A Survey on Feature Selection Using FAST Approach to Reduce High Dimensional Data

R.Munieswari [#1], S.Saranya [*2]
M.E Scholar, Department of Computer Science & Engineering,
Kumaraguru College of Technology, Coimbatore, Tamilnadu, INDIA

Abstract **--- Feature selection is a process of identifying the most useful subset of features.The survey summarising most of the feature selection methods and algorithms. Feature selection is the process of identifying a subset of most useful features. Typically the feature selection methods consist of four basic main steps and classify different existing feature selection algorithms .It is defined in terms of generation methods and evaluation functions. Most useful or representative methods are chosen from each category. The strength and weakness of different feature selection algorithms are explained. The aim here is to select some of the feature to form a feature subset. Feature selection has been effective technique in dimensionality reduction, removing irrelevant data, increasing learning accuracy, and improving comprehensibility. Increase in dimensionality of data imposes a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. To find a subset of features, the efficiency is related to time, the effectiveness is related to the quality of the subset of features. Existing feature selection algorithm removes only irrelevant features. But FAST algorithm removes both Irrelevant and redundant features. This survey mainly focuses on Comparison of various techniques and algorithms for feature selection process.**

Keywords **--- Feature selection, classification, Filter method, Hybrid method, redundant features, and irrelevant features.**

I.INTRODUCTION

Clustering has been recognized as an important and valuable capability in the data mining field. For high-dimensional data, traditional clustering techniques may suffer from the problem of discovering meaningful clusters due to the curse of dimensionality. A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In distance-based clustering, the similarity criterion is distance. Here two or more objects belong to the same cluster if they are "close" according to a given distance. In the Conceptual clustering two or more objects belongs to the same cluster if one defines a concept common to all that objects.

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some case. Clustering is an unsupervised learning method, and a common technique for statistical data analysis used in many fields such as Medical, Science, and Engineering. This Survey summarise various known feature selection methods to achieve classification accuracy by using subset of relevant feature. Due to the computational complexity the full original feature set cannot be used. Attribute selection is the process of selecting a subset of relevant features. Feature selection technique is based on the data contains many redundant and irrelevant features. Redundant features provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Subset of Feature selection technique is general field of feature extraction.

1.1 Feature subset selection

Feature subset selection is used to improve the accuracy and comprehensibility that has been explored in machine learning. Some features are dependent on other features and there is no need to include the feature or noisy. Some feature will randomly fit the data and hence the probability of over fitting increases.

There are 4 basic steps in any feature selection method. They are Generation, Evaluation, Validation and Stopping criterion. In generation process to select the candidate feature subset, In evaluation process to evaluate the generated candidate feature subset and output a relevancy value, where the stopping criteria will determine whether it is the defined optimal feature subset. If yes, the process end else the generation process will start again to generate the next

candidate feature subset. After the required number of features is obtained the process will be terminated.

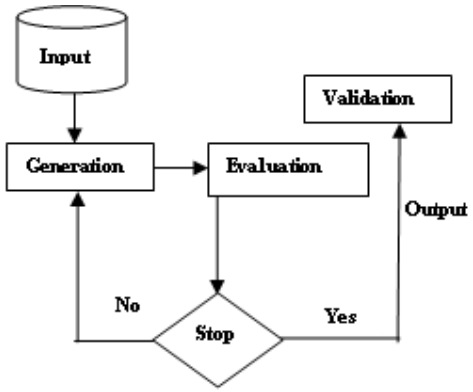

Fig.1 Feature Selection steps

1.2 Feature Selection Problems faced

The Feature Selection problem involves discovering a subset of features such that a classifier built only with this subset would have better predictive accuracy than a classifier built from the entire set of features. Feature selection algorithms will discover and select features of the data that are relevant to the task to be learned. In addition to irrelevant features, feature selection which may have a negative impact on the performance of learning systems such as redundant features and randomly class-correlated features. Irrelevant features do not contribute to the predictive accuracy of a particular target concept. Redundant features are relevant to a target concept. Randomly class-correlated features are correlated to the target class most of the time, and random otherwise. Thus, irrelevant, redundant and randomly class-correlated features are worthless and removing them can improve the learning process. The feature selection process can be seen alternatively as the process of identifying and removing as many irrelevant, redundant and randomly class-correlated features as possible.

