Fast Feature subset selection algorithm based on clustering for high dimensional data

Mrs. Komal Kate[1], Prof. Asha Pawar[2]

[1]PG Scholar, Department of Computer Engineering, ZES COER, pune, Maharashtra

[2]Assistant Professor, Department of Computer Engineering, ZES COER, pune, Maharashtra

Abstract— Feature selection algorithm can be used for removing irrelevant, redundant information from the data. Feature selection is divided into different categories, amongst them filter method is used because of its generality and is typically good choice when numbers of features are large. In cluster analysis, features are divided by using graph-theoretic clustering method. A Fast clustering bAsed feature Selection algoriThm (FAST) is based on minimum spanning tree which is constructed from weight complete graph. Then features are divided into forest where each tree represents a cluster and most representative feature is selected from each cluster. The clustering-based FAST algorithm produces a subset which contains independent features. Linear correlation based measure is used to identify highly correlated features amongst the final got subset of features from FAST algorithm.

Keywords—Graph-theoretic clustering; Minimum spanning tree; Feature selection; feature subset selection algorithm (FAST); High dimensional data; Filter method; Wrapper method; Embedded method.

INTRODUCTION

In data mining, data is analysed and summarized into useful information. The high-throughput technologies grow increasingly which result in exponential growth in the data with respect to dimensionality, storing and processing such a data becomes more challenging task. Feature selection is used as a pre-processing stage in machine learning. It is a procedure of choosing a subset of original features by removing irrelevant and redundant features so that the feature space is get reduced according to some criterion. For choosing a subset of good features with respect to the target concepts, feature subset selection is a useful way for reduction in dimensionality, removal of irrelevant data, increase in learning accuracy, and improvement in result comprehensibility. Feature selection is generally categorized into four models, namely: filter model, wrapper model, embedded model and hybrid model. Filter model methods do not make use of any clustering algorithm to test the quality of the features. They evaluate the value of each feature according to certain criteria. Then, it selects the features with the highest value. It is called the filter because it filters out the irrelevant features using given criteria. The wrapper model uses a clustering algorithm to evaluate the goodness of selected features by (1) finding a subset of features and then, (2) it evaluates the clustering quality using the selected subset. Finally, it repeats (1) and (2) until the desired quality of features is found. It is impossible to evaluate all possible subsets of features in high-dimensional datasets. Therefore, experience based search strategy is adopted to reduce the search space. The wrapper model is computationally expensive compared to filter model. However, it produces better clustering because it aims to select features that maximize the quality. The embedded methods integrate feature selection as a part of the training process and are usually specific to given learning algorithms, and thus may be more efficient than the other three categories. Drawbacks in filter and wrapper models are overcome in a hybrid model, which take benefit from the efficient filtering criteria and better clustering quality from the wrapper model. A hybrid process having following steps: it uses filtering criteria to select candidate subsets. Then, the quality of clustering of each candidate subsets is evaluated and the subset with highest clustering quality will be selected. Hybrid model usually produce better clustering quality than those of filter model and wrapper model.

Feature selection selects subset of highly differentiated features. In other words, it selects features that are capable of selecting samples that belong to different classes. Thus, if we have labelled samples as training samples in order to select these features, then this kind of learning is called supervised learning, which means that the dataset is labelled. In supervised learning, it is easy to differentiate the features in different classes. If sample data is unlabeled then selecting feature poses a challenge in feature selection task. In such cases, defining relevancy becomes unclear. However, we still consider that selecting subset(s) of features may help improving unsupervised learning in a same way to improving the supervised learning.

Feature selection algorithm used to select the subset of feature by removing irrelevant and redundant features. FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the minimum spanning tree into a forest such that each tree representing a cluster; and 3) and then the selection of representative features from the clusters[3].

In general, a good features are those which relevant to target class but not redundant to any of the other relevant features. Correlation is considered as goodness measure between two variables, then the above definition becomes that a good features are highly correlated to the target class but not highly correlated to any of the other features. In other words, if the correlation between a feature and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other relevant features, it will be regarded as a good feature for the classification task [2].

