ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# Subset Selection in High Dimensional Data by Using Fast Clustring Technique

**C.Pearley Vinitta Sharon*, K.Mohamed Amanullah**
M.Phil Scholars*, M.Sc., M.Phil,, Bishop Heber College, Trichy-17, Tamil Nadu, India.
pearleysharon@gmail.com

## Abstract

A feature subset selection is an effective method for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. This process enhanced by cluster based FAST Algorithm using MST construction. With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for dropping dimensionality, remove irrelevant data, rising learning accuracy, and improving result comprehensibility. Features in different clusters are moderately independent. The clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers such as the probability-based Naive Bays, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. We can build FAST algorithm with prim's algorithm based on MST Construction. Our investigational results show that improves the performances of the four types of classifiers.

**Keywords**: Feature Subset Selection, Cluster based FAST Algorithm, MST Construction.

## Introduction

### Creating a Cluster

When a system administrator desires to create a new cluster, the administrator will run a cluster installation utility on the system to become the first member of the cluster. For a new cluster, the database is created and the initial cluster member is added. The administrator will then configure any devices that are to be managed by the cluster software [3]. We now have a cluster with a single member. In the next step of clustering each node is added to the cluster by means of similarity on the basis of the resources used. The new node routinely receives a copy of the existing cluster database.

### Feature Selection for Clustering

The existence of irrelevant features in the data set may humiliate learning quality and consume more memory and computational time that could be saved if these features were removed. In addition, different relevant features may produce different clustering. Therefore, different subset of relevant features may result in dissimilar clustering, which greatly help discovering different hidden patterns in the data [1]. Motivated by these facts, dissimilar clustering techniques were proposed to utilize feature selection methods that remove irrelevant and redundant features while keeping relevant skin tone in order to improve clustering efficiency and quality. For simplicity and better organization, we are going to describe different feature selection for clustering (FSC) methods based on the domain. The following sections will be organized as follows: predictable FSC, FSC in text data, FSC in streaming data, and FSC link data. Similar to feature selection for supervised learning, methods of feature selection for clustering are categorized into filter wrapper, and mixture models. To alleviate the computational cost in the wrapper model [3], filtering criteria are utilized to select the candidate feature subsets in the hybrid model.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

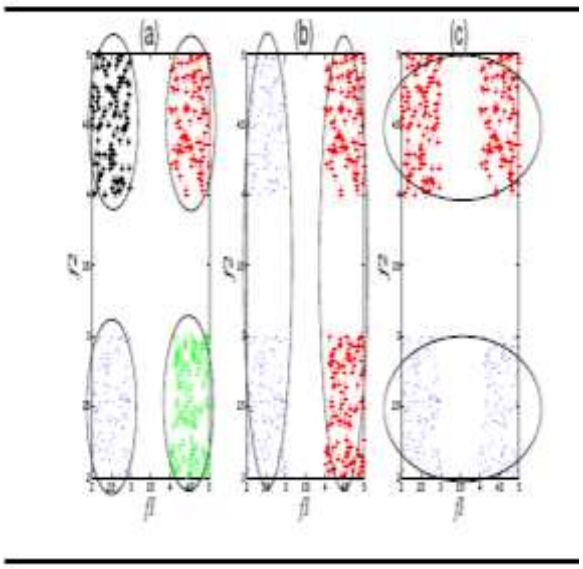



***Fig 1.Sub set classification.***

**Existing Work**

Subset selection evaluates a subset of features as a group for appropriateness. Subset selection algorithms can be broken into Wrappers, Filters and entrenched. Wrappers can be computationally costly and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded and specific to a model and also use the feature selection has been an active research area in pattern identification, statistics, and data mining communities.

The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can radically improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of analytical features is an important problem in its own right. The feature selection can be identified in clustering or unsupervised learning.

