ISSN:0975-9646

A Survey of Analyzing the Efficiency Based Feature Selection of High Dimensional Data using Fast Algorithm

S.Nagendrudu
Assistant Professor,Santhiram Engineering College,JNTUA
Dept: I.T,Nandyal,.INDIA

Chakali Subba Govardhan
Santhiram Engineering College,JNTUA
Dept: M.C.A,Nandyal ,INDIA

Abstract: There are unit many algorithms applied to seek out the potency and effectiveness. Here we have a tendency to take into account the potency because the time taken to retrieve the data's from the information and effectiveness is from the foremost datasets (or) subsets that area unit relevant to the users search. By mistreatment FAST algorithmic rule we are able to retrieve the data's while not the orthogonal options. Here the orthogonal options area unit disbursed by means that of assorted levels of the question input and therefore the output the relevant data is disbursed just in case of the set choice and clump ways. These is fashioned in well equipped format and therefore the time taken for retrieve the knowledge are going to be short time and therefore the FAST algorithmic rule calculate the retrieval time of the information from the dataset. This algorithmic rule formulates as per the information out there within the dataset. During this paper, in the main focus regarding the small array pictures that don't seem to be mentioned within the previous work. By analyzing the potency of the projected work and therefore the existing work, the time taken to retrieve the information are going to be higher within the projected by removing all the orthogonal options that area unit gets analyzed.

Keywords - Feature selection, Clustering, Filter, MST.

I. INTRODUCTION

In machine learning and statistics, feature choice additionally called variable choice, attribute choice or variable set choice. it's the method of choosing a set of relevant options to be used in model construction. The central assumption once employing a feature choice technique is that the info contains several redundant or unsuitable options. Redundant options square measure those which offer no additional info than the presently elite options, and unsuitable options offer no helpful info in any context[1-6]. Feature choice techniques square measure a set of the additional general field of feature extraction. Feature extraction creates new options from functions of the initial features[9,10], whereas feature choice returns a set of the options. Feature choice techniques square measure usually employed in domains wherever there square measure several options and relatively few samples or knowledge points. The prototypal case is that the use of feature choice in analyzing deoxyribonucleic acid microarrays, wherever there square measure several thousands of options, and a number of tens to many samples. Feature choice techniques offer 3 main edges once constructing prognosticative models: improved model interpretability, shorter coaching times, and increased by reducing over fitting.

A feature choice algorithmic rule will be seen because the combination of an enquiry technique for proposing new feature subsets, along side Associate in Nursing analysis live that scores the various feature subsets. the only algorithmic rule [13] is to check every doable set of options finding the one that minimizes the error rate. this is often Associate in Nursing complete search of the house, and is computationally recalcitrant for just about the littlest of feature sets. the selection of analysis metric heavily influences the algorithmic rule, and it's these analysis metrics that distinguish between the 3 main classes of feature choice algorithms: wrappers, filters and embedded methods[10].

Wrapper strategies use a prognosticative model to attain feature subsets. every new set is employed to coach a model, that is tested on a hold-out set. investigating the amount of mistakes created thereon hold-out set (the error rate of the model) provides the score for that set. As wrapper strategies train a brand new model for every set, they're terribly computationally intensive, however sometimes give the simplest acting feature set for that individual style of model.

Filter strategies use a proxy live rather than the error rate to attain a feature set. This live is chosen to be quick to cypher, while still capturing the quality of the feature set. Common measures include the Mutual Information, Pearson product-moment coefficient of correlation, and inter/intra category distance[14]. Filters are sometimes less computationally intensive than wrappers, however they manufacture a feature set that isn't tuned to a selected style of prognosticative model. several filters give a feature ranking instead of an exact best feature set, and also the cutoff purpose within the ranking is chosen via cross-validation.

Embedded strategies are a catch-all cluster of techniques that perform feature choice as a part of the model construction method. One different fashionable approach is that the algorithmic Feature Elimination algorithmic rule, unremarkably used with Support Vector Machines to repeatedly construct a model and take away options with low weights. These approaches tend to be between filters and wrappers in terms of procedure quality.

Subset choice evaluates a set of options as a bunch for suitableness. set choice algorithms will be broken into

Wrappers, Filters and Embedded. Wrappers use a research algorithmic rule to look through the area of doable options and appraise every set by running a model on the set.