1.3 Feature Selection Benefits

Feature selection method consist of potential benefits are

a. A reduction in the amount of training data needed to achieve learning.
b. The generation of learning models with improved predictive accuracy.
c. Learned knowledge more compact, simpler and easier to understand.
d. Reduced execution time required for learning.
e. Reduced storage requirements.

II. Classification

Classification is the process of predicts categorical labels. It is used to classify the data based on the training set and the values in a classifying attribute. The basic classification techniques are Decision tree induction; Bayesian classification, and Rule-based classification etc., Classification task plays an important role in clustering process. Classification is performed via following two step processes:

a) Model Construction**:** Describing a set of predetermined classes. Model is represented as classification rules, decision trees and mathematical formulae.

b) Model Usage**:** It is used to estimate the accuracy of model. Accuracy rate is the percentage of test set samples that are correctly classified by the model.

2.1 Classification methods

2.1.1 Bayesian Classification

Bayesian classifiers are statistical classifiers used to predict class membership probabilities. It is also known as naïve Bayesian classifier based on Bayes theorem. Compare to other classifiers it have the minimum error rate.

2.1.2 Decision tree Induction

Decision trees are constructed in a top-down recursive divide-and-conquer method. It consists of three algorithms such as ID3 (Iterative Dichotomiser), C4.5 (successor of ID3), CART (Classification and Regression Trees). The procedure employs an attribute selection measure such as gini index, information gain and gain ratio. Attribute selection measure [1] is used to separates the original data set (D) into individual classes.

2.1.3 Rule Based Classification

Rule-based classifiers use a set of rules for classification task. This method effectively produces the subset of features using different heuristic techniques.

III. Related Work

3.1 Feature Selection Approaches

A large number of algorithms have been used for the feature selection problem.

a) The search strategy used to determine the right subset of features.
b) Evaluation of each subset and feature selection algorithms are usually classified in three general groups: Filters, Wrappers and Hybrid solutions.

3.1.1 Filters

In a filter model [4], 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.

3.1.2 Wrappers

Filters, 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. 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. The main type of evaluation methods are

i. Distance (Euclidean distance measure).
ii. Information (entropy, information gain, etc.)
iii. Dependency (correlation coefficient).
iv. Consistency (min-features bias).
v. Classifier error rate (the classifier themselves).

After a feature subset is generated, it is feed into an evaluation process where the process will compute some kind of relevancy value. The generated subset candidate is feed into an evaluation function it will compute some relevancy value. The generation steps are able to categorise different feature selection method according to the way evaluation is carried out. The first four consider as a filter approach and the final one as a wrapper approach.

3.1.3 Hybrid Systems

The main goal of hybrid systems for feature selection [7] is to extract the good characteristics of filters and wrappers and combine them in one single solution. Hybrid algorithms achieve this behavior usually by pre-evaluating the features with a filter in a way to reduce the search space to be considered by the subsequent wrapper. The term "hybrid" refers to the fact that two different evaluation methods are used, a filter-type of evaluation and the classifier accuracy evaluation methods.

3.2 Feature Selection Algorithms

3.2.1 Feature selection based on Relief Algorithm

Kira and Rendell describe an algorithm called RELIEF that uses instance based learning to assign a relevance weight to each feature. Weight is assign for each feature's among the different class values. Features are ranked by weight and those feature exceed a user-specified threshold are selected to form the final subset of features. This method works by using randomly sample the instances from the training data. For each instance sampled the nearest instance of the same class (nearest hit) and opposite class (nearest miss) is found. An attribute's weight is updated according to how well its values distinguish the sampled instance from its nearest hit and nearest miss. An attribute that receive a high weight if it differentiates between instances from different classes and has the same value for instances of the same class.

