Clusters using High Dimensional Data for Feature Subset Algorithm

V. Abhilash

M. Tech Student Department of Computer Science and Engineering
SV College of engineering, Tirupati, Chittor, Andhra Pradesh, India
Email- velamuriabhilash@gmail.com

K. Shanthi

Associate Professor Department of Computer Science and Engineering
SV College of engineering Tirupathi, Chittor, Andhra Pradesh, India
Email- santhi@svcolleges.edu.in

Abstract: - The importance of bringing causality into play when designing feature selection methods is more and more acknowledged in the machine learning community. This paper proposes a filter approach based on information theory which aims to priorities direct causal relationships in feature selection problems where the ratio between the number of features and the number of samples is high. This approach is based on the notion of interaction which is shown to be informative about the relevance of an input subset as well as its causal relationship with the target. The resulting filters, called m-IMR (min-Interaction Max-Relevance), is compared with state-of-the-art approaches. Classification results on 25 real microarray datasets show that the incorporation of causal aspects in the feature assessment is beneficial both for the resulting accuracy and stability. A toy example of causal discovery shows the effectiveness of the filter for identifying direct causal relationships.

Index Terms—Feature subset selection, filter method, feature clustering, graph-based clustering.

1 INTRODUCTION

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.

V. Abhilash*, M .Tech Student, Department of CS, JNTUA, Anantapur/ SV College of Engineering Tirupati /India. (email: velamuriabhilash@gmail.com).*

K.Shanthi*, Associate Professor Department of CSE, JNTUA/ Anantapur, SV College of Engineering / Tirupati /India, (e-mail: santhi@svcolleges.edu.in).*

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.

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 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, we propose a Fast clustering-bAsed feature Selection algoriThm (FAST).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. The proposed feature subset selection algorithm FAST was tested upon 35 publicly available image, microarray, and text data sets. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers.

2 RELATED 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: (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. Our proposed FAST algorithm falls into the second group.

Traditionally, feature subset selection research has focused on searching for relevant features. A well known example is Relief [34], 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. 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. However, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well are examples that take into consideration the redundant features.

Hierarchical clustering also has been used to select features on spectral data. Van Dijk and Van Hullefor proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. 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 except that the former forces every cluster to contain consecutive features only. Both methods employed agglomerative hierarchical clustering to remove redundant features.

Quite different from these hierarchical clustering based algorithms, our proposed FAST algorithm uses minimum spanning tree based method to cluster features. Meanwhile, it does not assume that data points are grouped around centers or separated by a regular geometric curve. Moreover, our proposed FAST does not limit to some specific types of data.

3 FEATURE SUBSET SELECTION ALGORITHM

3.1 Framework and definitions

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.*"

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.

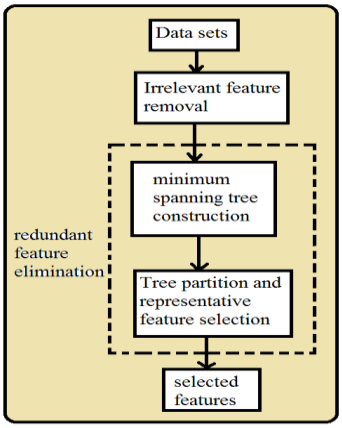

Fig. 1: Framework of the proposed feature subset selection algorithm

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, 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.

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 as follows.

Definition *1: (Relevant feature)* Fi is relevant to the target concept C if and only if there exists some $s'i$, fi and c, such that, for probability $p(S'i = s'i, Fi = fi) > 0$, $p(C = c \mid S'i = s'i, Fi = fi) \neq p(C = c \mid S'i = s'i)$. Otherwise, feature Fi is an *irrelevant feature*.

Definition 1 indicates that there are two kinds of relevant features due to different $S'i$: (i) when $S'i = Si$, from the definition we can know that Fi is directly relevant to the target concept; (ii) when $S'i \subsetneq Si$, from the definition we may obtain that $p(C|Si, Fi) = p(C|Si)$. It seems that Fi is irrelevant to the target concept. However, the definition shows that feature Fi is relevant when using $S'i \cup \{Fi\}$ to describe the target concept. There a son behind is that either Fi is interactive with $S'i$ or Fi is redundant with $Si - S'i$. In this case, we say Fi is indirectly relevant to the target concept.

Most of the information contained in redundant features is already present in other features. As a result, redundant features do not contribute to getting better interpreting ability to the target concept. It is formally defined by Yu and Liu based on Markov blanket. The definitions of Markov blanket and redundant feature are introduced as follows, respectively.