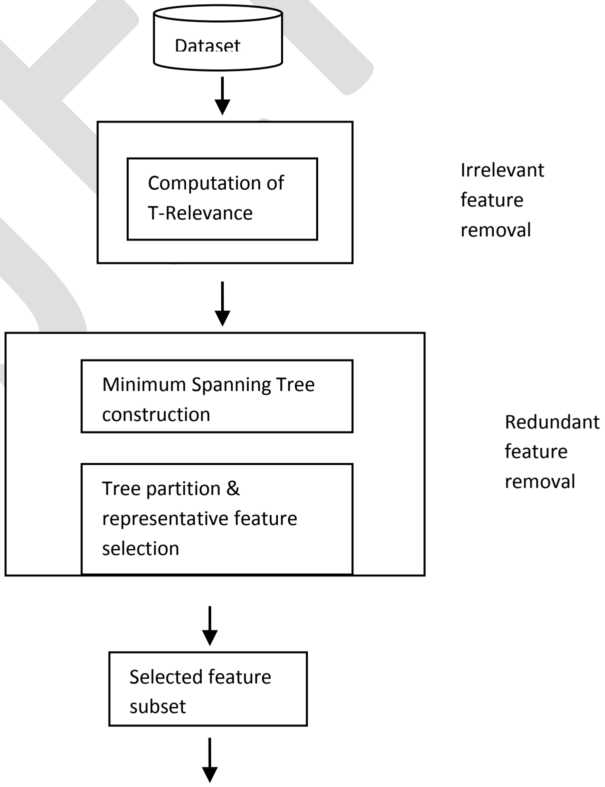
RELATED WORK

Feature subset selection can be the process of identifying and removing irrelevant that do not contribute to predictive accuracy and redundant features that they provide mostly information which is already present in other features as much as possible.

Many feature selection algorithms are present; some of them can be able to remove irrelevant features but not effective to handle redundant features. Yet some of the other can eliminate irrelevant feature while taking care of redundant features [1]. FAST algorithm falls in to second group. One of the feature selection algorithms is Relief [6], which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is useless at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted [7]. Relief-F [8] extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features. Redundant features also affect the accuracy and speed of learning algorithm; hence it is necessary to remove it. CFS [9], FCBF [10], and CMIM [12] are examples that take into consideration the redundant features. CFS [9] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. FCBF ([10], 11]) is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. CMIM [12] iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from above algorithms, FAST algorithm uses minimum spanning tree-based method to cluster features.

FEATURE SUBSET SELECTION ALGORITHM

Feature subset selection framework consists of two important component irrelevant feature removal and redundant feature removal. Accuracy of learning machines severely affected by irrelevant features, along with redundant features. Thus, irrelevant and redundant features should be identify and remove as much as possible. Moreover, “good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.” [5] A new algorithm is developed to efficiently and effectively deal with both irrelevant and redundant features, to obtain a good feature subset. A new feature selection framework (shown in Fig.1).


Dataset
Computation of T-Relevance
Irrelevant feature removal
Minimum Spanning Tree construction
Tree partition & representative feature selection
Redundant feature removal
Selected feature subset

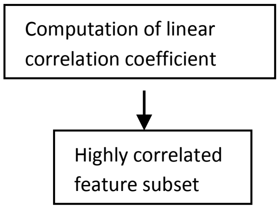

Computation of linear correlation coefficient
Highly correlated feature subset

Fig. 1. Framework of feature subset selection

The former obtains features by eliminating irrelevant ones and relevant to the target concept , and the latter choosing representative feature from different clusters by removing redundant features from relevant ones , and thus produces the final feature subset. In 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 [1].

Symmetric Uncertainty

The mutual information (MI) measures the amount of information that feature variable has about target class. This is a nonlinear evaluation of correlation between feature values or feature values and target classes. The symmetric uncertainty (SU) [20] 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 assess the goodness of features for classification by a number of researchers.

