Clustering Based Feature Subset Selection

Gurusamy P[1] Neethu Subash[2]
[1,2]Department of Computer Science and Engineering
[1,2]Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract— In data mining, the Feature 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 contain many redundant or irrelevant features. 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 CBFS algorithm eliminate them very efficiently. Traditionally, feature subset selection research has focused on searching for relevant features. The results on the five text and micro array data sets show that CBFS can effectively identify irrelevant and redundant features. And CBFS can not only efficiently reduce the feature space, but also can significantly improve the performance of the four well-known classifiers.

Key words: Clustering, Subset Selection, CBFS algorithm

I. INTRODUCTION

In machine learning, feature selection also known as variable selection, attribute selection or variable subset selection. It 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 hold many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no valuable 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 are often used in domains where there are many features and comparatively few samples or data points. The archetypal case is the use of feature selection in analysing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples. Feature selection techniques provide three main benefits when constructing predictive models: improved model interpretability, shorter training times, and enhanced by reducing over fitting. 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.

General procedure of feature selection has four key steps as shown in fig. 1. First step, Subset generation is essentially a process of heuristic search, with each state in the search space specifying a candidate subset for evaluation. Search may start with an empty set and successively add features (i.e., forward), or start with a full set and successively remove features (i.e., backward), or start with both ends and add and remove features simultaneously (i.e., bidirectional). Therefore, different strategies have been explored: complete [4], sequential [5], and random [3] search. Second step, Subset Evaluation is each newly generated subset needs to be evaluated by an evaluation criterion. An evaluation criterion can be independent criteria (in filter model), or dependent criteria (in wrapper model). Third step, Stopping Criteria is when the feature selection process should stop. Some criteria are as: when search completes, minimum number of features or maximum number of iterations, subsequent addition or deletion of any feature does not produce a better subset, sufficiently good subset is selected. Final step, Result Validation is done by prior knowledge or by some indirect methods by mining performance.

The proposed CBFS algorithm employs the clustering-based method to choose features. The general graph-theoretic clustering is simple: compute a graph of instances, then delete any edge in the graph according to some criterion. The result will generate a forest and each tree of the forest consider as a cluster. We use 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 Clustering Based Feature Selection algorithm (CBFS).

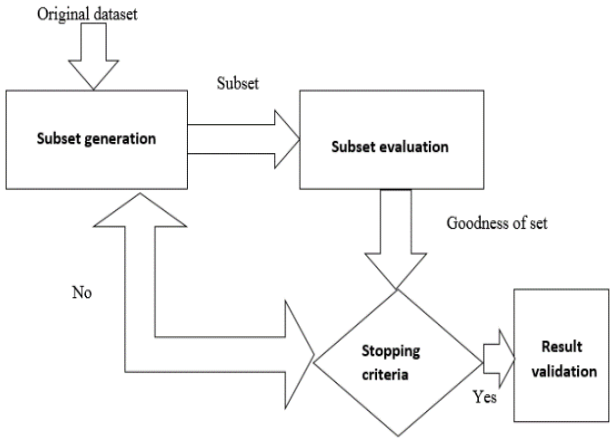


Fig. 1: Steps in feature selection algorithm

In this algorithm the features are divided into clusters by using graph-theoretic clustering methods. And then the most relevant feature that is strongly related to target classes is selected from each cluster to form the final subset of features. The proposed feature subset selection algorithm was tested upon publicly available 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.

II. RELATED WORK

Feature subset selection has been an active research topic since 1970s, and a great deal of research work has been published.

Feature subset selection can be noticed as the method of identifying and eliminating as many irrelevant and redundant features as possible. Because irrelevant features do not have the similarity with the target data, and redundant features do not redound to getting a better predictor. Of the many feature subset selection algorithms, some can efficiently eliminate irrelevant features but miscarry to handle redundant features [8], [9], and [10]. But proposed algorithm can eliminate the irrelevant while taking care of the redundant features as in [7], [11].

Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories as: Embedded, Wrapper, Filter, and Hybrid. A large number of algorithms have been used for the feature selection problem.

