variables symmetrically; it compensates for information gain's bias toward variables with more values and normalizes its value to the range [0,1]. A value 1 of $SU(X, Y)$ indicates that knowledge of the value of either one completely predicts the value of the other and the value 0 reveals that X and Y are independent. Although the entropy based measure handles nominal or discrete variables, they can deal with continuous features as well, if the values are discretized properly in advance.

Feature subset selection can be the process that identifies and retains the strong T-Relevance features and selects R-Features from feature clusters. The behind heuristics are that

- Irrelevant features have no/weak correlation with target concept;
- Redundant features are assembled in a cluster and a representative feature can be taken out of the cluster.

IV. IMPLEMENTATION

In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we firstly present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation

Pseudo code

Inputs: D (F1, F_2,…, F_m, C) - the given data set
θ - The T-Relevance threshold
Output: S- Selected feature subset
// Irrelevant feature removal//
1. for i=1 to m do
2. T-Relevance = SU (F_i, C)
3. if T-Relevance >θthen
4. S = S ∩ {F_i}
// Part-2: Minimum spanning Tree construction //
5. G = NULL; // G is a complete graph
6. for each pair of features { F'_i , F'_j } ⊂ 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 the minimum spanning tree;
10. Forest = minSpanTree
11. for each edge E_{ij} ϵ Forest do
12. if SU(F'_i , F'_j) < SU(F'_i , C) ∧ SU(F'_i , F'_j) < SU(F'_i , C) then
13. Forest = Forest - E_{ij}
14. S = ∅
15. for each tree T_iϵ Forest do
16. F^j_R= argmax$_{F'k \epsilon Ti}$ SU (F'_k, C)
17. S = S ∪ {F^j_R} ;
18. return S

V. CONCLUSION

In this project, novel filter feature selection method based on information theory. The idea of using the symmetric uncertainty is used to evaluate the correlation among the features and the features and the target classes. The method's complexity, efficiency, and other required are considered. By observing the methods, suitable method for feature selection is chosen.

The research work has been done on various aspects of feature selection and how it can be performed efficiently. Some authors discuss on how sampling of datasets can be done for efficient extraction of useful patterns. Examination of two categories of feature selection filter and wrapper method is performed and an algorithm which uses hybrid method has been performed. Different classification methods have been examined. One of the efficient methods to form a cluster is the Minimum Spanning Tree (MST). Various methods of Graph-based clustering have been considered.

Mutual information (MI) is used to evaluate the features and select informative subset to be used for neural network classifier. MI measures the dependencies between random variables. MI based features selection algorithms have been considered for survey. Some works which considers the existing algorithm RReliefF, its implementation, experimental analysis have been analyzed. Artificial Neural Network (ANN) have been considered to perform feature selection using both filter and wrapper method.

One of the main challenges of the data mining is the reduction in the dimensionality. One of them is the feature selection. Some of the data sets like image, text, microarray are high dimensional in nature and contain irrelevant and redundant features which would increase the complexity and the processing time.

ACKNOWLEDGEMENT

I take this Opportunity to express my profound gratitude and deep regards to my guide Prof, Mr.Prakasha S Assistant Professor, RNS Institute of technology, Bangalore, for his exemplary guidance, and constant encouragement throughout.

REFERENCES

[1] Hu Min, Wu Fangfang, Filter-Wrapper Hybrid method on feature selection, IEEE 2010
[2] Andrew Y. Ng, On feature selection: Learning with Exponentially many irrelevant features as training examples, IEEE Transactions on knowledge and Data Engineering
[3] L. Douglas Baker, Andrew Kachites McCallum, Distributional clustering of words for text classification, ACM 1998
[4] Inderjit S. Dhillon, SubramanyamMallela, Rahul kumar, A Divisive Information-Theoretic Feature clustering algorithm for text classification, Journal of machine learning 2003
[5] Fernando Pereira, NaftaliTishby, Lillian Lee,. Distributional clustering of english words, ACL 1993.
[6] Jerzy W. Jaromczyk, Godfried T. Toussaint, Relative neighborhood Graphs and their Relatives, IEEE Transactions on knowledge and Data Engineering, 2003
[7] Isabelle Guyon, Andre Elisseeff, An introduction to Variable and feature selection, Journal of Machine Learning Research, 2003
[8] Roberto Battiti, Using mutual information for selecting features in supervised Neural Net Learning, IEEE 2004
[9] Richard Butterworth, Gregory Piatetsky-Shapiro, Dan A. Simovici, On Feature Selection through clustering, IEEE 2005
[10] George H. John, Ron Kohavi, Karl Pfleger, Irrelevant features and the subset selection problem, IEEE Transactions on knowledge and Data Engineering 1994
[11] Luis Carlos Molina, Lluis Belanche, Angela Nebot, Feature Selection Algorithms: A survey and experimental Evaluation, IEEE, 2002
[12] Theoretical and Empirical Analysis of ReliefF and RReliefF, Marko Robnik-Sikonja, Igor Kononenko, 2003 Machine Learning Journal.
[13] M. Scherf, W. Brauer, Feature Selection by Means of a Feature Weighting Approach, IEEE Transaction on Knowledge and Data Engineering 2005.

