International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

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

Mrs. Komal Kate[1], Prof. S. D. Potdukhe[2]

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

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

Abstract— A Feature selection algorithm employ for removing irrelevant, redundant information from the data. Amongst feature subset selection algorithm filter methods are used because of its generality and are usually good choice when numbers of features are large. In cluster analysis, graph-theoretic clustering methods to features are used. In particular, the minimum spanning tree (MST)-based clustering algorithms are adopted. A Fast clustering bAsed feature Selection algoriThm (FAST) is based on MST method. In the FAST algorithm, features are divided into clusters by using graph-theoretic clustering methods and then, the most representative feature that is strongly related to target classes is selected. Features in different clusters are relatively independent. A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. Traditionally, feature subset selection research has focused on searching for relevant features. The clustering-based strategy of FAST having a high probability of producing a subset of useful and independent features.

Keywords— Cluster analysis, Graph-theoretic clustering, Minimum spanning tree, Feature selection, feature subset selection algorithm (FAST), High dimensional data, Filter method.

INTRODUCTION

Data mining is a process of analyzing data and summarizes it into useful information. In order to achieve successful data mining, feature selection is an essential component. In machine learning feature selection is also known as variable selection or attributes selection. The main idea of feature selection is to choose a subset of features by eliminating irrelevant or no predictive information. It is a process of selecting a subset of original features according to specific criteria. Feature selection is an important and frequently used technique in data mining for dimension reduction. It employ for removing irrelevant, redundant information from the data to speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility. Supervised, unsupervised and semi-supervised feature selection algorithms are developed as result of process of feature selection algorithm. A supervised feature selection algorithm determines features' relevance by evaluating their correlation with the class or their utility for achieving accurate prediction, and without labels, an unsupervised feature selection algorithm may exploit data variance or data distribution in its evaluation of features' relevance and a semi-supervised feature selection algorithm uses a small amount of labelled data as additional information to improve unsupervised feature selection [2].

Feature subset selection methods can be divided into four major categories: Embedded, Wrapper, Filter, and Hybrid. The embedded methods has a feature selections as a part of the training process and are usually specific to given learning algorithms, and thus possibly more efficient than the other three categories. Machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. Wrapper methods assess subsets of variables according to their relevance to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. Filter methods are pre-processing methods. They attempt to assess the useful features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by ranking them through compression techniques or by computing correlation with the output. 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. The important part of hybrid method is combination of filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods [1].

In cluster analysis, graph theoretic approach is used in many applications. In general graph-theoretic clustering a complete graph is formed by connecting each instance with all its neighbours. Zahn's clustering Algorithm

1. Construct the MST for the set of n patterns given.
2. Identify inconsistent edges in MST.
3. Remove the inconsistent edges to form connected components and call them clusters.

In the FAST algorithm, features are divided into clusters by using graph-theoretic clustering methods and then, the most representative feature that is strongly related to target classes is selected. Features in different clusters are relatively independent. A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. Traditionally, feature subset selection research has focused on searching for relevant features. The clustering-based strategy of FAST

having a high probability of producing a subset of useful and independent features.

RELATED WORK

Feature selection is aim at a choosing a subset of features by eliminating irrelevant or no predictive information. It is a process of selecting a subset of original features according to specific criteria. Irrelevant features do not contribute to the accuracy and redundant features mostly provide the information which is already present in other features.

There are many feature selection algorithm present, some of them are useful at removing 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 [3], 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 [4]. Relief-F [5] 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 [6], FCBF [7], and CMIM [9] are examples that take into consideration the redundant features. CFS [6] 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 ([7], [8]) is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. CMIM [9] 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

