ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

Survey On: Comparison of Clustering Based Feature Subset Selection Algorithms for High Dimensional Data

Vishnu M. Tore [1], Prof. P. M. Chawan [2], Prof. S. A. Khedkar [3], Prof. Kishor Dongarwar [4]

[*] M.Tech Student, Department of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

Associate Professor, Department of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

Assistant Professor, Department of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

Assistant Professor, Department of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

ABSTRACT: In data mining Feature selection is the area which is mostly used as input for high dimensional data for effective data mining. Feature selection is used to identify most relevant feature amongst all.This selection can be measured in terms of effectiveness and efficiency. While effectiveness concerns the quality of the subset of features and efficiency is related to time required to find subset of features. The main idea while using a feature selection algorithm is that input data contains many redundant as well as irrelevant features. There are some or less drawbacks of these algorithms, amongst these some algorithms can eliminate irrelevant features but fail to handle redundant features and others can remove the irrelevant features taking care of the redundant features. Based on these constraint, a FAST (fast clustering-based feature selection algorithm is proposed).The FAST algorithm has two steps. 1. features are divided into clusters by using non-linear clustering methods. 2. the most relevant feature that is strongly related to target classes is selected from each cluster to form a subset of features.For this we use MST(Minimum Spanning Tree) using Kruskal's Algorithm clustering based method.

KEYWORDS: Data mining, Feature subset selection, Feature selection, relevant features, redundant features.

I. INTRODUCTION

Clustering can be defined as combining a set of data objects into classes of same properties. The performance, robustness, and use of clustering algorithms is depending on searching similarities between data according to the characteristics of data and arranging relevant data objects into clusters. The quality of a clustering depends on the similarity measure method and its implementation. FAST clustering subset selection algorithm can be used for increasing quality and accuracy. Feature selection method is same as feature extractions but first one is more generic. In feature extraction, new feature set is formed from the features which are already present in the data set. While in feature selection process, it gives the subsets of feature which is most relevant to and useful for required search. In this algorithm searching complexity is very less as compared to other methods and the results obtained with this are free from redundancy. The selection of these features can be done with the help of various algorithms such as best search, greedy forward selection and greedy backward elimination algorithm, genetic algorithm[8]..

Till now there are different feature subset selection algorithms have been proposed for machine learning applications. They can be categorized into four categories: the Embedded, Filter, Wrapper, and Hybrid approaches. The wrapper method used to determine the quality of the selected subsets, accuracy of this algorithm is ordinarily high. But the computational complexity is more. In contrast The filter method are having good generality. Their computational complexity is less, but the correctness of the algorithms is not guaranteed [11]. The wrapper methods are computationally expensive and not useful on small training sets. The filter methods are usually a good choice when the number of features is very large(high dimensional data). To reduce dimensionality and to obtained quality features

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

FAST algorithm is used by removing irrelevant data with most suitable way. To obtain the goal of FAST clustering algorithm, it works in two parts. In the first part, it divides the input dataset into clusters using graph methods. For this purpose Minimum Spanning Method(MST) is used. In second part, the subsets that are having great accuracy or relative with the required search are selected from cluster and form a feature subset. Feature subset identifies and removes as many irrelevant and redundant features as possible. All the related feature set and unrelated feature set are clustered in different sets. This technique is useful in taking take care of related and unrelated feature sets for prediction as per the requirement of target. Even though some of existing algorithms has capability to remove the unrelated feature set, some do not involves quality assurance and accuracy to achieve goal.
Adopted Minimum Spanning Tree Computes a neighbourhoodness amongst graph of instances, then deleting any edge in the graph that is much longer or shorter than its neighbours. The result of this computation will be a forest of tree(part of tree).Each tree in that forest represents a cluster of feature. This resulting cluster is used for feature subset selection [10].

II. RELATED WORK

The following papers explains different views regarding the feature selection and its techniques. Also they elaborate the different feature selectioin methods.

Qinbao Song, Jingjie Ni and Guangtao Wang [1] presented paper stating that FAST clustering algorithm is effective and efficient related to quality and time.The paper also explained about Wrapper,Filter,Hybrid techniques and its drawbacks.

