IJRRECS/May 2014/Volume-2/Issue-5/2225-2230 **ISSN 2321-5461**

**INTERNATIONAL JOURNAL OF REVIEWS ON RECENT ELECTRONICS AND COMPUTER SCIENCE**

# ALGORITHM OF FEATURE SELECTION INTENDED FOR HIGH DIMENSIONAL DATA

**B.Sankara Babu[1], Dr. K.Rajasekhar Rao[2]**

**[1]Associate Professor, Dept of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, A.P, India**

**[2]Professor, Dept of CSE, Koneru Laxmaiah University, Guntur, A.P, India**

**ABSTRACT:**

Feature subset selection can be imagined as the progression of identifying and elimination, as many inappropriate and redundant features as promising. Numerous feature subset selection methods have been planned and considered for machine learning applications. Feature subset selection can be analyzed as the process of recognizing and eliminating as many inappropriate and redundant features as promising since: inappropriate features do not put in to the predictive accurateness and redundant characteristics do not redound to getting an enhanced predictor for that they make available mainly information which is by now present in previous feature. We build up a novel algorithm that can capably and efficiently deal with both inappropriate and redundant characteristics, and get hold of a superior feature subset. Based on the minimum spanning tree method, we recommend a FAST algorithm. The algorithm is a two steps process in which, characteristics are divided into clusters by means of using graph-theoretic clustering means. In the subsequent step, the mainly used representative feature that is robustly related to target classes is particular from each cluster to structure the final subset of features. Features in altered clusters are comparatively autonomous; the clustering-based scheme of FAST has a high possibility of producing a subset of constructive and independent characteristics. In our projected FAST algorithm, it entails the building of the minimum spanning tree from a subjective inclusive graph; the separation of the minimum spanning tree into a forest by means of every tree signifying a cluster; and the collection of representative features from the clusters.

***Keywords: Feature subset selection, fast clustering-based feature Selection algorithm, Minimum spanning tree, Cluster.***

## 1. INTRODUCTION:

Feature subset selection is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Selection of Feature subset is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Since inappropriate features do not put in to the predictive accuracy and redundant characteristics do not redound to reaching an improved predictor for that they afford for the most part of information which is previously present within other attribute. Numerous feature subset selection methods have been planned and considered for machine learning applications [4]. They can be separated into four major categories such as: the Wrapper, Embedded, and Filter and Hybrid methods. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centres or separated by means of a normal geometric curve and have been extensively used in tradition [8]. FAST algorithm makes use of minimum spanning tree based scheme to cluster features. The projected feature subset selection algorithm FAST was evaluated with other various types of feature subset selection algorithms, the algorithm not only decrease the number of features, but also advances the performances of the renowned various types of classifiers [1]. It does not presume that data points are clustered around centres otherwise separated by means of a standard geometric curve. FAST does not limit to several specific types of data. Feature subset selection is the process of recognizing and eliminating as many inappropriate and redundant features as promising in view of the fact that: inappropriate features do not put in to the predictive accurateness and redundant characteristics do not redound to getting an enhanced predictor for that they make available mostly the earlier information [12]. Based on the minimum spanning tree scheme, we recommend a FAST algorithm which is a two steps process in which; characteristics are divided into clusters by means of using graph-theoretic clustering means and in the subsequent step, the

mainly used representative feature that is robustly related to target classes is particular from each cluster to structure the final subset of features. The clustering-based system of FAST has a high opportunity of producing a subset of productive and independent characteristics [3].