FRAMEWORK

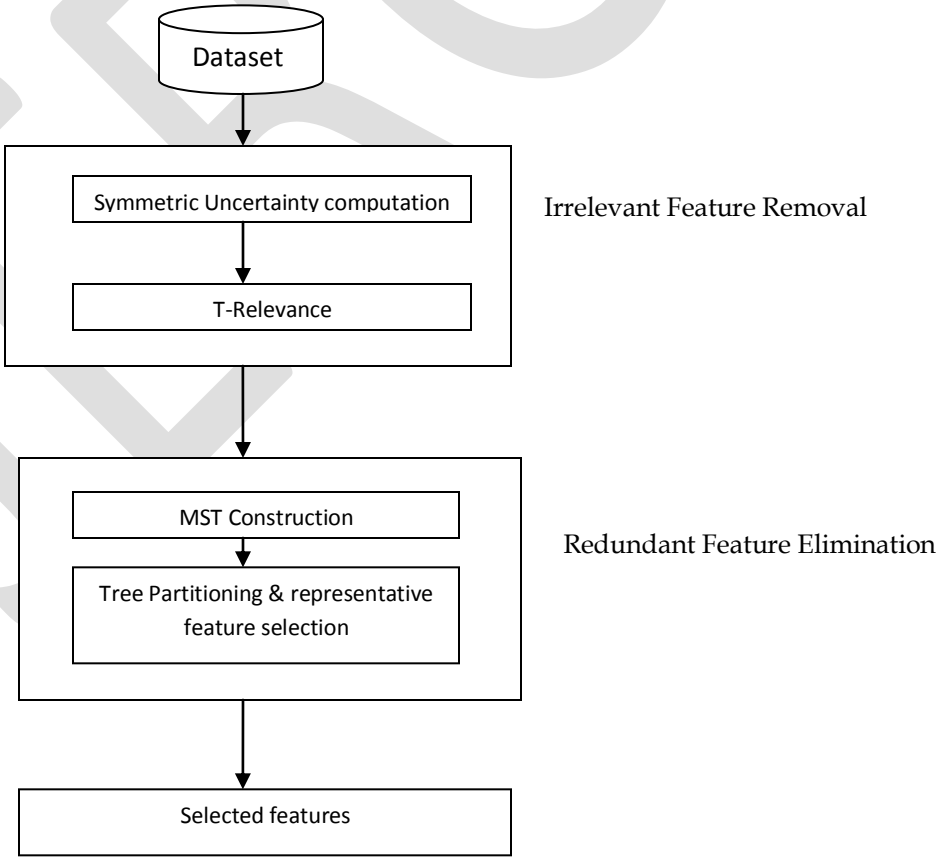

Fig.1. Framework of feature subset selection algorithm

Feature subset selection algorithms are aim at identifying and removing irrelevant and redundant features as much as

possible. "*Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other*." [10]

Feature selection framework can deal with effectively and efficiently deal with irrelevant and redundant features. It is made up of 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 [1].

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.

Relevant features have strong correlation with target concept hence they are always needed for a best subset, while redundant features are not needed because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally defined in terms of feature correlation and feature-target concept correlation.

SYMMETRIC UNCERTAINTY

Mutual information majors how much feature values and target classes differ from each other. This is nonlinear estimation of correlation between feature values or feature values and target classes [1]. The symmetric uncertainty (SU) [11] 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 (e.g., Hall [6], Hall and Smith [10], Yu and Liu [7], [8], Zhao and Liu [12], [13]).

The symmetric uncertainty is defined as follows

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

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

The relevance between the feature Fi ∈ 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, then Fi is a strong T-Relevance feature. After finding the relevance value, the redundant attributes will be removed with respect to the threshold value.

F-CORRELATION

The correlation between any pair of features Fi and Fj (Fi,Fj ∈ F^ i ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj). 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.

MINIMUM SPANNING TREE

Viewing features Fi and Fj as vertices and (Fi, Fj) ($i \neq j$) as the weight of the edge between vertices Fi and , a weighted complete graph G = (V,E) is constructed. As symmetric uncertainty is symmetric further the F-Correlation (Fi, Fj) 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 interwoven. Moreover, the decomposition of complete graph is NP-hard [15]. 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 [14]. The weight of edge (Fi, Fj) is F-Correlation (Fi, Fj). After building the MST, we first remove the edges E, whose weights are smaller than both of the T-Relevance (Fi, C) and (Fj, C), from the MST. Each deletion results in two disconnected trees $T1$ and $T2$.

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)= 0.6 and SU(F4, C)= 0.7. This makes the MST is clustered into two clusters. The details of the FAST algorithm are shown in algorithm1.