The symmetric uncertainty is defined as follows

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

Where,

$$Gain(X|Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(X|Y)$$

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

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

Where, p(x) is the probability density function and p (x|y) is the conditional probability density function.

T-Relevance

Relevance between the feature F_i ∈ 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). F_i is a strong T-Relevance feature only when SU (F_i,C) is greater than a predetermined threshold,. After finding the relevance value, the redundant features will be removed with respect to the threshold value.

F-Correlation

The correlation between any pair of features F_i and F_j (F_i, F_j ∈ $F^{\wedge}{}_{i} \neq j$) is called the F-Correlation of F_i and F_j, and denoted by SU(F_i, F_j). The same equation of symmetric uncertainty which is used for finding the relevance between the feature and the target class is again applied to find the similarity between two attributes with respect to each label.

IJERGS

Minimum spanning tree

The complete graph G shows the correlations among all the target-relevant features and it has v vertices and $v(v-1)/2$ edges. In case of high dimensional data, it is heavily dense and the edges are strongly interweaved with different weights. Thus for graph G, minimum spanning tree is constructed, which connects all vertices such that the sum of the weights of the edges is the minimum, using Prim algorithm. The weight of edge (F'_i, F'_j) is F-Correlation $\mathcal{S}(F'_i, F'_j)$. After building the MST, we first remove the edges E, whose weights are smaller than both of the T-Relevance (F_i, C) and (F_j, C), from the MST. Each deletion of edge results in two disconnected trees T 1 and T 2.

This can be illustrated by an example. Suppose fig.2 shows MST which is generated from complete graph. We first travel all the edges then decide to remove the edge (F0, F4) because its weight SU (F0, F4) = 0.2 is smaller than both SU (F0, C) = 0.3 and SU(F4, C)= 0.4. This makes the MST is clustered into two clusters. Each cluster contains relatively independent features. Then representative features from each cluster selected to form a subset of features. FAST algorithm is shown in algorithm1.

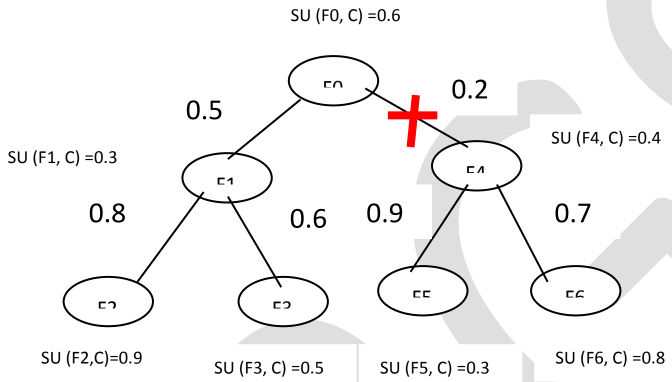


Fig. 2. Example of clustering

ALGORITHM 1: FAST

```
Inputs: S ($F_1$, $F_2$, …, $F_n$, C) - the given data set
        θ- The T-Relevance threshold.
Output: X- selected feature subset.
//------ Irrelevant Feature Removal ------

for i=1 to n do
   T-Relevance= SU ($F_i$, C)
   if T-Relevance >θ then
   S= S U {$F_i$};

//------ Minimum spanning tree construction-----

G= NULL; // G is a complete graph
for each pair of features {$F'_i$, $F'_j$} ⊂ X do
     F-Correlation = SU ($F'_i$, $F'_j$)
     Add $F'_i$ and/ or $F'_j$ to G with F-Correlation as the weight of the corresponding edge;
 minSpanTree = Prim(G); //Using Prim Algorithm to generate the minimum spanning tree

//-------Tree Partition and Representation Feature Selection----

Forest= minSpanTree
for each edge $E_{ij}$ ϵ Forest do
    if SU($F'_i$, $F'_j$) < SU($F'_i$, C) ∧ SU($F'_i$, $F'_j$) < SU($F'_j$, C) then
           Forest= Forest- $E_{ij}$
X= ϕ
for each tree $T_i$ ϵ Forest do
     $F^j_r$ = argmax $F'_k$ ϵ $T_i$ SU($F'_k$, C)
     X= X U { $F^j_r$};
```

```
return X
```

