A Fast Clustering-Based Feature Subset Selection Algorithm

Varsha S. Sonwane

ME (C.S.E)
Computer science and Engineering
SYCET,Aurangabad,Maharashtra.

Abstract*—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. Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. 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 a 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. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results, on 35 publicly available real-world high dimensional image, microarray, and text data, demonstrate that FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

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

INTRODUCTION

With the aim of choosing a subset of good features with respect to the target concepts, feature subset se-lection 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 spe- cific to given learning algorithms, and therefore may be more efficient than the other three categories[11]. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of em- bedded 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. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed [5].The hybrid methods are a combination of filter and wrapper methods [7] 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 com- plexity of the filter methods. The wrapper methods are computationally expensive and tend to overfit on small training sets [5], [7]. 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.With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algo- rithms. Pereira et al Baker et al. [4], and Dhillon et al. [8] employed the distributional clustering of words to reduce the dimensionality of text data.

In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their re- sults have, sometimes, the best agreement with human performance [13]. The general graph-theoretic cluster- ing is simple: Compute a neighborhood graph of in- stances, 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, fea- tures are divided into clusters by using graph-theoretic clustering methods. In the second step, the most repre- sentative 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 se- lection 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. The rest of the article is organized as follows: In Section 2, we describe the Literature survey. In Section 3, we describe related works, Finally, in Section 4, we summarize the present study and draw some conclusions.

2.LITERATURE SURVEY

Existing system:

1) RELIEF

It is a feature selection algorithm used in binary classification. Relief is ineffective at removing redundant features. RELIEF is a feature selection algorithm used in binary classification (generalisable to polynomial classification by decomposition into a number of binary problems) proposed by Kira and Rendell in 1992.[1] Its strengths are that it is not dependent on heuristics, requires only linear time in the number of given features and training instances, and is noise-tolerant and robust to feature interactions, as well as being applicable for binary or continuous data; however, it does not discriminate between redundant features, and low numbers of training instances fool the algorithm. Kononenko et al. proposed some updates to the algorithm (RELIEFF) in order to improve the reliability of the probability approximation, make it robust to incomplete data, and generalising it to multi-class problems.[2]

2) Relief-F

It is a feature selection strategy that chooses instances randomly. It cannot identify redundant features. ReliefF is a simple yet efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes. In practice, ReliefF is usually applied in data pre-processing as a feature subset selection method.

3) Hierarchical clustering:

Hierarchical clustering algorithm is of two types:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm are exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.

Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods.

Disadvantages of Existing system

1) Hierarchical clustering Algorithm can never undo what was done previously.

2) Hierarchical clustering Time complexity of at least O(n2 log n) is required, where ‘n’ is the number of data points.

4) No objective function is directly minimized in Hierarchical clustering

5) Sometimes it is difficult to identify the correct number of clusters by the dendogram in Hierarchical clustering.

6)The generality of the selected features is limited

7)The computational complexity is large.

8)Accuracy is not guaranteed.

9)Ineffective at removing redundant features

3.PROPOSED SYSTEM

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 [14], 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 [10], [12].

3.1.FEATURE SUBSET SELECTION ALGORITHM

3.1.1 Framework and dentitions:Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines [12]. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and re-relevant to the target concept; Moreover, “good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.”

