iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XI **Month of publication:** November 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A FAST Algorithm for High Dimensional Data using Clustering-Based Feature Subset Selection

Puppala Priyanka[1], M Swapna[2]
[1]Department of CSE, M.Tech, AVN Inst. of Engg. & Tech., Hyderabad.
[2]Associate Professor, Department of CSE, AVN Inst. Of Engg. & Tech., Hyderabad.

Abstract: — Feature subset clustering is a powerful technique to reduce the dimensionality of feature vectors for text classification and involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A novel approach called supervised attribute clustering algorithm is proposed to improve the accuracy and check the probability of the patterns. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. Efficiency is related to the time required to find a subset of features while the effectiveness is related to quality of subset of features.Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree clustering method.
Index Terms—Feature subset selection, Filter method, feature clustering, graph-theoretic clustering, MST

I. INTRODUCTION

The aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Feature subset selection is an effective way for reducing dimensionality, eliminating irrelevant data and redundant data, increasing accuracy. There are various feature subset selection methods in machine learning applications and they are classified into four categories: Embedded, wrapper, filter and hybrid approaches. Based on the MST method, we propose a Fast clustering-bAsed feature Selection algoriThm (FAST).The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features.

Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The proposed feature subset se- lection algorithm FAST was tested upon 35 publicly available image, microarray, and text data sets. 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. 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.

II. FEATURE SELECTION

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many 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. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

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 filtermethods. Wrapper methods are widely recognized as a superior alternative in supervised learning problems, since by employing the inductive algorithm to evaluate alternatives they have into account the particular biases of the algorithm. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, 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.

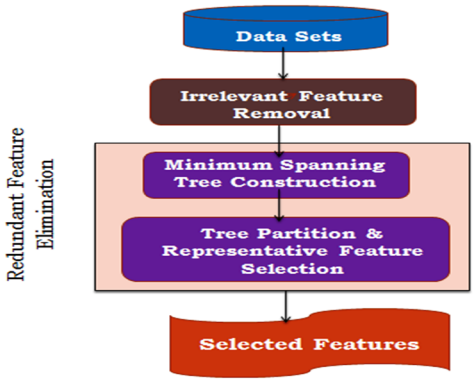
III. PROBLEM DEFINITON

Several algorithms 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. The systems like Distortion and blocking 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, once the user get confused then they can never get the data back. 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. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

IV. FEATURE SUBSET SELECTION ALGORITHM

The central assumption when using a feature selection technique is that the data contains many 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. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). Feature selection techniques provide three main benefits when constructing predictive models: Improved model interpretability, Shorter training times, enhanced generalization by reducing over fitting. Feature selection is also useful as part of the data analysis process, as shows which features are important for prediction, and how these features are related.

A. 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. The algorithm for feature selection that

clusters attributes using a special metric and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. The main interest of our technique resides in the improved understanding of the structure of the analyzed data and of the relative importance of the at-tributes for the selection process.

B. Graph-theoretic clustering:

Graph-theoretic clustering are partition vertices in a large graph into different clusters. Both coarse clustering and fine clustering are based on this algorithm called dominant-set clustering. It produces fine clusters on incomplete high dimensional data space. These algorithms that are held to execute well with respect to the indices explain as in the previous section are outlined. The first iteratively emphasise the intra-cluster over inter-cluster connectivity and the second is repeatedly refines an initial partition based on intra-cluster conductance. While together essentially work locally, we also suggest another, more global method. In all three cases, the asymptotic worst-case running time of the algorithms based on certain parameters known as input. However, see that for important choices of these parameters, the time complexity of the novel algorithm GM is superior than for the other two algorithms.

V. IRRELEVANT FEATURES REMOVAL

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, 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 develop 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 elimination is a bit of sophisticated.

In our proposed FAST algorithm, it involves-

1) The construction of the minimum spanning tree from a weighted complete graph;

2) The partitioning of the MST into a forest with each tree representing a cluster; and

3) The selection of representative features from the clusters.