CORRELATION BASED MEASURE

Generally, good feature is relevant to the class concept, but not redundant to other features. We use correlation between any two features to measure the goodness, hence above statement becomes that a feature is good if it is highly correlated to the class but not highly correlated to any of the other features. Classical linear correlation is one of the approach exist. Linear correlation coefficient can be calculated for selected feature subset to obtain highly correlated features. For a pair of variables (P, Q), the linear correlation coefficient r is given by the formula

Where p_i is the mean of P, and q_i is the mean of Q. The value of r lies between -1 and 1, inclusive. If P and Q are completely correlated, r takes the value of 1 or -1; if X and Y are totally independent, r is zero. It is a symmetrical measure for two variables. There are some benefits of linear correlation as a feature goodness measure for classification. Linear correlation helps to remove features with near zero linear correlation to the class concept. It helps to reduce redundancy among selected features. It is known that if data is linearly separable in the original representation, it is still linearly separable if all but one of a group of linearly dependent features are removed (Das, 1971).

DATA SOURCE

With the aim of evaluating the performance and efficiency of FAST algorithm, verifying whether or not the method is potentially useful in practice, some publicly available data sets were used. These data sets cover application domain such as text data classification.

EXPERIMENTAL SETUP

To evaluate the performance of FAST algorithm, set up for performance of the feature subset selection algorithms, three metrics, (i) the proportion of selected features (ii) the time to obtain the feature subset, (iii) the classification accuracy, and (iv) the Win/Draw/Loss record [20], are used. The proportion of selected features is the ratio of the number of features selected by a feature selection algorithm to the original number of features of a data set [1]. The Win/Draw/Loss record shows three values on a given measure, i.e. the numbers of data sets for which FAST algorithm obtains better, equal, and worse performance. The measure can be the proportion of selected features, the runtime to obtain a feature subset, and the classification accuracy, respectively.

RESULT

A FAST feature selection algorithm requires a parameter θ that is the threshold of feature relevance. Different θ values might obtain different classification results. When determining the value of θ, classification accuracy, the proportion of the selected features also considered. Because unacceptable proportion of the selected features results in a large number of features are taken, and further affects the classification efficiency.

Fast Clustering-Based Feature Subset Selection

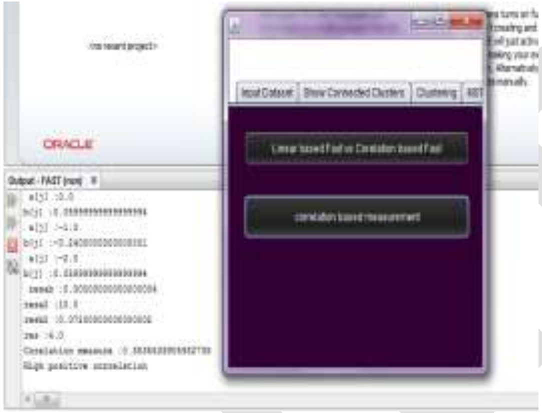
Fig. 3. Input Dataset

Fast Clustering-Based Feature Subset Selection

Fig. 4. Irrelevant Feature Removal

Fig. 5. Minimum Spanning Tree

Fig. 6. Final Feature Subset



```
 a[j] :0.0
b[j] :0.[illegible]
 a[j] :-1.0
b[j] :-0.2400000000000[illegible]
 a[j] :-2.0
b[j] :0.0[illegible]
 sumab :0.3000000000000004
sumaa :10.0
sumbb :0.0716000000000002
rss :4.0
Correlation measure :0.[illegible]
High positive correlation
```

