International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Fast Clustering Based Feature Selection Algorithm for High Dimensional Data

Swapnil A. Sutar[1], Prof. Devendra P. Gadekar[2]

[1]Research scholar, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune

[2] Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune

Abstract: ***Data mining has main problem of partitioning a group of objects into a number of subsets, such that similarity in each subset or cluster is increased and effective result should be obtained. The feature selection method is more generalized form of feature extraction. Feature selection gives useful feature from data while the feature extraction creates new feature set according to existing feature sets. The main concept of this fast clustering feature selection algorithm is to cluster subset with most similar characteristic while removing irrelevant subset from that cluster. This FAST algorithm concern with both efficiency and accuracy for finding required set. The FAST algorithm requires two steps for its working, in first step algorithm uses minimum spanning tree (MST) to divide data into different clusters and in second step, it removes irrelevant sets and gives accurate and efficient result with similar sets. All clusters resulted in FAST algorithms are relatively independent of each other. So this may useful for most effective results.***

Keywords: Feature extraction, Feature selection, FAST clustering Algorithm, irrelevant subset, Minimum Spanning Tree

1. Introduction

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. It is a common and important task that finds many applications in IR and other places. The performance, robustness, and usefulness of clustering algorithms are depends on Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. So for increasing quality and accuracy FAST clustering subset selection algorithm is used.

Feature selection method is general form of feature extractions. In feature extraction, new feature set is generated from the data features which are already preset. While in feature selection process, it gives the subsets of feature which are useful for required search. It provides advantages including less time for searching, optimal results etc. The feature subset selection can be done with the help of various algorithms such as best search, greedy forward selection algorithm, greedy backward elimination algorithm, genetic algorithm [12].

Many 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 [1]. The wrapper method used to determine the goodness of the selected subsets, the accuracy of this algorithm is ordinarily high. But the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the algorithms is not guaranteed [1]-[9]-[10].

The wrapper methods are computationally expensive and not useful on small training sets [1]-[9]. The filter methods are usually a good choice when the number of features is very large i.e. for high dimensional data.

FAST algorithm is an effective way to reduce dimensions, removing irrelevant data and to produce result in high effective manner.

To achieve goal of FAST clustering algorithm, it works in two steps. In the first step, it divides data into clusters using graph methods. For this purpose Minimum Spanning Method is used. In second step of FAST, the subsets that are most accurate or relative with the required search are selected from cluster and form a feature subset [1] (figure1). Feature subset identifies and removes as many irrelevant and redundant features as possible. As we are clustering all the features with their relations, all the related feature set and unrelated feature set are clustered separately. This clustering helps to take care of related and unrelated feature sets for prediction as per the required target search. Although some of existing system or algorithm has capability to remove the unrelated feature set, some do not involves efficiency and effectiveness to achieve goal.

Adopted Minimum Spanning Tree Computes a neighborhood graph of in-stances, then delete any edge in the graph that is much longer or shorter (according to some criterion) than its neighbors. The outcome will be a forest and each tree in that forest represents a cluster of feature. This cluster is used to feature subset selection [10].

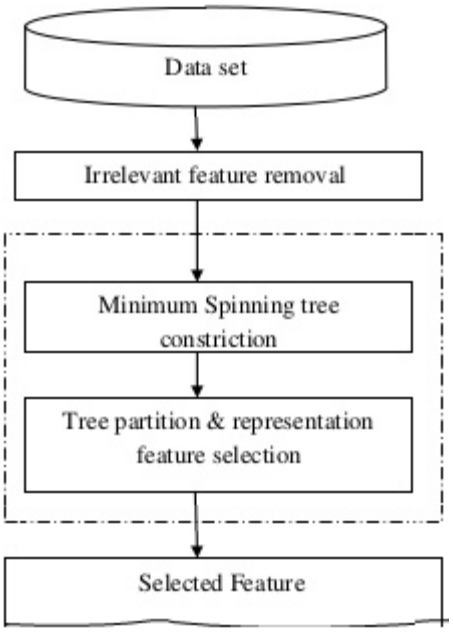

Figure 1: Basic Framework of FAST Algorithm.

