www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 7 July, 2014 Page No. 7102-7105

Feature Subset Selection Algorithm for Elevated Dimensional Data By using Fast Cluster

B.Swarna Kumari, M.Doorvasulu Naidu
M.Tech: Department of CSE
SISTK, Puttur, INDIA
swrarna1259@gmail.com.

Assistant Professor
Department of CSE
SISTK, Puttur, INDIA
doorvasulunaidum@gmail.com

Abstract- ***Feature selection involves recognizing a subset of the majority helpful features that produces attuned results as the unique set of features. Feature selection algorithm can be evaluated from mutually efficiency and effectiveness points of vision. FAST algorithm is***
Proposed and then experimentally evaluated in this paper. FAST algorithm mechanism considering two steps. In the primary step, features are separated into clusters by means of graph-theoretic clustering methods. In the subsequent step, the majority delegate feature that is robustly connected to target classes is chosen from each cluster form a subset of features. The Features in unusual clusters are relatively self-governing; the clustering-based approach of FAST has a elevated possibility of producing a subset of useful features. in the direction of guarantee to the efficiency of FAST, we implement the efficient minimum-spanning tree clustering technique. general experiments are approved to contrast FAST and some delegate feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, by admiration to four types of famous classifiers, specifically, the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER and following feature selection.

Keywords-feature subset selection, graph-theoretic clustering , feature selection;

I. RELATED WORK

Feature subset selection can be viewed as the method of identifying and removing a lot of unrelated and unnecessary features as probable. This is the reason that: (i) immaterial features do not give the predictive correctness [1], and (ii) unnecessary features do not redound to receiving a superior predictor for that they give main data which is previously there in additional feature(s).

There are numerous feature subset selection algorithms , a few can successfully remove immaterial features but not succeed to hold unnecessary features [2], [3], [4], [5], [6], [7], however a few of others can remove the immaterial while taking concern of the unnecessary features [8], [9], [10],[11]. Our proposed FAST algorithm cascade into the subsequent group.

usually feature subset selection study has been alert on searching for important features. A famous example is Relief [5], it weighs every feature according to its capability to classify instances under dissimilar targets based on the distance-based criteria purpose. though Relief is unsuccessful at removing unnecessary features as two predictive but greatly correlated features are occurred and both are highly weighted [12]. Relief-F [4] extends Relief, this technique is enabling to work with noisy and incomplete data sets and it can deals with multi-class problems, but still cannot recognize unnecessary features.

though, along with immaterial features, unnecessary features also change the speed and correctness of learning algorithms, and thus could be eliminated as well[12], [13], [3]. CFS [9], FCBF [11] and CMIM [14] are

the examples that capture into concern the unnecessary features. CFS [9] is achieved by the assumption that a good feature subset is that contains features are greatly correlated with the target, yet uncorrelated with each other. FCBF ([11], [15]) is a fast filter technique which can recognize important features as well as redundancy among important features without pair off correlation analysis. CMIM [14] iteratively picks features which makes most of their common information with the class to predict, conditionally
to the comeback of any feature that has already picked. FAST algorithm is Different from these algorithms, FAST algorithm employs clustering based technique to select the features.

newly, hierarchical clustering has been adopted in word collection in the context of text classification (e.g.,[16],[17], and [18]). Distributional clustering is helpful to cluster words into groups based on their involvement in particular grammatical relations with additional words by Pereira et al. [16] or on the sharing of class labels linked with each word by BakeandMcCallum [17]. in natural history distributional clustering of words are
Agglomerative , and result in sub-optimal word clusters and elevated computational price, Dhillon et al. [18]proposed a latest information-theoretic divisive algorithm for word clustering and useful it to text classification .Butterworth et al. [19] proposed to cluster features using a unique metric of Barthelemy-Montjardet distance. and then makes use of the dendrogram of the resulting cluster. in addition, the obtained correctness is lesser when compared with other feature selection methods.

our proposed FAST algorithm is different from these hierarchical clustering based algorithms and it make use of minimum spanning tree based technique to cluster features .in the meantime, it does not suppose that data points are grouped around centers or separated by a ordinary geometric curve. our proposed FAST does not limit to some exact types of data.

II INTRODUCTION

with respect to the target concepts, the aim of selecting a subset of good features. for reducing dimensionality and removing immaterial data subset selection is considering as an effective way .which can increasing learning accuracy, and improving result clarity[20],[21]. Feature selection algorithms can be divided into four broad categories: they are Embedded, Wrapper, Filter, and Hybrid approaches.