L. Yu and H. Liu [2] proposed an approach that data mining for high dimensional data is big problem.Which is focused on removing irrelevant data. It uses correlation-based approach for feature selection. This helps to reduce features which are having zero linear correlation and reduce redundancy amongst selected features.

M. Dash, H. Liu, and H. Motoda [3] They have mainly concentrated on consistency measure for feature selection. This paper have elaborated the consistency with properties and greaterness of this with other present techniques for selection of features.

In [4] The paper contains main idea of working of FAST clustering based feature subset algorithm. It also cover idea that if more than one feature are joint and if matching with target feature then declare it as relevant. It also elaborate the distributed clustering and time complexity of Prim's algorithm.

In [5] Dingcheng Feng, Feng Chen_, and Wenli Xu have explained how to calculate optimized feature set on the basis of classification and clustering. In this paper the drawbacks of backward greedy elimination algorithm by leave one out strategy.

III. FEATURE SELECTION

Procedure for Feature Selection:

- Produce Subset: Produce candidate subset from original feature set.
- Estimation : Estimate the obtained subset.
- Calculation : Comparing with user defined threshold value.

Verification: Test out whether the subset is valid.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

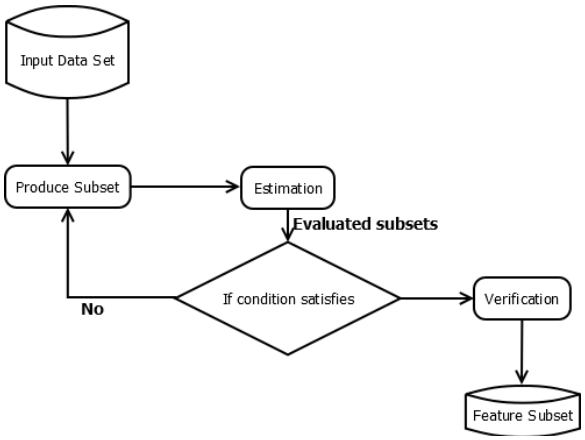

Fig: Steps involved in feature selection

IV. CLASSIFICATION

It is used to classify the data based on the training set and its attribute. It has two steps as follows
Model Construction-Classification rules, decision trees and formulaes.
Model Usage-To estimate the accuracy of the model

Classification Methods
Bayesian Classification: It is used for the prediction of class membership probabilities. It have minimum error rate as compared to other classifiers.

Decision tree: It is constructed in top down approach. It is having three algorithms ID3,C4.5,CART

Rule Based Classification: It uses set of rules for classification. This method produces the subset of features using different techniques.

V. CLUSTERING TECHNIQUES

Minimum-Spanning Tree:

It is connected, undirected and weighted graph spanning tree of minimum weight. It is started with empty graph and added one edge at a time to form subset of some minimum spanning tree. It's disadvantage is that it requires more time to form cluster. It has various applications:

1. Computer Network
2. Cluster Formation
3. Transportation Network
4. Ricognition of Handwriting

Graph Clustering:

It is simple to evaluate the neighborhood graph of instances and deletion of edge that is short or long as compared to other neighbors. It result into a forest and each tree in the forest represent a cluster.

Consistency Search:

It mainly reduces the size of dataset and locate the relevant feature subset to improve performance in terms of accuracy. It has two steps one is calculating inconsistency rate(IR) and second is comparing with fixed value which is threshold.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

If inconsistency rate is less than this value then that subset is consistent. It is having various searching strategies as Exhaustive searching, Complete searching, Heuristic searching, Probabilistic search, Hybrid search[9].

Relief Algorithm:

This algorithm estimates a good feature set. It estimates the quality of feature by comparing nearest feature with selected features. It has two terms Hit(H) and Miss(M). H refers to hit from same class and M refers to miss from other class. So based on these terms quality of features I evaluated. Relief method is used to find out most accurate and relevant feature set[6].

Hierarchical Clustering:

It is defined as grouping data objects into a cluster tree. It is having two types: 1.Bottom Up Approach: this is also known as Agglomerative Clustering in which clustering starts with each object resulting into separate group and merging these groups into a larger group. This is done until all objects are in the same group. 2.Top Down Approach: This is called as Divisive Clustering. It starts with all of objects in the same cluster. In each step, a cluster is divided into smaller clusters, until a specific condition arrives. Tree formation symbolize the process of hierarchical clustering.

