ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Survey on Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

Akshay S. Chavan
Department of Computer Science and Engineering
Nanded, Maharashtra, India.

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. 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.***

Keywords: Feature subset selection, filter method, feature clustering, graph-based clustering.

I. 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. 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 overfit 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.

II. LITERATURE SURVEY

Learning Boolean concepts in the presence of many irrelevant features, Artificial Intelligence

The use of feature selection can improve accuracy, efficiency, applicability and understand ability of a learning process. For this reason, many methods of automatic feature selection have been developed. Some of these methods are based on the search of the features that allows the data set to be considered consistent. In a search problem we usually evaluate the search states, in

the case of feature selection we measure the possible feature sets. This paper reviews the state of the art of consistency based feature selection methods, identifying the measures used for feature sets. An in-deep study of these measures is conducted, including the dentition of a new measure necessary for completeness. After that, we perform an empirical evaluation of the measures comparing them with the highly reputed wrapper approach. Consistency measures achieve similar results to those of the wrapper approach with much better efficiency.

We have presented a survey on the use of data set consistency measures for feature selection. To begin with, the feature selection problem and their main applications are reviewed. After that, based on a previous work on categorization of feature selection methods, we have introduced a modular decomposition of the feature selection process illustrating its relation with some well known methods. We hope this modular view can provide new views for researching in feature selection, as well as a skeleton for possible new methods. Then, our study is centered on the evaluation function—one of the modules of the decomposition—and more precisely on those measures based on consistency.

The state of the art of consistency measures for feature selection is reviewed, describing the three identified measures: the monotonic consistency measure proposed by Liu et al. (1998) for feature selection, the generic consistency measure from Rough Set Theory, and one measure defined from the ideas of some previous consistency based feature selection methods, that we consider necessary to define as a measure to fill a natural gap in this field. All these measures are carefully analyzed and compared, considering their properties and interpretation. We have identified their limit values and their use comparing data sets, revealing the relation between Liu's measure and the majority concept. We have also presented a review of other feature selection methods based on consistency as they are the basis of measures. Finally, an empirical evaluation of these measures and the wrapper approach has been performed, paying special attention to accuracy and reduction of the number of features.

We have shown that consistency measures can be very useful in many feature selection problems for the following reasons. First, they can achieve similar accuracy results than the wrapper approach, while being much more efficient. Second, they can achieve a higher feature reduction. And finally, being independent of the classifier used, they may be more practical in some circumstances, for example using various algorithms on the same problem, or assessing experts. For these reasons, we can conclude that the use of the filter approach for feature selection is an interesting choice. When efficiency is a requirement, the wrapper approach is usually not suitable, but the filter approach with consistency measures is your choice. Moreover, even in situations where the wrapper approach could be used, the filter approach can render more accurate results.

The three consistency measures compared achieve pretty similar results, thus making a choice among them is difficult. In case we are interested in a high feature reduction for a classification problem, we may choose Inconsistent Example Pairs measure, while if we are interested in maximal accuracy Liu's measure may be a better choice. As the three measures are very efficient, it also possible to apply all of them and to take the one which fits best to our problem, probably in the same time that it would take to run other measures. The results suggest that consideration of continuous features and regression problems deserve a deeper study to improve accuracy, because while the consistency measures provide a much more efficient way of selecting features than the wrapper approach, the accuracy is slightly worse using the former approach. Finally, there is an open field of research in the combination of feature selection and discretization.

A feature set measure based on relief

Feature subset selection is an important subject when training classifiers in Machine Learning (ML) problems. Too many input features in a ML problem may lead to the so-called "curse of dimensionality", which describes the fact that the complexity of the classifier parameters adjustment during training increases exponentially with the number of features. Thus, ML algorithms are known to suffer from important decrease of the prediction accuracy when faced with many features that are not necessary. In this paper, we introduce a novel embedded feature selection method, called ESFS, which is inspired from the wrapper method SFS since it relies on the simple principle to add incrementally most relevant features. Its originality concerns

the use of mass functions from the evidence theory that allows to merge elegantly the information carried by features, in an embedded way, and so leading to a lower computational cost than original SFS. This approach has successfully been applied to the domain of image categorization and has shown its effectiveness through the comparison with other feature selection methods.