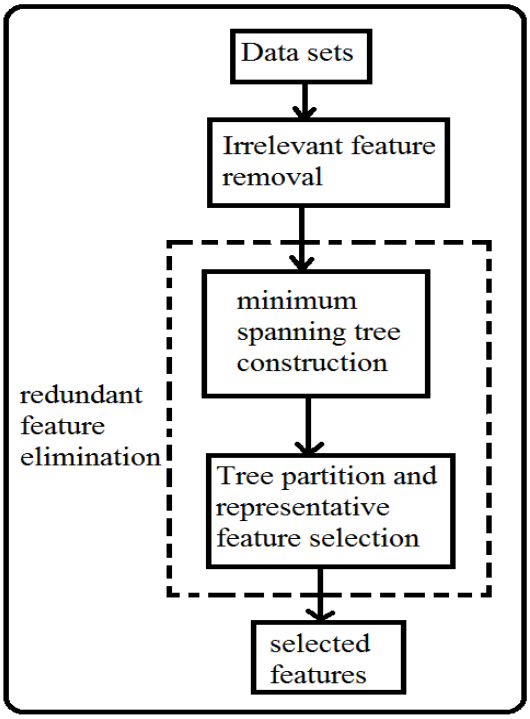


**Fig1: An overview of feature subset selection algorithm**

## 2. METHODOLOGY:

Inappropriate features, all along with redundant description, severely have an effect on the correctness of the learning machines. Consequently, feature subset assortment should be able to recognize and take away as much of the unrelated and redundant information as probable [7]. In addition, superior feature subsets enclose features extremely linked with the class, so far uncorrelated with each other. We build up a novel algorithm shown in fig1 which can capably and efficiently deal with both inappropriate and redundant characteristics, and get hold of a superior feature subset. We attain this all the way through a novel characteristic selection construction which composed of the two associated components of removal of irrelevant features and elimination of redundant feature [2] [14]. The removal of irrelevant features obtains features applicable to the target notion by means of removing inappropriate ones, and the elimination of redundant feature removes redundant characteristics from applicable ones by means of preferring representatives from various feature clusters, and consequently produces the concluding subset [5]. The removal of irrelevant feature is uncomplicated formerly the right significance assess is defined, although the elimination of redundant feature is a bit of complicated. FAST algorithm entails the building of the minimum spanning tree from a subjective inclusive graph; the separation of the minimum spanning tree into a forest by means of every tree signifying a cluster;

and the assortment of representative features from the clusters [10]. For the most part of the information contained in redundant characteristics is by now present in other characteristics. Consequently, redundant features do not add to getting better interpreting capability to the target idea. In order to more exactly commence the algorithm, and for the reason that features subset selection structure involves inappropriate feature removal and redundant feature removal. Appropriate features have tough correlation with target idea so are constantly essential for a best subset, whereas redundant characteristics are not since their values are entirely simultaneous with each other [9] [13]. Consequently, notions of feature redundancy and feature significance are normally in terms of feature association and feature-target concept association. Mutual information computes how much the allocation of the feature values and target classes are at variance from statistical freedom [6]. This is a nonlinear inference of association between target classes as well as feature values. The symmetric uncertainty is derived from the shared information by means of regularizing it to the values of feature entropies as well as target classes, and has been used to assess the integrity of features intended for classification [11]. Symmetric uncertainty was chosen as the measure of correlation among two features in addition to the target concept. Symmetric uncertainty cares for a pair of variables symmetrically and it compensates for gain of information bias toward variables with added values.

## 3. RESULTS:

FAST executes extremely well on the microarray data and obtains first rank of for microarray data. Microarray data has the environment of the large number of characteristics other than small sample size, which can cause curse of dimensionality. In the presence of numerous features, researchers become aware of that a large number of characteristics are not instructive because they are moreover inappropriate or superfluous with respect to the class concept. Consequently, choosing a small number of discriminative genes from numerous genes is necessary for booming sample categorization. FAST efficiently filters out a mass of inappropriate features which reduces the likelihood of inappropriately bringing the inappropriate features into the succeeding analysis. FAST eliminates a large number of outmoded

features by means of choosing a single representative characteristic from each cluster of outmoded features. Consequently, only a very small number of discriminative characteristics are selected.

## 4. CONCLUSION:

On the basis of minimum spanning tree method, we recommend a FAST algorithm which is a two steps process in which, characteristics are divided into clusters by means of using graph-theoretic clustering means. The process of recognizing and eliminating as many inappropriate and redundant features as promising can be examined by the feature subset selection. The clustering-based system of FAST has a high prospect of producing a subset of practical and self-governing characteristics. FAST algorithm necessitates the building of the minimum spanning tree from a subjective inclusive graph; the separation of the minimum spanning tree into a forest by means of every tree signifying a cluster; and the collection of representative features from the clusters. FAST removes a large number of outmoded features by means of choosing a single representative characteristic from each cluster of outmoded features.

## REFERENCES:

[1] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96- 103, 1998.

[2] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., and Caligiuri, M. A., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439), pp 531-537, 1999.

[3] Yu J., Abidi S.S.R. and Artes P.H., A hybrid feature selection strategy for image defining features: towards interpretation of optic nerve images, In Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 8, pp 5127-5132, 2005.

[4] Hall M.A. and Smith L.A., Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper, In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.

[5] Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162, 2007.

[6] Usama M. F. and Keki B., Irani: Multi-interval discretization of continuousvalued attributes for classification learning, In Proceedings of 13th International Joint Conference on AI, pp 1022-1027, 1993.

[7] Park H. and Kwon H., Extended Relief Algorithms in Instance-Based Feature Filtering, In Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), pp 123-128, 2007.

[8] Van Dijk G. and Van Hulle M.M., Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, International Conference on Artificial Neural Networks, 2006.

[9] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.

[10] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.

[11] Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.

[12] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[13] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.

[14] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279- 305, 1994.