The Relief algorithm [2] weight is assigned for each feature according to its relevance to the classification task. Initially all weights are set to zero and then updated iteratively. In this process two groups of instances are selected: some closest instances belonging to the same class and some belonging to a different class.

Relief-F extends Relief enabling this algorithm to work with noisy and incomplete datasets and to deal with multi-class problems. Relieve D is a deterministic version of Relief. It uses all instances and all near-hits and near-misses of each instance. This results in the equivalent of running Relief for an infinite amount of time. The EUBAFES (Euclidean Based Feature Selection) algorithm weights and selects features similarly to the Relief algorithm. It is also a distance-based approach that reinforces the similarities between instances that belong to the same class while deteriorating similarities between instances in different classes.

Consistency measure focuses to locate the optimal subset of related feature for improve the overall accuracy of classification task and deduce the size of the dataset. This method based on inconsistency rate over the dataset for a given feature [2, 3] set. Apply the consistency measure to feature selection task, first they calculate the inconsistency rate IR(S). Inconsistency rate is less than user threshold value then the subsets (S) are known as consistent. Consistency Measure use different Search Strategies such as Exhaustive, Complete, Heuristic, Probabilistic, and Hybrid. This method is monotonic, fast and Suitable for remove irrelevant and redundant features. Complete Search: ABB: ABB is Automatic Branch and Bound, extensions of Branch & Bound method. ABB algorithm having its bound set to the inconsistency rate of the original feature set.

3.2.2 Feature subset selection based on FOCUS method

The Focus system starts by searching through individual features that perfectly represents the original dataset. That is feature will be found when the inconsistency count for this

feature is equal to the one of the initial dataset. If no single feature is found to meet this criterion, combinations of features of size two, three and so on are considered until a perfect subset is discovered. Unlike LVF, this inconsistency measure simply counts the number of instances with same feature values but belonging to different classes.

A drawback of Focus is that it is computationally prohibitive if the number of features to consider is not small. The exhaustive search may allow Focus to find the optimal subset of features. Focus is a deterministic solution. ABB is a Branch and Bound algorithm with its bound set to the inconsistency rate ± of the data set with the full set of features. It starts with the full set of features, removes one feature from this set to generate new subsets until no more feature can be removed.

3.2.3 Feature selection based on LVF algorithm

Feature selection based on LVF algorithm LVF [4] randomly searches the space of subsets using Las Vegas algorithm. During iteration a new subset is randomly generated. If the number of features in subset is smaller than or equal to that of the current best subset, subset is evaluated with the use of an inconsistency count. This count is based on the intuition that the most frequent class label among the instances matching subset of features is the most probable class label. An inconsistency threshold is used to reject subsets that have greater inconsistency rate. This process is repeated. LVF is a non-deterministic algorithm, being able to output several partial results during processing. The LVS algorithm was designed to make LVF scalable.

QBB applies a combination of LVF and ABB (Automatic Branch and Bound). First, it runs LVF and records all returned subsets. Then it runs ABB over each one of this subset by using them as starting sets. The algorithm keeps the minimum sized subsets so that ABB is more focused.

Only feasible subsets can be chooses as final solution. The Beam Search (BS) feature selection algorithm works similarly to the Best-First algorithm, except that it limits the scope of the search with a bounded queue. Here, subsets will be placed on the queue from best to worse. At every iteration, the best subsets are extracted from the queue. All possible subsets by adding a feature to it are generated and placed back in the queue in their proper position.