Features election for high-dimensional data Pearson redundancy based filter

Selection of Feature subset is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Numerous feature subset selection methods have been planned and considered for machine learning applications. Feature subset selection can be analysed 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. We build up a novel algorithm that can capably and efficiently deal with both inappropriate and redundant characteristics, and get hold of a superior feature subset. 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. 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.

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 analysed 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.

Using decision trees to improve case-based learning

This paper addresses the problem of han- dling skewed class distributions within the case-based learning(CBL)framework.We first present as a baseline an informationgain-weighted CBL algorithm and apply it to three data sets from natural language processing (NLP) with skewed class distributions. Although overall performance of the baseline CBL algorithm is good, we show that the algorithm exhibits poor performance on minority class instances. We then present two CBL algorithms designed to improve the performance of minority class predictions. Each variation creates test-case-specic feature weights by rst observing the path taken by the test case in a decision tree created for the learning task, and then using pathspecic information gain values to create an appropriate weight vector for use during case retrieval. When applied to the NLP datasets, the algorithms are shown to signicantly increase the accuracy of minority class pre- dictions while maintaining or improving overall classication accuracy.

Insummary, we have investigated the use of test-casespecic feature weighting to aid in the recovery of mi- nority class instances in skewed class distributions. We presented two case-based learning algorithms that use decision trees to create the case-specic weight vectors. Each variation composes the vector in two stages. In a feature selection stage, the path that would be taken by the test case in a decision tree created for the learning task is noted. Any feature tested along the path is included in the case representation; all other features are ignored. Weights for the selected features are then determined using path-specic

information gain values. On three data sets with skewed distri- butions from the natural language processing domain,the algorithms are shown to signicantly increase the accuracy of minority class predictions while maintaining or improving overall classication accuracy. Given our initial results, we believe that this is a promising approach for dealing with skewed distributions in other domains.

Mining of Attribute Interactions Using Information Theoretic Metrics

Feature interaction is an important issue in feature subset selection. However, most of the existing algorithms only focus on dealing with irrelevant and redundant features. In this paper, a propositional FOIL rule based algorithm FRFS, which not only retains relevant features and excludes irrelevant and redundant ones but also considers feature interaction, is proposed for selecting feature subset for high dimensional data. FRFS merges the features appeared in the antecedents of all FOIL rules, achieving a candidate feature subset which excludes redundant features and reserves interactive ones. Then, it identifies and removes irrelevant features by evaluating features in the candidate feature subset with a new metric Cover Ratio, and obtains the final feature subset. The efficiency and effectiveness of FRFS are extensively tested upon both synthetic and real world data sets, and it is compared with other six representative feature subset selection algorithms, including CFS, FCBF, Consistency, Relief-F, INTERACT, and the rule-based FSBAR, in terms of the number of selected features, runtime and the classification accuracies of the four well-known classifiers including Naive Bayes, C4.5, PART and IB1 before and after feature selection. The results on the five synthetic data sets show that FRFS can effectively identify irrelevant and redundant features while reserving interactive ones.

Feature Selection for Classification

Feature selection has been the focus of interest for quite some time and much work has been done. With the creation of huge databases and the consequent requirements for good machine learning techniques, new problems arise and novel approaches to feature selection are in demand. This survey is a comprehensive overview of many existing methods from the 1970's to the present. It identifies four steps of a typical feature selection method, and categorizes the different existing methods in terms of generation procedures and evaluation functions, and reveals hitherto unattempted combinations of generation procedures and evaluation functions. Representative methods are chosen from each category for detailed explanation and discussion via example. Benchmark datasets with different characteristics are used for comparative study. The strengths and weaknesses of different methods are explained. Guidelines for applying feature selection methods are given based on data types and domain characteristics.