The embedded methods integrate feature selection as a part of the training process and are usually specific to given learning algorithms. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. Embedded techniques integrate the search for an optimal feature subset into the process of building a classifier. The benefits of the embedded techniques are as: they model feature dependencies; they interact with the classifier; and in terms of computational complexity, they are better than wrapper methods. But they are computationally more expensive than filter-based subset evaluation techniques, and they are classifier dependent [1], [2].

In a filter [6] model, the feature selection is performed as a pre-processing step to classification. Selection process is performed independently which is used to induce the classifier. In order to evaluate a feature, or a subset of features, filters apply an evaluation function that measures the discriminating ability of the feature or the subset to differentiate class labels .Filters are generally much less computationally expensive than wrapper and hybrid algorithms. They may suffer from low performance if the evaluation criterion does not match the classifier well. Filter-based feature selection techniques are divided into two subcategories: (1) each individual feature in the case of filter based feature ranking (known as ranker or univariate techniques) or (2) the entire feature subsets in the case of filter-based subset evaluation (known as subset-evaluation or multivariate techniques)[1]. The multivariate techniques are the model of feature dependencies, they are independent of the classifier, and they are more efficient than wrapper methods in terms of computational complexity. But they are slower than univariate (ranker) techniques; they are less stable than univariate techniques; and they ignore interaction with the classifier, so they are outperformed by the wrapper technique in general. The Correlation-based Feature Selection (CFS) [7] and INTERACT algorithm are the filter models.

Wrappers [6] do use the learning algorithm as an integral part of the selection process. The selection of features should consider the characteristics of the classifier. Then, in order to evaluate subsets, wrappers use the classifier error rate induced by the learning algorithms as its evaluation function (Euclidean distance measure, entropy, information gain, correlation coefficient, min-features bias, etc.).This aspect of wrappers results in higher accuracy performance for subset selection than simple filters. Wrappers have to train a classifier for each subset evaluation, they are often much more time consuming. With wrapper based subset evaluation, however, features are evaluated in context, i.e., dependencies and correlations between features are considered. This is useful for discovering which features are either redundant or correlated with each other. The wrapper-based subset evaluation techniques are feature dependencies, and they interact with the classifier, so they have the potential to outperform rankers. Since they are computationally expensive and probably prohibitive for datasets with high-dimensionality, and they have higher risk of over fitting.

The main goal of hybrid systems for feature selection is to extract the good characteristics of filters and wrappers and combine them in one single solution. Hybrid algorithms achieve this behaviour usually by pre-evaluating the features with a filter in a way to reduce the search space to be considered by the subsequent wrapper.

Our proposed CBFS algorithm uses minimum spanning tree-based method to cluster features and fall in the filter category. Meanwhile, it does not limit to some specific types of data.

III. PROPOSED SYSTEM

Irrelevant features and redundant features severely affect the accuracy of the learning machines [8], [10]. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. And resulting feature subsets contain features highly correlated with the target class.

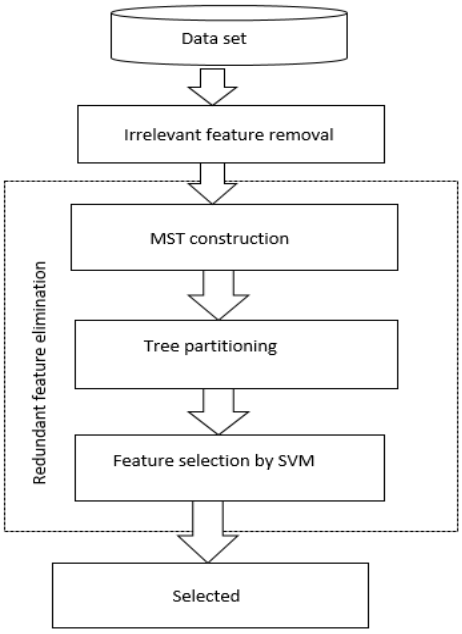


Fig. 2: Proposed feature selection method.

Proposed CBFS algorithm mainly consists of four steps.1) Removing irrelevant features, 2) constructing an

MST and 3) Partitioning the MST and 4) Selecting features. The proposed system contains the following components.