3.2.4 Feature Selection Based on FAST algorithm

Feature Subset selection framework involves irrelevant feature removal and redundant feature elimination, the traditional definitions of relevant and redundant features, then provided definitions based on variable correlation. In FAST method irrelevant feature are removed by using threshold value at the same time redundant features are eliminated by using minimum spanning tree. Feature subset selection can be the process that identifies irrelevant features has no/weak correlation with target concept. Redundant features are assembled in a cluster and a representative feature can be taken out of the cluster. The advantages of FAST method have Good feature subsets contain features highly correlated with (predictive of) the class yet uncorrelated with each other.

i. Construction of MST: The Minimum Spanning Tree (MST) is constructed by using Prim's algorithm. This algorithm works better than existing Kruskal's algorithm The symmetric uncertainty (SU) 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 Therefore, choose symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept.

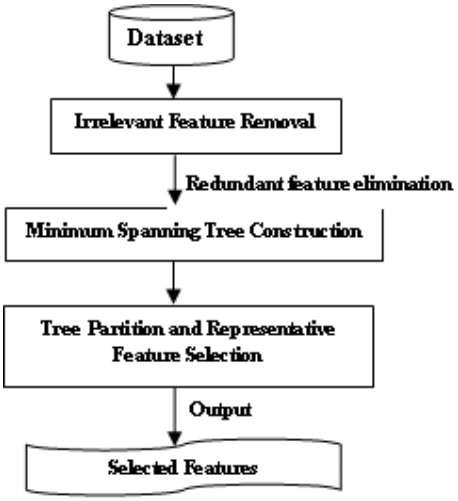

Fig.2 Feature subset selection using FAST method

ii. Partitioning MST: After MST remove the edges whose weights are smaller than both of the T-Relevance from MST deletion results in two disconnected trees T1 and T2. The set of vertices in any one of the final trees to be V (T) have the property that for each pair of vertices guarantees the features in V (T) are redundant. Removing all the unnecessary edges, a forest is obtained. Each tree T Forest represents a cluster that is denoted as V (T), which is the vertex set of T as well. As the features in each cluster are redundant, so for each cluster V (T) choose a representative feature.

iii. Classification Accuracy: FAST Algorithm has better classification accuracy according to the four classifiers, such as Naive Bayes, C4.5, IB1, and RIPPER.

TABLE
Comparison of Different Algorithms/Techniques

S. NO	Techniques (or) Algorithms	Pros	Cons
1.	FAST Algorithm	Improve the performance of classifiers	Required more time
2.	Consistency Measure	Fast, Remove noisy and irrelevant data	Unable to handle large volumes of data
3.	Wrapper Approach	Accuracy is high	Computational complexity is large
4.	Filter Approach	Suitable for very large features	Accuracy is not guaranteed
5.	Hybrid Approach	Reduce Complexity	Decrease the Quality when dimensionality become high
6.	INTERACT Algorithm	Improve Accuracy	Only deal with irrelevant data
7.	Relief Algorithm	Improve efficiency and Reduce Cost	Powerless to detect Redundant

IV. Conclusion

In this survey, feature selection approaches which helped classify and describe several filter and wrapper selection algorithms according to their generation procedure and evaluation criterion. In addition classifying hybrid feature selection algorithms by taking into consideration both the type of filter evaluation measure and the classifier used by such methods is developed. From the study of this method the irrelevant feature removal is simple. The FAST algorithm removes both irrelevant and redundant features. Relevant features have strong correlation with target concept. So we conclude that FAST algorithm works better than existing feature selection methods.

REFERENCES

[1] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transaction on Knowledge and Data, Engineering, Vol. 25, No. 1, January 2013.

[2] AlmuallimH.and. Dietterich T.G, "Algorithms for Identifying Relevant Features", Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.

[3] Arauzo-Azofra A.,. Benitez J.M, and Castro J.L., "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[4] Biesiada J. and Duch W., "Features selection for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[5] Das S, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[6] Dash M. and Liu H., "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.

[7] Kohavi R. and. John G.H, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324, 1997.

[8] Souza J., "Feature Selection with a General Hybrid Algorithm," PhD dissertation, Univ. of Ottawa, 2004.