Statistical comparison of classifiers over multiple data sets

While methods for comparing two learning algorithms on a single data set have been scrutinized for quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple data sets, which is even more essential to typical machine learning studies, has been all but ignored. This article reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly introduced CD (critical difference) diagrams

III. Feature Subset Selection Framework And Algorithm

Framework

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."

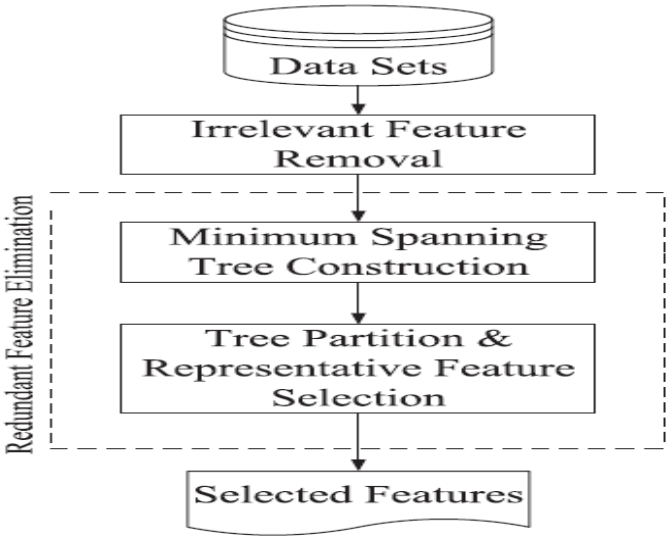

Fig. 1: Framework of the proposed feature subset selection algorithm

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 re-dundant feature elimination, we firstly present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation as follows. presented a definition of relevant features. Suppose F to be the full set of features, $F \in F$be a feature, $S\ i = F$ -{$F\ i$ } and $S'\ i \subseteq S\ i$. Let s'be a valueassignment of all features in $S'i$, fii a value-assignment of feature F, and c a value-assignment of the target concept C.

The proposed FAST algorithm logically consists of tree steps: (i) removing irrelevant features, (ii) constructing a MST from relative ones, and (iii) partitioning the MST and selecting representative features. For a data set D with m features F = {F1,F2...Fm} and class C, we compute the T-Relevance $SU(F$i,$C)$ value for each feature Fi (1 <= i< = m) in the first stepThe features whose $SU(F$i,$C)$ values are greater than a predefined threshold θ comprise the target-relevant feature subset F'={F'1,F'2, ...F'k}(k<=m).

In the second step, we first calculate the F-Correlation ($F'i$,'j) value for each pair of feature F'i and F'j ($F'i$,'$j \in F' \wedge i$ / = j). Then, viewing features F'iand F'j as vertices and $SU(F$'i,F'j)(i≠j) as the weight of the / edge between vertices $F'i$ and F'j, a weighted complete graph G =(V, E) is constructed where V = {F'i | F'i ∈ F' ⋀ i ∈ [1,K]} and E={(F'i,F'j) | (F'i,F'j ∈ F' ⋀ i, j ∈ [1,k] ⋀ i≠j}.As symmetric uncertainty is symmetric further the F-correlation $SU(F$'i,F'j) is symmetric as well, thus G is an undirected graph.

The complete graph G reflects the correlations among all the target-relevant features. Unfortunately, graph G has k vertices and (k-1)/2 edges. For high dimensional data, it is heavily dense and the edges with different weights are strongly interweaved. Moreover, the decomposition of complete graph is NP-hard. Thus for graph G, we build a MST, which connects all vertices

such that the sum of the weights of the edges is the minimum, using the well-known Prim algorithm. The weight of edge ($F'i,'j$) is F-Correlation (F'i,F'j).

After building the MST, in the third step, we first remove the edges $E = \{(F'i,'j) \mid (F'i,'j \in F \wedge i, j \in [1,k] \wedge i \neq j\}$, whose weights are smaller than both of the T-Relevance $SU(F'i,C)$ and $SU(F,C)$, from the MST.Each deletion results in two disconnected trees T'1 and T2.