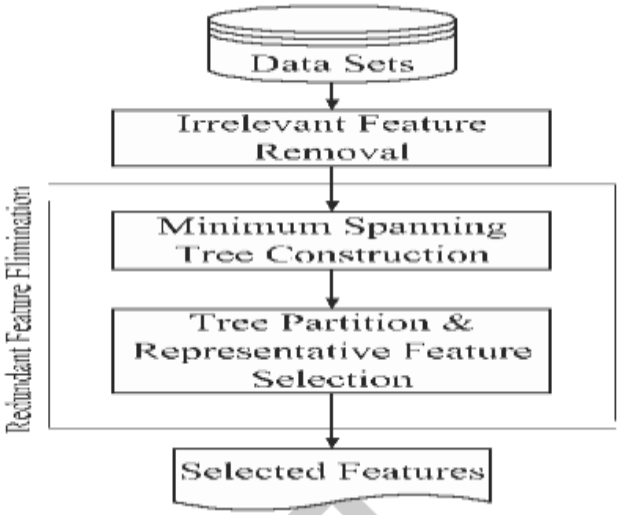

Fig. 1: Framework of the proposed feature subset selection algorithm

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. 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. John et al. [14] presented a definition of relevant features. Suppose F to be the full set of features, $Fi \in$ be a feature, $Si = F - \{Fi\}$ and $S'i \subseteq Si$. Let $s'i$ be a value assignment of all features in S', fi a value-assignment of feature Fi, and c a value-assignment of the target concept C. The definition can be formalized as follows.
The definition can be formalized 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 $p(S'i = si, Fi = fi) > 0$, $p(C = c \mid Si = s'i, Fi = fi) \neq p(C = c \mid S'i = s'i)$. 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 $p(C \mid Si, Fi) = p(C \mid Si)$.

***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. The *symmetric uncertainty* is defined as follows

$(X, Y) = 2 \times (X/Y)/H(X) + H(Y)$. (1)Where, 1) (X) is the entropy of a discrete random variable X. Suppose (x) is the prior probabilities for all values of X, (X) is defined by $(X) = - \sum_{x \in X} p(x) \log 2\ p(x)$. (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 [55] which is given by

$$Gain(X/Y) = H(X) - H(X/Y)$$
$$= H(Y) - H(Y/X).$$

Where $(X/)$ 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 (x) is the prior probabilities for all values of X and (x/y) is the posterior probabilities of X given the values of
$(X/)$ is defined by

$(X/Y) = -\Sigma(y) \quad \Sigma p(x/y) \log 2\, p(x/y)$.
$y \in Y \qquad x \in X$

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 (Fi, C). If (Fi, C) is greater than a predetermined threshold θ, we say that Fi is a strong *T-Relevance* feature.

Definition *5: (F-Correlation)* The correlation between any pair of features Fi and Fj $(Fi, Fj \in F \wedge i \neq j)$ is called the *F-Correlation* of Fi and Fj, and denoted by $SU(Fi, Fj)$.

Definition *6: (F-Redundancy)* Let $S = \{F1, F2, ..., Fi,..., Fk<|F|\}$ be a cluster of features.if $\exists Fj \in S, SU(Fj, C) \geq SU(Fi, C) \wedge SU(Fi, Fj) > SU(Fi, C)$ is always corrected for each $Fi \in S$ $(i \neq j)$, then Fi are redundant features with respect to the given Fj (i.e. each Fi is a *F-Redundancy*).

Definition *7: (R-Feature)* A feature $Fi \in S = \{F1, F2, ..., Fk\}$ $(k < |F|)$ is a representative feature of the cluster S (i.e. Fi is a *R-Feature*) if and only if, $Fi = \text{argmax}_{Fj \in S} SU(Fj, C)$. This means the feature, which has the strongest *TRelevance*, can act as a *R-Feature* for all the features in the cluster.

4. CONCLUSION

Obtaining a suitable rank as to where the FAST algorithm exactly stands amongst few other existents. In this, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data.In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. 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.For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space. In feature we are going to classify the high dimensional data. In this paper, we have presented a novel clustering-based. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime.For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space.

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] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
[6] Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.
[7] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection,In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.
.[8] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3, pp 1265-1287, 2003.
[9]Dougherty, E. R., Small sample issues for microarray-based classification. Comparative and Functional Genomics, 2(1), pp 28-34, 2001.
[10]Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.
[11] Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003. Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.Learning Research, 3, pp 1157-1182, 2003. [29] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.
[12]Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.
[13] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and Their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517, 1992.
[14]John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.
[15] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif. Intell., 97(1-2), pp 273-324, 997.