In first step, we calculate the T-Relevance of each feature in the input data set using the Symmetrical Uncertainty (SU). In second step, we compute F-Correlation value for each pair of feature which provide the value of feature to feature affiliation value. If two feature are more related to each other then they have high value of F-correlation. Third step remove the edges, whose weight is less than both of the T-Relevance and SU (F_j, C) from the MST. After removing the edges, a forest is obtained. Every tree of the forest represents a cluster. In step four, from the cluster the feature are carefully chosen by applying the SVM classifier.

The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of the feature values and target classes.

$$SU(x_k, \omega) = 2\frac{I((x_k,\omega)}{H(x_k)+H(\omega)} \quad (3.1)$$

Where the entropy of variable X is found by

$$H(X) = -\sum P(x_i)Log_2P(x_i) \quad (3.2)$$

The relevance between the features F_i U 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). Relevant Features are defined as, F_i is relevant to the target concept C if and only if there exists some S_i', f_i and c, such that, for probability P ($S_i' = s_i', F_i' = f_i) > 0$, P $(C = c| S_i' = s_i', F_i' = f_i) \neq$ P $(C = c \ |S_i' = s_i')$. And the correlation between any pair of features F_i and F_j (F_i, F_j ϵ F and i ≠ j) is called the F-Correlation of F_i and F_j, and denoted by SU (F_i, F_j). If SU(F_i,C) is greater than a predetermined threshold Θ then F_i is a strong T-Relevance feature and they are used to further processing and the rest are removed as irrelevant feature.

To construct the MST we use the T-relevance value as vertices of the graph and the F-correlation value as the weighted edge between the vertices $F_i and\ F_j$ of the graph. The MST is built through the prim's algorithm. Because the prim's algorithm has, the less complexity, then the other algorithm. A weighted complete graph G = (V, E) is constructed. As symmetric uncertainty is symmetric further the F correlation SU (F_i, F_j) is symmetric as well, thus G is an undirected graph.

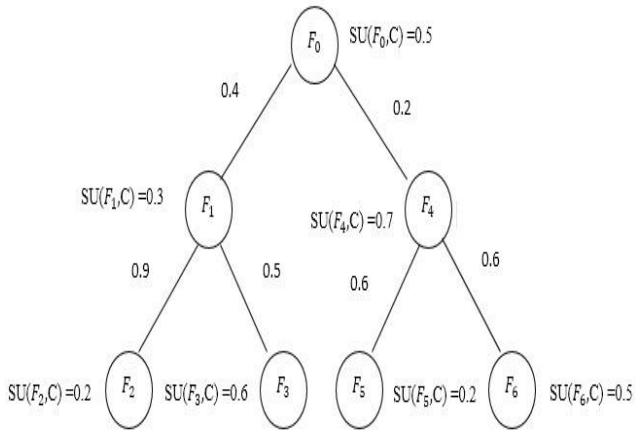
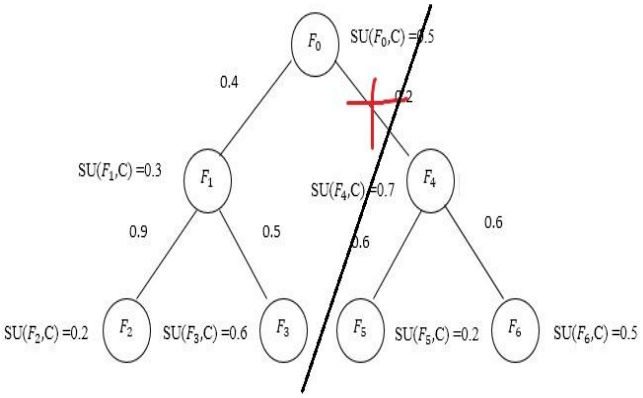


Fig. 3: MST construction using Prim's algorithm

The third step first eliminate the edges $E = \{(F_i', F_j') \mid (F_i', F_j' \in F' \wedge) \text{ I, j} \in [1, k] \wedge i \neq j\}$ whose weights are lesser than both of the T-Relevance SU(F_i,C) and SU(F_j,C) from the MST. Each deletion results in two disconnected trees T1 and T2.



T1 T2

Fig. 4: Graph clustered into two trees T1 and T2.

Support vector machines (SVMs) have been widely used as a classification tool. The support vector machine (SVM) is a hyper-plane that separates two different sets of samples with maximum distance of hyper-plane to nearest samples from both sets [18].The SVM classifier classifies the feature of each cluster into two class as one have maximum relevant feature and the other one is fewer relevant feature. Subsequently the best one is selected as the final feature from the clusters.