VI. FEATURE SELECTION MEHODS

Filters
In a filter model 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. it will result in low performance if the evaluation criterion does not match the classifier well.

Wrappers
Filters, wrappers 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 a evaluation function called as error rate. 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 Distance, Information, Dependency, Consistency, Classifier rate.
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 categorize different feature selection method according to the way evaluation is carried out. The first four considered as a filter approach and the last one is of wrapper method approach.

Hybrid Systems
The main goal of hybrid systems for feature selection 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” means that two different evaluation methods are used for the evaluation of feature, a filter-type of evaluation and the classifier accuracy evaluation methods.

FAST Algorithm

Feature Subset selection framework involves irrelevant feature removal and duplicate feature elimination, the former 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

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

minimum spanning tree. The advantages of FAST method have Good feature subsets contain correlated features within the class but uncorrelated with each other.

Feature Selection Based on FAST:

Firstly based on input dataset irrelevant and redundant features are removed. Based on obtained features the next step is minimum spanning tree construction. The construction is done based on Prim's Algorithm. This algorithm works better than Kruskal's Algorithm.

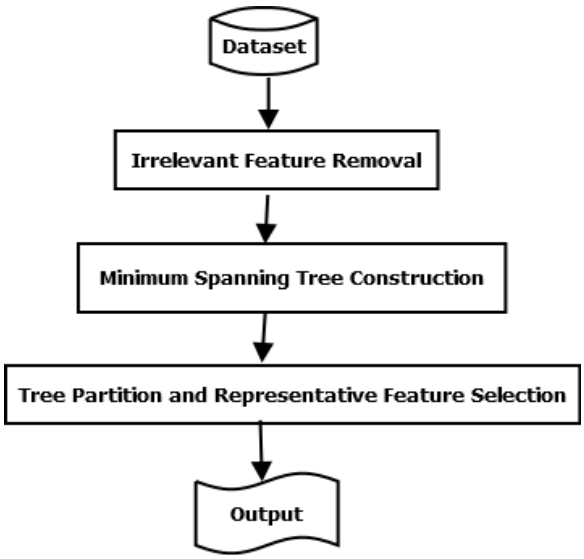

Fig. Feature Subset Selection using FAST Algorithm[11].

COMPARISON OF VARIOUS ALGORITHMS AND TECHNIQUES

Sr.No	Algorithm	Advantages	Disadvantage
1	FAST Algorithm	Greater Classification	More Time
2	Wrapper	High Accuracy	More Computation Complexity
3	Filter	Good for large features	Accuracy Problem
4	Consistency Measure	Fast, remove irrelevant features	Only for small datasets
5	Relief Algorithm	Improve efficiency and reduce cost	Fails to detect redundant features
6	Distributed Clustering	Higher Classification Accuracy	Evaluation Difficulty
7	Agglomerative	Less Complexity	Quality Assurance Fails

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 2, February 2016

VII. CONCLUSION

Feature selection method is an efficient way to improve the accuracy of classifiers, dimensionality reduction, removing both irrelevant and redundant data. In this paper, we have made a comparative study of various feature selection methods and algorithms. Discussions of those techniques are reviewed and the benefits and drawbacks of feature selection methods and algorithms are summarized.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast clustering based feature subset selection algorithm for high dimensional data", In proceedings of the IEEE Transactions n Knowledge and data engineering, 2013.
[2] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.
[3] M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.
[4] A New Clustering Based Algorithm for Feature Subset Selection (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5272-5275.
[5] Dingcheng Feng, Feng Chen_, and Wenli Xu "Efficient Leave-One-Out Strategy for Supervised Feature Selection" TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 09/10 pp629 635 Volume 18, Number 6, December 2013.
[6] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int"l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.
[7] H. Liu, H. Motoda, and L. Yu, "Selective Sampling Approach to Active Feature Selection," Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.
[8] 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.
[9] M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.
[10] Jesna Jose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.
[11] K..Revati, T.Kalai Selvi,"Survey: Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, Issue 12, December 2013.