After surveying various embedded approaches, FAST clustering produces results which are efficient and effective.

2. Literature Survey

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. :***

They stated the FAST clustering based algorithm is effective and efficient. Efficiency refers to time required to search a particular feature set through large data and effective concerns with quality or accuracy of the selected feature set. FAST uses MST for feature selection. It works in two different steps including partitioning data into clusters and then finding appropriate feature set. The paper also explained wrapper, filter, hybrid methods and its limitations. They have experimented FCBF, Relief F, CFS, Consist, and FOCUS-SF techniques on 35 different datasets and conclude that FAST algorithm is more effective than all others. The exact working of FAST algorithm is explained in this paper. As FAST also works on removing irrelevant data, it is more useful in finding accurate results.

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. :***

Data mining of high dimensional data is big problem. Feature selection from high dimensional data is basically focused on removing irrelevant features i.e. not related to the required search. But it is also difficult to remove irrelevant data. This paper provides a study of feature redundancy in high-dimensional data and proposes a novel correlation-based approach to feature selection within the filter model. Classical linear correlation helps to remove features with near zero linear correlation to the class and reduce redundancy among selected features. It uses FCBF algorithm. This algorithm firstly calculates symmetrical uncertainty SU and then calculate list of symmetrical and relevant data.

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. :***

They stated that the feature selection algorithms uses various measures to determine the usefulness and effectiveness of search result. The paper mainly concentrated on consistency measure for feature selection. They have explained the consistency, its properties and comparison with other present measures for feature selection. Also we studied how to calculate inconsistency measure, distance measure, consistency measure, information measure for better search.

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:***

This paper also contains the main idea of the FAST clustering based feature selection algorithm and its step for working of algorithm. They have also explained the main challenge that if more than one feature are joint and they suit the target feature then it can be treated as relevant. Feature interface is the new challenge for identifying the applicable feature. FAST algorithm extract only targeted features out of many features. They don't measure the irrelevant and redundant data because irrelevant and redundant data affects the competence and effectiveness of the algorithm. It also explains distributed clustering and time complexity of prim's algorithm.

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:***

This paper shows how to select optimized feature sets on the basis of clustering as well as classification. They have explained greedy backward elimination algorithm for better optimization. It states that finding related feature is more difficult than to find irrelevant data sets or non-related feature. Hence leave-one-out strategy may be more helpful and cost effective. The limitation of greedy backward elimination algorithm has overcome through this leave-one-out strategy.

6] ***Houtao Deng, George Runger "Feature Selection via Regularized Trees" The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012 :***

The author proposed tree regularization framework for feature subset selection. This model or framework gives effective results as tree models can deal with variables of numerical and categorical, different scales between variables, missing data, interactions and nonlinearities etc. The tree regularization framework provides an effective and efficient feature selection solution for many practical problems. The algorithm includes the splitting of tree from first node up to the last. Then the forest trees are generated which are

referred as regularized random forest (RRF) and regularized boosted random trees (RBoost).

This method can deal with either strong or weak classifiers.

7] *Yijun Sun, Sinisa Todorovic "Local Learning Based Feature Selection for High Dimensional Data Analysis" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 32, NO. 9, SEPT.2010:*

This paper states the partitioning of complex non-linear problems into simple local linear sets with the help of local learning. Then through numerical analysis and machine learning, feature relevance is obtained globally. This algorithm is used in high dimensional data. In this algorithm, first step is to calculate the margin by distance function i.e. first find two neighbors of every sample, one from nearest hit class and other from nearest miss class. Through this margin, the noise or irrelevance of features can be obtained. The algorithm has to compute margin and distance function locally every time. then it finds hidden variables. It also states major problem with RELIEF algorithm [11]-[13].that the nearest neighbors of a given sample are predefined in the original feature space, which typically yields erroneous nearest hits and misses in the presence of copious irrelevant features.