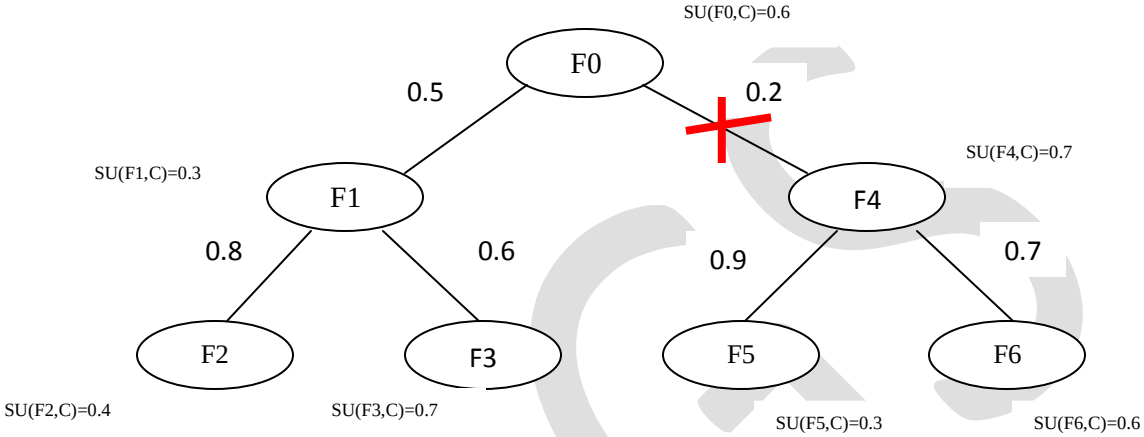

Fig.2. Example of clustering

ALGORITHM 1: FAST

Inputs: D(F1, F2, ... Fm, C)- the given data set
θ- the T-Relevance threshold.
Output: S- selected feature subset.
//-------Part1: Irrelevant Feature Removal ------

```
1    for i=1 to m do
2       T-Relevance= SU(Fi, C)
3       if T-Relevance >θ then
4       S= S U {Fi};

//------Part2: Minimum spanning tree construction-----

5    G= NULL; // G is a complete graph
6    for each pair of features {F'i, F'j} ⊂ 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

//-------Part3: Tree Partition and Representation Feature Selection------

10    Forest= minSpanTree
11    for each edge Eij ϵ Forest do
12          if SU(F'i, F'j) < SU(F'i, C) ∧ SU(F'i, F'j) < SU(F'j, C) then
13                Forest= Forest- Eij
14    S= ϕ
15    for each tree Ti ϵ Forest do
16          F_jr = argmax F'k ϵ Ti SU(F'k, C)
17          S= S U { F_jr};
18    return S
```

TIME COMPLEXITY

The first part of the algorithm has a linear time complexity (m) in terms of the number of features m. When $1 < k \leq m$, the second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is $O(k2)$, and then generates a MST from the graph using Prim algorithm whose time complexity is $O(k2)$. The third part partitions the MST and chooses the representative features with the complexity of (k). Thus when $1 < k \leq m$, the complexity of the algorithm is (m+$k2$).

CONCLUSION

FAST cluster based subset selection algorithm involves three important steps: 1.Removal of irrelevant features. 2. Elimination of Redundant features using minimum spanning tree. 3. Partitioning the MST and collect the selected features. Each cluster consists of redundant features and which is treated as single feature, so that dimensionality is reduced. A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. The clustering-based strategy of FAST produces a subset of useful and independent features. The FAST algorithm can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

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] Almuallim H. and Dietterich T.G., "Algorithms for Identifying Relevant Features", In Proceedings of the 9th Canadian Conference on AI, pp 38-45,1992.
[3] 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.
[4] Koller D. and Sahami M., "Toward optimal feature selection", In Proceedings of International Conference on Machine Learning, pp 284-292, 1996.
[5] Kononenko I.," Estimating Attributes: Analysis and Extensions of RELIEF", In Proceedings of the 1994 European Conference on Machine Learning, pp171-182, 1994.
[6] Hall M.A., "Correlation-Based Feature Subset Selection for Machine Learning", Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.
[7] 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.
[8] 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.
[9] Fleuret F., "Fast binary feature selection with conditional mutual Information",Journal of Machine Learning Research, 5, pp 1531-1555, 2004.
[10] 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.
[11] Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., "Numerical recipes in C". Cambridge University Press, Cambridge, 1988.
[12] Zhao Z. and Liu H., "Searching for interacting features", In Proceedings of the 20th International Joint Conference on AI, 2007.
[13] Zhao Z. and Liu H.," Searching for Interacting Features in Subset Selection", Journal Intelligent Data Analysis, 13(2), pp 207-228, 2009.
[14] Prim R.C., "Shortest connection networks and some generalizations", Bell System Technical Journal, 36, pp 1389-1401, 1957.
[15] Garey M.R. and Johnson D.S., "Computers and Intractability: a Guide to the Theory of Np-Completeness". W. H. Freeman & Co, 1979.
[16] 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.
[17] 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
[18] 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.
[19] Battiti R., "Using mutual information for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.

[20] 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.
[21] Biesiada J. and Duch W., "Features election for high-dimensional datała Pearson redundancy based filter", AdvancesinSoftComputing, 45, pp 242C249,2008.
[22] 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.
[23] Cardie, C., "Using decision trees to improve case-based learning", In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.
[24] 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