Fig. 7.Correlation Based Measure

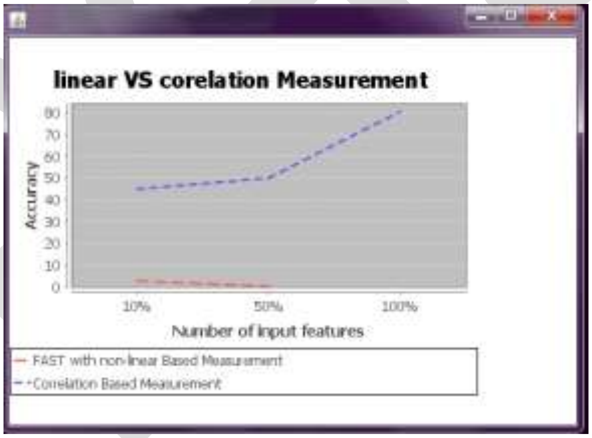


Fig. 8. Non Linear Vs Linear Measure

Conclusion

FAST subset selection algorithm based on clustering contains three important steps: Removal of irrelevant features that do not contribute to predictive accuracy; Elimination of Redundant features using minimum spanning tree; partitioning the MST into clusters and collect the selected features. Each cluster consists of redundant features and which is treated as single feature, which result in significant reduction in dimensionality. The FAST algorithm can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. To obtain highly correlated features linear correlation coefficient is used, which identify highly correlated features with corresponding class. Performance of FAST algorithm is compared on different datasets for text data with three different aspects of proportion of selected features, runtime and classification accuracy. For future work, linked data can be use as input and algorithm can be deal with unsupervised leaning.

REFERENCES:

1. Qinbao Song, Jingjie Ni and Guangtao Wang "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data" IEEE transactions on knowledge and data engineering vol:25 no:1 year 2013
2. Yu L. and Liu H.," Feature selection for high-dimensional data: a fast correlation-based filter solution", in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.
3. Komal Kate, Prof S. D. Potdukhe,"Fast Feature subset selection algorithm based on clustering for high dimensional data", International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014

4. Jiliang Tang,Huan Liu, "An Unsupervised Feature Selection Framework for Social Media Data" , In Proceeding of the 18th ACM SIGKDD international conference on knowlegde discovery and data mining, pp 904-912, 2012.

5. Hall M.A. and Smith L.A., "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper", In Proceedings of the Twelfth international Florida Artificial intelligence Research Society

6. Kira K. and Rendell L.A.," The feature selection problem: Traditional methods and a new algorithm", In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.

7. Koller D. and Sahami M., "Toward optimal feature selection", In Proceedings of International Conference on Machine Learning, pp 284-292, 1996.

8. Kononenko I.," Estimating Attributes: Analysis and Extensions of RELIEF", In Proceedings of the 1994 European Conference on Machine Learning, pp171-182, 1994.

9. Hall M.A., "Correlation-Based Feature Subset Selection for Machine Learning", Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.

10. Yu L. and Liu H., "Feature selection for high-dimensional data: a fast correlation-based filter solution", in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.

11. Yu L. and Liu H., "Efficient feature selection via analysis of relevance and redundancy", Journal of Machine Learning Research, 10(5), pp 1205-1224, 2004.

12. Fleuret F., "Fast binary feature selection with conditional mutual Information",Journal of Machine Learning Research, 5, pp 1531-1555, 2004.

13. Hall M.A. and Smith L.A., "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper", In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.

14. Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., "Numerical recipes in C". Cambridge University Press, Cambridge, 1988.

15. Zhao Z. and Liu H., "Searching for interacting features", In Proceedings of the 20th International Joint Conference on AI, 2007.

16. Zhao Z. and Liu H.," Searching for Interacting Features in Subset Selection", Journal Intelligent Data Analysis, 13(2), pp 207-228, 2009.

17. Prim R.C., "Shortest connection networks and some generalizations", Bell System Technical Journal, 36, pp 1389-1401, 1957.

18. Garey M.R. and Johnson D.S., "Computers and Intractability: a Guide to the Theory of Np-Completeness". W. H. Freeman & Co, 1979.

19. 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.

20. Webb G.I., "Multiboosting: A technique for combining boosting and Wagging", Machine Learning, 40(2), pp 159-196, 2000.

21. Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., Numerical recipes in C. Cambridge University Press, Cambridge, 1988

IJERGS