8] *Sriparna Saha "Feature selection and semi-supervised clustering using multiobjective optimization" SpringerPlus 2014, 10.1186/2193-1801-3-465.:*

The paper stated the multi-objective optimization to overcome the problem of automatic feature selection and semi-supervised clustering. cluster centers and features are encoded in the form of a string. The stated technique archived multi-objective simulated annealing (AMOSA) is utilized to detect the appropriate subset of features, appropriate number of clusters as well as the appropriate partitioning from any given data set. The simulated annealing (SA) technique has limitations that it gives single solution for feature selection after single run. Hence for multi-dimensional data, it is not useful to find feature set.
The newly developed AMOSA uses multiple distance indexes in string form to find appropriate feature data set.

9] *R.Munieswari,"A Survey on Feature Selection Using FAST Approach to Reduce High Dimensional Data",IJETT, Volume 8 Number 5- Feb 2014 :*

It stated the survey on different feature selection techniques or algorithms including wrapper, filter, fast clustering algorithm, hybrid approach and relief algorithm. The comparison among all these algorithms is explained. The Relief algorithm concerns with giving weight to each cluster. Then reorder the cluster according to max weight. When the cluster weight crossed the threshold value, the particular cluster is taken as feature set. All other methods are studied in previous and other papers mentioned above.

They have also stated the FAST algorithm with use of MST and gives advantage of FAST over all other techniques.

10] *Jesna Jose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014:*

The wrapper, filter relevance analysis models are surveyed in this paper. Analysis of relevance and redundancy classified the features on relevance basis as whether is of category strong relevance, weak relevant or irrelevant. Strong relevance of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all. FAST works on all including removal of irrelevant sets. Hence it can be more useful than any other.

3. Overall Survey

By studying various paper and journals, the advantages and limitations of different techniques used for feature selection are found. The comparison shows the advantages and limitations of particular feature selection technique (Table1).

Table 1: Comparison of different feature section methods

Algorithm	*Advantage*	*Disadvantage*
Wrapper Approach	High Accuracy	Large computational complexity
Filter Approach	Suitable for very large features	Accuracy is not guaranteed
Distributional Clustering	Higher Classification Accuracy	Difficult to Evaluation
Relief Algorithm	Improve efficiency, Reduces cost	Powerless to detect redundant features
Simulating Annealing	Accuracy, Useful for small datasets	Single feature for single turn.
FAST Algorithm	Efficient, Effective	Takes more time

4. Future Enhancement

In future scope, different correlation measures along with fuzzy logic can be included in the present algorithm to improve performance of a system. We can enhance this work by extending the symmetric uncertainty for extracting the feature subset selection.

5. Conclusion

We surveyed various papers and journal regarding feature selection and its techniques to find appropriate feature set. Through comparison between wrapper, filter, hybrid approach for feature selection, FAST algorithm is more efficient and accurate than others. FAST algorithm uses Minimum Spanning Tree to divide data into clusters and then relevant feature set is obtained excluding the irrelevant feature set. It deals with removing of irrelevant and redundant data or feature set that leads to provide high accurate feature as per required target class.

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] Houtao Deng, George Runger "Feature Selection via Regularized Trees" The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012.
[7] Yijun Sun, Sinisa Todorovic "Local Learning Based Feature Selection for High Dimensional Data Analysis" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 32, NO. 9 SEPT.2010.
[8] Sriparna Saha "Feature selection and semi-supervised clustering using multiobjective optimization**"** *SpringerPlus* 2014, 10.1186/2193-1801-3-465.
[9] R.Munieswari,"A Survey on Feature Selection Using FAST Approach to Reduce High Dimensional Data",IJETT, Volume 8 Number 5- Feb 2014.
[10] Jesna Jose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.
[11] S. Chikhi and S. Benhammada, "ReliefMSS: A Variation on a Feature Ranking Relieff Algorithm," Int'l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.
[12] L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.
[13] I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," Proc. European Conf. Machine Learning, pp. 171-182, 1994.