Algorithm 1:

Input: D ($F_1, F_2, F_3, \ldots, F_n$) - Data set
Θ - Threshold value
Output: S –selected feature subset

Algorithm CBFS **{**
//part1: Irrelevant feature removal
Irrelevant feature removal
for i =1 to m do
T-Relevance =SU (F_i, C)
if T-Relevance > Θ then
S=SU { F_i}
// part2: Minimum spanning tree construction
G=NULL // G is a complete graph
for each pair of(F_i', F_j') ∈S do
F-Correlation = SU(F_i', F_j')
Add F_i' and / or F_j' to G with the F-correlation as weight of the edge
MinSpanTree= Prims (G) //Using the prims algorithm to generate the minimum
Spanning tree
// part 3: Tree partition
Forest =MinSpanTree
for each $E_{i,j}$ element of Forest do
if SU(F_J,C) ≥SU(F_i,C) ^ SU (F_i, F_J) > SU(F_i,C) then
Forest = Forest -$E_{i,j}$
// part 4: Required feature subset for the machine learning
S =Φ
for each tree T_i ∈ Forest do
F_R^j= SVM (T_i)
S= S U {F_R^j} **}**
return S

IV. EVALUATION

Table 1: Summary of the 5 Benchmark Data Sets

Data Id	Data Name	F	I	T	Domain
1	Chess	37	3196	2	Text
2	mfeat-fourier	77	2000	10	Text
3	coil2000	86	9822	2	Text
4	elephant	232	1391	2	Micro Array
5	arrythmia	280	452	16	Micro Array

For the purposes of evaluating the performance and effectiveness of the proposed CBFS algorithm, verifying whether or not the method is useful in practice, 5 publicly available data sets were used. The numbers of features of the 5 data sets vary from 37 to 49. The dimensionality of the 54.3 percent data sets exceed 5,000, of which 28.6 percent data sets have more than 10,000 features. The 5 data sets cover a range of application domains such as text and bio microarray data classification. Table 1 shows the corresponding statistical information

V. CONCLUSION

Feature Selection techniques are studies and classified and define the better response of feature subset selection which is the search algorithm. Feature selection has been a research topic with practical significance in many areas such as pattern recognition, statistics, machine learning. The objectives of feature selection include: building simpler and more intelligible models, improving data mining performance, and helping prepare, clean, and understand data. In this system, the presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm comprises 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and 4) selecting features. In the proposed algorithm, a cluster consists of features. Each cluster has very few feature and thus dimensionality is reduced. Then compared the performance of the proposed algorithm with those of the five familiar feature selection algorithms ReliefF, CFS, Consist, and FOCUS-SF on publicly available microarray, and text data from the different aspects of the proportion of selected features, classification accuracy of a given classifier.

REFERENCES

[1] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.
[2] Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol 3, pp. 1157-1182, 2003.
[3] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41,No. 2, pp. 175-195, 2000.
[4] J. Biesiada and W. Duch, "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in SoftComputing, vol. 45, pp. 242-249, 2008.
[5] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf.Data Mining, pp. 581-584, 2005.
[6] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.
[7] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.
[8] 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.
[9] Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," Proc. European Conf. Machine Learning, pp. 171-182, 1994.
[10] K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," Proc. 10th Nat'l Conf.Artificial Intelligence, pp. 129-134, 1992.
[11] 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.
[12] AlmuallimH. and. Dietterich T.G, "Algorithms for Identifying Relevant Features", Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-5, 1992.
[13] Asuncion, D. Newman, UCI machine learning repository, 2007.
[14] U. Fayyad, K. Irani, Multi-interval discretization of continuous-valued attri- butes for classification learning, in: Proceedings of Thirteenth International Joint Conference on Artificial Intelligence, 1993, pp. 1022–1027.
[15] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, vol. 1, Citeseer, 1995, pp. 338–345.
[16] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[17] E. Frank, I.H. Witten, Generating accurate rule sets without global optimiza- tion, in: Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 144–151.
[18] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Machine Learning 6 (1) (1991) 37–66.
[19] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, vol. 14, Citeseer, 1995, pp. 1137–1145.