The symmetric uncertainty () 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 by a number of researchers. Therefore, we choose symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept.

Symmetric uncertainty of variables and, the relevance T-Relevance between a feature and the target concept, the correlation F-Correlation between a pair of features, the feature redundance F-Redundancy and the representative feature R-Feature of a feature cluster can be defined as follows.

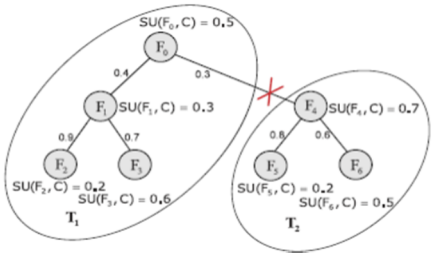


Fig. F-Correlation

Definition1: (T-Relevance)

The relevance between thefeature $F_i \in F$ and the target concept C is referred to as the T-Relevance of Fiand C , and denoted by (,).

If (,)is greater than a predetermined threshold, we say that is a strong T-Relevance feature.

Definition 2: (F-Correlation)

The correlation between any pair of features and (, ∈∧ ≠) is called the F-Correlation of and , and denoted by (,).

According to the above definitions, feature subset selection can be the process that identifies and retains the strong T-Relevance features and selects R-Features from feature clusters.

The behind heuristics are that

1) Irrelevant features have no/weak correlation with target concept;
2) Redundant features are assembled in a cluster and a representative feature can be taken out of the cluster.

VI. CONCLUSION

The purpose of cluster analysis has been established to be more effective than feature selection algorithms. Since high dimensionality and accuracy are the two major concerns of clustering, we have considered them together in this paper for the finer cluster for removing the irrelevant and redundant features. The proposed supervised clustering algorithm is processed for high dimensional data to improve the accuracy and check the probability of the patterns. Retrieval of relevant data should be faster and more accurate. This displays results based on the high probability density thereby reducing the dimensionality of the data.

VI. ACKNOWLEDGMENT

We would like to thank everyone who has motivated and supported us for preparing this manuscript.