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. Conventional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a programmed learning algorithm to determine the goodness of the selected subsets, the accurateness 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 self-governing of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. Feature selection is a process that chooses a subset of original features. The optimality of a feature subset is measured by an evaluation criterion. As the dimensionality of a domain expands, the number of features N should increases. Finding an optimal feature subset is usually intractable and many problems related to feature selection have been shown to be NP-hard.

A typical feature selection process consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and result that produce candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. The hybrid methods are a mixture of filter and wrapper methods by using a filter method to reduce search space that will be considered by the following wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time difficulty of the filter methods.

FAST can be very useful for enhancing customer relationship management (CRM) because standard application of cluster analysis uses the complete set of features or a pre-selected subset of features based on the prior knowledge of market managers. Thus it cannot provide new marketing models that could be efficient but have not been considered. Our approach provides possible feature subset sizes and numbers of clusters. The generalization of the selected features is limited and the computational complexity is large.

- Their computational complexity is low, but the correctness of the learning algorithms is not guaranteed.
- The hybrid methods are a grouping of filter and wrapper methods by using a filter method

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

to reduce search space that will be considered by the succeeding wrapper.

## Related Work

We can propose the feature subset selection can be viewed as the process of identify and removing as many irrelevant and redundant features as possible. This is because unrelated features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s) and construct minimum spanning trees to evaluate whether two sets of n-dimensional data are from the same distribution.

Unrelated features, along with redundant features, severely affect the accuracy of the learning machines. Thus the 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**.** Keeping these in mind, we expand 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 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 removal A minimum spanning tree is built across the data points, and edges which connect data from one distribution to the other are removed. If many edges are removed, then the data from the distributions are mixed up together, and so they must come from the same distribution. A minimum-spanning tree is a sub-graph of a weighted, connected and undirected graph. It is acyclic, connects all the nodes in the graph, and the sum of all of the weight of all of its limits is minimum. That is, there is no other spanning tree, or sub-graph which connects all the nodes and has a smaller sum. This approach can easily be applied in feature subset selection.

Instead of attempt to determine whether the sets of data come from different distributions, we try to find a feature subset which best shows that the sets of data come from different classes. Given a feature subset to evaluate, a minimum spanning tree is built across the sample data. The edges leading from one class to another are removed, and tallied. The more edges that are removed, the worse the feature subset is. When construction of a minimum spanning tree on a complete graph, an algorithm which has a complexity based on the number of boundaries must have a complexity better than O(M) to beat Prim's algorithm. The just about O (M), but it is far more complex than this algorithm. As a result, we are using Prim's algorithm to construct minimum spanning trees in our criterion function.

- Good feature subsets contain skin tone highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.
- The professionally and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.
- Generally all the six algorithms achieve significant reduction of dimensionality by select only a small portion of the original features.
- The null proposition test is that all the feature selection algorithms are equivalent in terms of runtime.

## Upload Datasets

We upload the datasets. A dataset is a collection of data. Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where each column of the table represents a particular variable, and each row corresponds to a given member of the dataset in question. The dataset list out values for each of the variables, such as height and weight of an object, for each component of the dataset. Each value is known as a datum. The dataset may include data for one or more members, corresponding to the number of rows.

## Preprocessing

Data pre-processing is an important step in the data mining process. It express garbage in, garbage out is particularly applicable to data mining and machine learning projects. Data-gathering methods are often slackly controlled, resulting in out-of-

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

range values, impossible data combinations missing values, etc. Analyzing data that has not been suspiciously screened for such problems can produce misleading results. Thus, the demonstration and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge detection during the training phase is more difficult.

Data pre-processing includes cleaning, normalization, which transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

## MST construction

A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every added spanning tree. which is a union of minimum spanning trees for its connected components. Finding the smallest edge can be done at the same time as updating a values of each components.

Minimum spanning tree is When building a minimum spanning tree on a complete graph , an algorithm which has a complexity based on the number of edges must have a complexity better than O(M) to beat Prim's algorithm.

## Tree partition

Each tree in the MST represent a cluster. In this module, we apply graph theoretic clustering methods to features. In particular, we adopt 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.

## Feature Selection

Feature subset selection was split up into two parts, subset searching and criterion functions. For both parts, the common algorithms were introduced and analyzed. Feature selection, also known as variable selection, attribute selection, variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central postulation when using a feature selection technique is that the data should may contains any redundant or irrelevant features.

Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful 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.

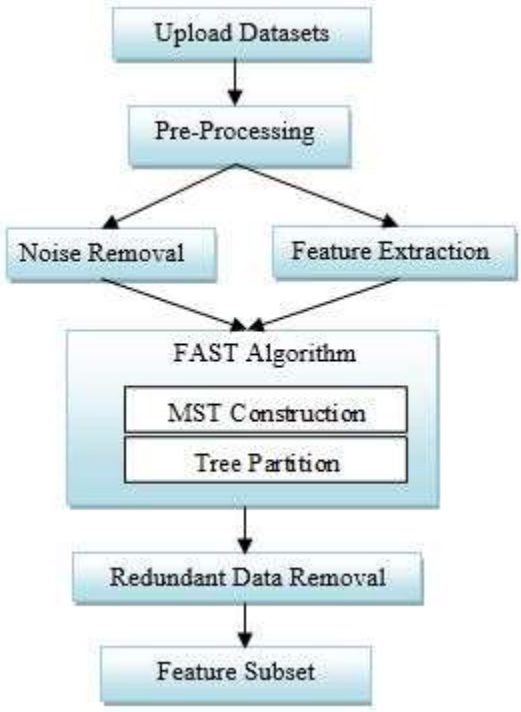
## Performance Evaluation

The experimental results show that, compared with other five different types of feature subset collection 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.

Danish Jamil, Is Ethical Hacking Ethical?

1. Kumar Utkarsh, System Security and Ethical Hacking Volume 1.Issue 1. 2013
2. Marilyn Leathers, Closer Look at Ethical Hacking and Hackers
3. Dinesh Babu S, Ethical Hacking, Volume.3.Jan-2012.
4. Kenneth Einar Himma ,The Ethics of Tracing Hacker Attacks through the Machines of Innocent Persons, Volume.2, 11/2004
5. Regina D. Hartley, .Rationale for a hacking methodological approach to network security
6. Jeffrey Livermore, Walsh College, Member, IEEE Computer Society, What Are Faculty Attitudes toward Teaching Ethical Hacking and Penetration Testing? June2007.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852




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.

## Conclusion

We conclude that our proposed algorithm analyze the features. 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. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best categorization accuracy for IB1. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high possibility of producing a subset of useful and independent features. Our proposed algorithm performs efficient all types of datasets.

## *References*

1. *Almuallim H. and Dietterich T.G., Algorithms for finding relevant Features, In Proceedings of the 8th Canadian Conference on AI, pp 38-45, 1996.*
2. *Almuallim H. and Dietterich T.G., Learning boolean concept in the occurrence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279- 305, 1996.*
3. *Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set calculate based on relief, In Proceedings of the fifth international conference on Recent advance in Soft Computing, pp 104-109, 2003.*
4. *Baker L.D. and McCallum A.K., Distributional cluster of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR meeting on Research and enlargement in information Retrieval, pp 96- 103, 1998.*
5. *Battiti R., Using common information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.*
6. *Bell D.A. and Wang, H., formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.*
7. *Biesiada J. and Duch W., Features election for high-dimensionaldata Pearson redundancy based sorting, Advances in Soft Computing, 45, pp 242C249, 2008.*
8. *Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature collection through Clustering, In dealings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.*
9. *Cardie, C., Using decision trees to improve case-based learn, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.*
10. *Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute communications Using Information Theoretic Metrics, In Proceedings of IEEE international meeting on Data Mining Workshops, pp 345-350.*