Wrappers will be computationally dearly-won and have a risk of over fitting to the model. Filters are kind of like Wrappers within the search approach, however rather than evaluating against a model, a less complicated filter is evaluated. Embedded techniques ar embedded in and specific to a model[7].

The Fast clustering Based Feature Selection algorithm (FAST) works into 2 steps. In the first step, features are divided into clusters by using graph theoretic clustering methods. In the second step, the most representative feature i.e., strongly related to target class is selected from each cluster to form the final subset of features. A feature in different clusters in relatively independent, the clustering based strategy of FAST has high probability of producing a subset of useful and independent features. 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 four well known different types of classifiers. A good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. Different from these algorithms, our proposed FAST algorithm employs clustering based method to features.

This paper discusses about feature selection, FAST algorithm, Text classification and so on. The next section is literature review.

II. LITERATURE REVIEW

A. An efficient approach to clustering in large multimedia databases with noise.

Several clustering algorithms can be applied to clustering in large multimedia databases. The effectiveness and efficiency of the existing algorithms, however is somewhat limited since clustering in multimedia databases requires clustering high-dimensional feature vectors and since multimedia databases often contains large amounts of noise. Using DENCLUE algorithm we can remove noise from the large databases. The main advantage includes, that the clustering can be done efficiently in high dimensional database.

B. Feature subset selection using the wrapper method: over fitting and dynamic search space topology

In the feature subset selection, a search for an optimal set of features is made using the induction algorithm as a black box. The best-first search is to find good feature subsets. The over fitting problems can be reduced by using the best first search. The relevant and optimal features can be easily selected in this method, and also the over fitting can be reduced.

C. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

Partitioning an outsized set of objects into unvaried clusters may be a basic operation in data processing. The k-means formula is best fitted to implementing this operation thanks to its potency in clump giant information sets. clump ways partition a collection of objects into clusters specified objects within the same cluster are a lot of almost like one another than objects in numerous clusters. the foremost distinct characteristic of information mining is that it deals with terribly giant data sets (gigabytes or maybe terabytes). This needs the algorithms utilized in data processing to be scalable. However, most algorithms presently utilized in data processing don't scale well once applied to terribly giant information sets as a result of they were at the start developed for different applications than data processing that involve little information sets. We tend to quick clump formula accustomed cluster categorical information. The most blessings of this methodology is that, will simply reason the objects and therefore the dissimilar objects are often removed simply.

D. Fast and Effective Text Mining Using Linear-time Document Clustering

Clustering is a powerful technique for large-scale topic discovery from text. It involves two phases: first, feature extraction maps each document or record to a point in high-dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Document clustering helps tackle the information overload problem in several ways. One is exploration; the top level of a cluster hierarchy summarizes at a glance the contents of a document collection. Also the features are extracted efficiently.

E. Irrelevant Features and the Subset Selection Problem

We address the problem of ending a subset of features that allows a supervised induction algorithm to induce small high accuracy concepts. We examine notions of relevance and irrelevance_ and show that the definitions used in the machine learning literature do not adequately partition the features into useful categories of relevance. The features selected should depend not only on the features and the target concept_ but also on the induction algorithm. We describe a method for feature subset selection using cross validation that is applicable to any induction algorithm. In this the relevant features alone are extracted.

III. EXISTING SYSTEM

In the past approach there are several algorithm which illustrates how to maintain the data into the database and how to retrieve it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology.

A Distortion algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records.

A Blocking algorithm make propagation to the above problem, and reduce the problems occurred in the existing distortion algorithm, but here also having the problem called data overflow, once the user get confused then they can never get the data back.

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. Traditional machine

learning algorithms like decision trees or artificial neural networks are examples of embedded approaches.

The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy 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 independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. 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. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

A. Drawbacks of Existing System

- Lacks speed
- Security Issues
- Performance Related Issues
- The generality of the selected features is limited and the computational complexity is large.
- Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

So the focus of our new system is to enhance the throughput for any basis to eliminate the data security lacks therein and make a newer system prominent handler for handling data in an efficient manner.

IV. PROPOSED SYSTEM

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant 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). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features.

Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F 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.

A. Advantages

1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

B. In our proposed FAST algorithm, it involves

1. The construction of the minimum spanning tree (MST) from a weighted complete graph;
2. The partitioning of the MST into a forest with each tree representing a cluster;
3. The selection of representative features from the clusters.