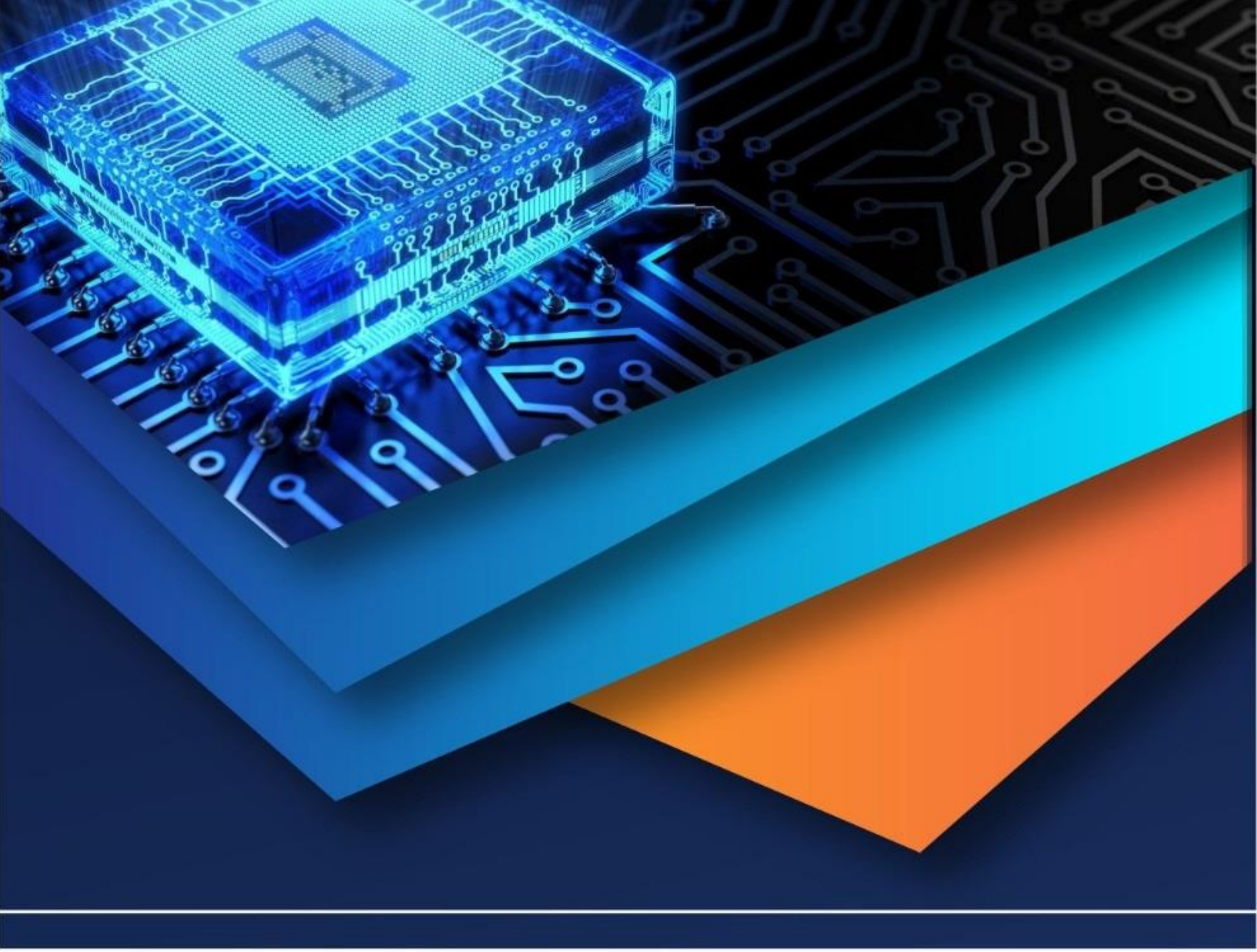
REFERENCES

[1] Dash M. and Liu H., Consistency-based search in feature selection. Artificial Intelligence, 151(1-2), pp 155-176, 2003.
[2] Demsar J., Statistical comparison of classifiers over multiple data sets, J.Mach. Learn. Res., 7, pp 1-30, 2006.
[3] Fleuret F., Fast binary feature selection with conditional mutual Information,Journal of Machine Learning Research, 5, pp 1531-1555, 2004.
[4] Forman G., An extensive empirical study of feature selection metrics fortext classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003..
[5] Garcia S and Herrera F., An extension on "Statistical Comparisons ofClassifiers over Multiple Data Sets" for all pairwisecomparisons, J. Mach.Learn. Res., 9, pp 2677-2694, 2008.
[6] Guyon I. and Elisseeff A., An introduction to variable and feature selection,Journal of Machine Learning Research, 3, pp 1157-1182, 2003.
[7] Hall M.A., Correlation-Based Feature Selection for Discrete and NumericClass Machine Learning, In Proceedings of 17th International Conferenceon Machine Learning, pp 359-366, 2000.
[8] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Machine Learning Research, vol. 3,pp. 1289-1305, 2003
[9] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of Relief and ReliefF," Machine Learning, vol. 53, pp. 23-69, 2003.
[10] 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.
[11] H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.
[12] M. Last, A. Kandel, and O. Maimon, "Information-Theoretic Algorithm for Feature Selection," Pattern Recognition Letters, vol. 22, nos. 6/7, pp. 799-811, 2001.
[13] C. Krier, D. Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data," Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162,2007

AUTHORS

Puppala Priyanka[1] , I am pursuing Post Graduate in Master of Technology with specialization of Computer Science &Engg. at AVN Inst. of Engg.& Tech, Hyderabad, AP, India. My interested research area is Data warehousing & Data Mining, Network Security and Data Structures.

Mrs. M Swapna[2] is the Associate Professor, Dept. of CSE, AVN Inst.OfEngg.& Tech, Hyderabad, AP, India. She received his M.Tech. from JNTU Hyderabad . Her expertise areas are Network Programming and Networking Security.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY