IJRRECS/October 2013/Volume-1/Issue-6/1295-1300 **ISSN 2321-5461**

INTERNATIONAL JOURNAL OF REVIEWS ON RECENT ELECTRONICS AND COMPUTER SCIENCE

A NOVEL FEATURE SUBSET ALGORITHM FOR HIGH DIMENSIONAL DATA

S.Swetha[1], A.Harpika[2]

[1]M.Tech Student, Dept of CSE, TKR College of Engineering & Technology, Hyderabad, A.P, India

[2]Assistant Professor, Dept of CSE, TKR College of Engineering & Technology, Hyderabad, A.P, India

ABSTRACT:

Selection of Feature subset is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Numerous feature subset selection methods have been planned and considered for machine learning applications. Feature subset selection can be analysed 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. Numerous feature subset selection methods have been planned and considered for machine learning applications [2]. 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 [13]. The projected feature subset selection algorithm FAST was tested and the investigational results demonstrate that, evaluated with other various types of feature subset selection algorithms, the projected algorithm not only decrease the number of features, but also advances the performances of the renowned various types of classifiers [5] [9]. Feature subset selection can be analysed 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 [8] [11]. 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 [1]. 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 [3].

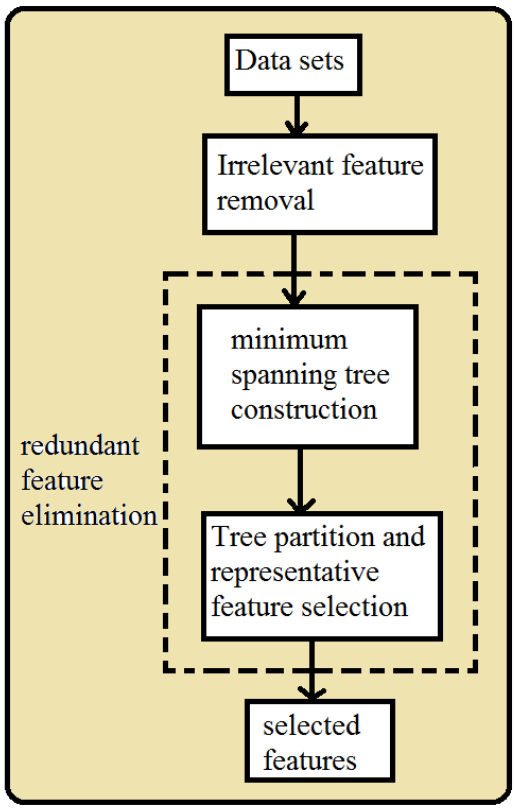

Fig1: An overview of feature subset selection algorithm

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. In addition, superior feature subsets enclose features extremely linked with the class, so far uncorrelated with each other [6]. 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 [14]. The earlier obtains features applicable to the target notion by means of removing inappropriate ones, and the concluding removes redundant characteristics from applicable ones by means of preferring representatives from various feature clusters, and consequently produces the concluding subset [4] [10]. 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. 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 [12]. 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 our projected feature 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 [15]. 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 [7]. This is a nonlinear inference of association among feature values and target classes.

3. RESULTS:

FAST executes extremely well on the microarray data. The explanation lies in both the features of the data set itself and the property of the projected algorithm. 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 it is general 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. Our projected FAST efficiently filters out a mass of inappropriate features in the initial step which reduces the likelihood of inappropriately bringing the inappropriate features into the succeeding analysis. Then, in the subsequent step, FAST eliminate 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:

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. Feature subset selection can be analysed 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. 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. The projected feature subset selection algorithm FAST was tested and the investigational results demonstrate that, evaluated with other various types of feature subset selection algorithms, the projected algorithm not only decrease the number of features, but also advances the performances of the renowned various types of classifiers.

REFERENCES:

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45,
1992.

[2] 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.

[3] 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.

[4] 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.

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

[6] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset

selection, Machine Learning, 41(2), pp 175-195, 2000.

[7] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.

[8] 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.

[9] Cardie, C., Using decision trees to improve case-based learning, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.

[10] 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.

[11] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.

[12] Cohen W., Fast Effective Rule Induction, In Proc. 12th international Conf. Machine Learning (ICML'95), pp 115-123, 1995.

[13] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[14] Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.

[15] 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.