V. FEATURE SUBSET SELECTION ALGORITHM

Framework and Definitions

Feature subset selection should be able to identify and eliminate irrelevant and redundant information as possible. Because irrelevant and redundant features severely affect the accuracy of the learning machines. So we develop a novel algorithm to deal with both irrelevant and redundant features. Finally, it will obtain a good feature subset.

Figure l shows the feature subset selection using our proposed FAST algorithm. In this, irrelevant and redundant features are removed. To eliminate the irrelevant and redundant features, our FAST algorithm involves three steps. 1) MST is constructed 2) MST is partitioned into a forest 3) Selecting most representative features from each clusters.

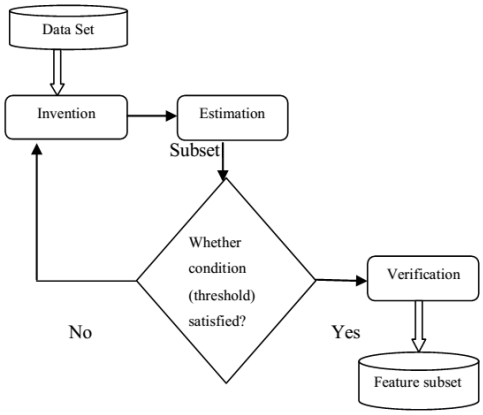

Figure1. Framework for the proposed feature subset selection algorithm

Table 1: Comparison of various techniques and algorithms

S.NO	Techniques (or)Algorithms	Advantages	Disadvantages
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.	Agglomerative linkage algorithm	Reduce Complexity	Decrease the Quality when dimensionality become high
6.	INTERACT Algorithm	Improve Accuracy	Only deal with irrelevant data
7.	Distributional clustering	Higher classification accuracy	Difficult to evaluation
8.	Relief Algorithm	Improve efficiency and Reduce Cost	Powerless to detect Redundant features

VI. CONCLUSION

For the entire Fast algorithm in hands with association rule implementation gives flexible results to users, like removing irrelevant features from the Original Subset, and constructing a minimum spanning tree from the relative subset whatever present in the data store. By partitioning the minimum spanning tree we can easily identify the text representation from the features. Association Rule Mining gives ultimate data set with header representation as well as FAST algorithm with applied K-Means strategy provides efficient data management and faster performance. The revealing regulation set is significantly smaller than the association rule set, in particular when the minimum support is small. The proposed work has characterized the associations between the revealing regulation set and the non-redundant association rule set, and discovered that the enlightening regulation set is a subset of the non-redundant association rule set.

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. H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.
3. H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.
4. 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.
5. L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
6. Christos Boutsidis, Michael W. Mahoney, Petros Drineas, "An Improved Approximation Algorithm for the Column Subset Selection Problem" S. Chen and J. Wigger," Fast Orthogonal Least Squares Algorithm for Efficient Subset Model Selection", IEEE TRANSACTIONS ON SIGNAL PROCESSING. VOL. 43. NO 1, JULY 1995 1713.
7. Charu C. Aggarwal Cecilia Procopiuc IBM T. J. Watson, "Fast Algorithms for Projected Clustering", Research Center Duke University Yorktown Heights, NY 10598 Durham, NC 27706.
8. Alexander Hinneburg, Daniel A.Keim," An efficient approach to clustering in Large Multimedia Databases with Noise" Institute of Computer Science, University of Halle, Germany.
9. Ron Kohavi and Dan Sommer_eld, "Feature Subset Selection Using the Wrapper Method: Over_tting and Dynamic Search Space Topology", Appears in the First International Conference on Knowledge Discovery and Data Mining (KDD-95)
10. Bjornar Larsen and Chinatsu Aone,"Fast and Effective Text Mining Using Linear-time Document Clustering" SRA International, Inc.
11. Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining",Cooperative Research Centre for Advanced Computational Systems.
12. Saowapak Sotthivirat and Jeffrey A. Fessler "Relaxed ordered-subset algorithm for penalizedlikelihood image restoration" Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 S. Sotthivirat and J. A. Fessler Vol. 20, No. 3/March 2003/J. Opt. Soc. Am. A 439.
13. Bin Zhang, Member, IEEE, and Sargur N. Srihari, Fellow, IEEE,"Fast k-Nearest Neighbor Classification Using Cluster-Based Trees", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 26, NO. 4, APRIL 2004.
14. Jihoon Yang and Vasant Honavar, "Feature Subset Selection Using A Genetic Algorithm Artificial Intelligence Research Group, Department of Computer Science, Iowa State University.