Assuming the set of vertices in any one of the final trees to be $V(T)$, we have the property that for each pair of vertices ($F'i,'j \in V(T)$), $(F'i,'j) = (F,C) \vee SU(F,C)$ always holds.

Algorithm and analysis:

Algorithm 1: FAST

inputs: $D(F_1, F_2, ..., F_m, C)$ - the given data set
θ - the T-Relevance threshold.
output: S - selected feature subset .
//==== Part 1 : Irrelevant Feature Removal ====
1 **for** $i = 1$ to m **do**
2 T-Relevance = SU (F_i, C)
3 **if** T-Relevance $> \theta$ **then**
4 $S = S \cup \{F_i\}$;
//==== Part 2 : Minimum Spanning Tree Construction ====
5 G = NULL; //G is a complete graph
6 **for** each pair of features $\{F'_i, F'_j\} \subset S$ **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
//==== Part 3 : Tree Partition and Representative Feature Selection ====
10 Forest = minSpanTree
11 **for** each edge $E_{ij} \in$ Forest **do**
12 **if** $SU(F'_i, F'_j) < SU(F'_i, C) \wedge SU(F'_i, F'_j) < SU(F'_j, C)$ **then**
13 Forest = Forest $- E_{ij}$
14 $S = \phi$
15 **for** each tree $T_i \in$ Forest **do**
16 $F^j_R = \text{argmax}_{F'_k \in T_i} SU(F'_k, C)$
17 $S = S \cup \{F^j_R\}$;
18 return S

Removal of Irrelevant features:

An effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed for machine learning applications. if we take a Dataset 'D' with m features F={F1,F2,..,Fn} and class C, automatically features are available with target relevant feature. The generality of the selected features is limited and the computational complexity is large. 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.

T-Relevance, F-correlation calculation:

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. According to the

above definitions, 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

1. Irrelevant features have no/weak correlation with target concept.

2. Redundant features are assembled in a cluster and a representative feature can be taken out of the cluster.

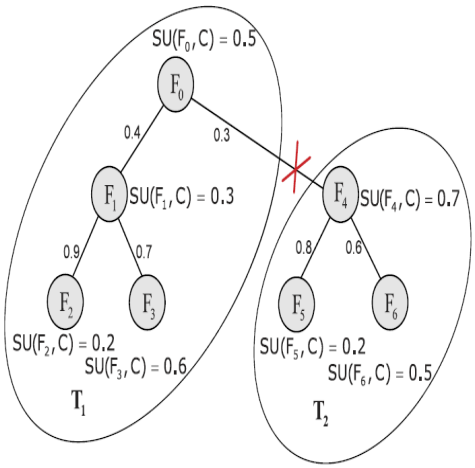

Fig.2 Minimum Spanning Tree

MST construction

To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) 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. We construct a Minimal spanning tree with weights. a MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well-known Prim algorithm.

Relevant feature calculation:

After tree partition unnecessary edges are removed. each deletion results in two disconnected trees(T1,T2).After removing all the unnecessary edges, a forest Forest is obtained. Each tree represents a cluster. Finally it comprises for final feature subset. then calculate the accurate/relevant feature.

IV. CONCLUSION

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

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.

We also found that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy of the four different types of classifiers, and CFS is a good alternative. At the same time, FCBF is a good alternative for image and text data. Moreover, Consist and FOCUSSF are alternatives for text data.

References

1. 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.
2. Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249,2008.
3. Cardie, C., Using decision trees to improve case-based learning, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.
4. Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.
5. Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009..
6. Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
7. Demsar J., Statistical comparison of classifiers over multiple data sets, J. Mach. Learn. Res., 7, pp 1-30, 2006.

AUTHOR(S) PROFILE

Akshay S. chavan , received BE in information and technology and pursuing master of engineering in computer science &engineering. Asst. lecturer in MPGI school of engineering Nanded.