Definition *2: (Markov blanket)* Given a feature $Fi \in F$, let $Mi \subset F (Fi \notin Mi)$, Mi is said to be a Markov blanket for Fi if and only if $p(F - Mi - \{Fi\}, C \mid Fi, Mi) = p(F - Mi - \{Fi\}, C \mid Mi)$.

Definition *3: (Redundant feature)* let S be a set of features, a feature in S is redundant if and only if it has a Markov Blanket within S.

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.

The *symmetric uncertainty* is defined as follows

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}. \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where,

1) $H(X)$ 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

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

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 gain which is given by

$$\begin{aligned} Gain(X|Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

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

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y). \quad \ldots\ldots\ldots (4)$$

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 discredited properly in advance.

Given $SU(X, Y)$ the symmetric uncertainty of variables X and Y , the relevance T-Relevance between a feature and the target concept C, the correlation F Correlation between a pair of features, the feature redundancy F-Redundancy and the representative feature R-Feature of a feature cluster can be defined as follows.

Definition 4: (T-Relevance) The relevance between the feature $Fi \in F$ and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by $SU(Fi, C)$. If $SU(Fi, C)$ is greater than a predetermined threshold θ, we say that Fi is a strong T-Relevance feature.

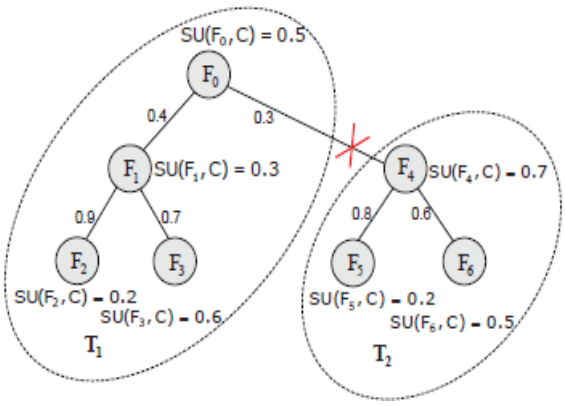

Fig. 2: Example of the clustering step

After removing all the unnecessary edges, a forest *Forest* is obtained. Each tree $Tj \in Forest$ represents a cluster that is denoted as $V(Tj)$, which is the vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster $V(Tj)$ we choose a representative feature FjR whose *T-Relevance* $SU(FjR, C)$ is the greatest. All FjR (j = 1...*|Forest|*) comprise the final feature subset $\cup FjR$.

4. CONCLUSION:

Based on the minimum spanning tree method, we recommend a FAST algorithm. The algorithm is a two steps process in which, characteristics are divided into clusters by means of using graph-theoretic clustering means. Feature subset selection can be analyzed as the process of recognizing and eliminating as many inappropriate and redundant features as promising since: inappropriate features do not put in to the predictive accurateness and redundant characteristics do not redound to getting an enhanced predictor for that they make available mainly information which is by now present in previous feature. In the subsequent step, the mainly used representative feature that is robustly related to target classes is particular from each cluster to structure the final subset of features. Features in altered clusters are comparatively autonomous; the clustering based scheme of FAST has a high possibility of producing a subset of constructive and independent characteristics. In our projected FAST algorithm, it entails the building of the minimum spanning tree from a subjective inclusive graph; the separation of the minimum spanning tree into a forest by means of every tree signifying a cluster; and the collection of representative features from the clusters. The projected feature subset selection algorithm FAST was tested and the investigational results demonstrate that, evaluated with other various types of feature subset selection algorithms, the projected algorithm not only decrease the number of features, but also advances the performances of the renowned various types of classifiers.

5. REFERENCES

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.
[2] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994.
[3] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.
[4] 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.
[5] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537-550, 1994.
[6] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.
[7] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.
[8] 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

ABOUT AUTHORS

V.Abhilash received the B.Tech Degree in Computer Science and Engineering from KMMITS, University of JNTUA in 2012. He is currently working towards the Master's Degree in Computer Science, in SV College of Engineering University of JNTUA. His interest lies in the areas of Data Mining, SQL, DBA, DB2 DBA.

K. Shanthi Received B.Tech and M.Tech Degrees in Computer Science and Engineering from in Vignan's engineering college, JNTUH, JNTU Hyderabad in 2003 & 2006 respectively. Currently she is an Associate Professor in the Department of Computer Science and Engineering at SV College of Engineering-Tirupati.

She has published a paper titled **"A survey on Image Retrieval using Data Mining based on Future Clustering Techniques".** in Journals and refereed Conference Proceedings. Her current interests include Data Mining, Operating Systems, Computer Networks and Information Security.