The embedded methods include feature selection as a part of the training process. the examples of embedded approaches are Traditional machine learning algorithms like decision trees or artificial neural networks [22]. To determine the goodness of the selected subsets the wrapper method is used. the correctness of the learning algorithms is usually high. However, the simplification of the selected features is limited and the computational complexity is elevated. The filter methods are self-governing of learning algorithms, with good simplification. By combining filter and wrapper methods the hybrid method occurred.

graph-theoretic methods have been considered in cluster analysis and used in many applications. Their outcomes gives the best agreement with human performance [23]. The general graph-theoretic clustering is uncomplicated. Compute a neighborhood graph of instances, then by deleting any edge in the graph that is shorter (according to some criterion) than its neighbors. The outcome can be in the form of a cluster. In this research, we concern graph theoretic clustering methods to features. Here assume the minimum spanning tree (MST) based on clustering algorithms.

subset of features. By considering the MST method, Fast clustering-bAsed feature Selection algoriThm (FAST) is proposed. The FAST algorithm mechanism has two steps. In the primary step, features are separated into clusters with the help of graph-theoretic clustering methods. In the subsequent step, the the majority representative Feature that is powerfully related to target classes is selected from every cluster to appearance the final subset of features.

III WORKING OF ALGORITHM

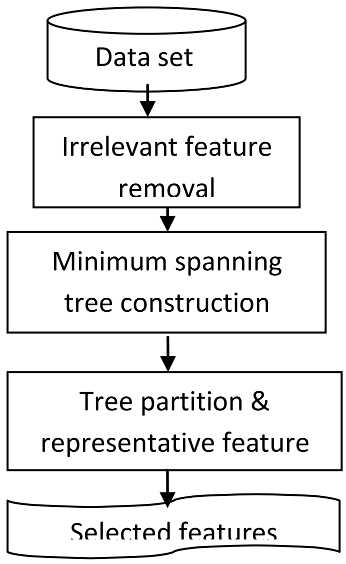

Fig: Framework of the proposed feature subset selection algorithm

The proposed FAST algorithm sensibly consists of three steps: (i) removing immaterial features, (ii)a MST is constructed from relative ones, and (iii) the MST is portioned and then selecting representative features.

A. First step:

the data set D *with* m *features* $F = \{F1, F2, ..., Fm\}$ and class C, we compute the *T-significance* $SU(Fi, C)$ value for every feature Fi $(1 \leq i \leq m)$.

B. Second step:

Here first calculate the *F-Correlation* $SU(F'i, F'j)$ value for each pair of features $F'i$ and $F'j$ Then, seeing features $F'i$ and $F'j$as vertices and $SU(F'i, F'j)$ the edge between vertices $F'i$ and $F'j$, a weighted complete graph $G = (V,E)$ is constructed. And it is an undirected graph. The complete graph reflects the correlations among the target-relevant features. Thus the edges shown as minimum, using the well-known Prims algorithm [24].

C. Third step:

Here unnecessary edges can be removed each tree $Tj \in$ *Forest* shows a cluster that is denoted as $V(Tj)$, which is the vertex set of Tj. for each cluster $V(Tj)$. select a representative feature FjR whose *T-Relevance* $SU(FjR, C)$ is the highest. All FjR (j = 1...| *Forest*|) consist of the final feature subset $\cup FjR$.

IV. CONCLUSION

In this paper, a novel clustering-based feature subset selection algorithm is presented for elevated dimensional data. The algorithm includes (i) removing immaterial features, (ii) constructing a minimum spanning tree from comparative ones, and (iii) MST is partioned and then choosing representative features. the cluster consists of features . Each cluster is considering as a single feature and thus dimensionality is reduced.

The performance of the proposed algorithm can be compared with five famous feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the 35 openly accessible image ,microarray ,and text data from the four different aspects of the section of selected features, the proposed algorithm having the best proportion of selected features, the best runtime, and the best classification correctness for Naive Bayes,C4.5, and RIPPER.

With the FAST algorithm it is easy to originate the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification correctness of
the four different types of classifiers

REFFERENCES

[1] John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

[2] Forman G., An extensive empirical study of feature selection metrics fortext classification, Journal of Machine Learning Research, 3, pp 1289-1305, 2003.

[3] Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.

[4] Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF,In Proceedings of the 1994 European Conference on Machine Learning, pp171-182, 1994.

[5] Kira K. and Rendell L.A., The feature selection problem: Traditional method sand a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.

[6] Modrzejewski M., Feature selection using rough setstheory, In Proceedingsof the European Conference on Machine Learning, pp 213-226, 1993.

[7] Scherf M. and Brauer W., Feature Selection By Means of a FeatureWeighting Approach, Technical Report FKI-221-97,Institut fur Informatik,Technische Universitat Munchen, 1997.

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

[9] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.

[10] Liu H. and Setiono R., A Probabilistic Approach to Feature Selection: A Filter Solution, in Proceedings of the 13th International Conference on Machine Learning, pp 319-327, 1996.

[11] Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.

[12] Yu L. and Liu H., Feature selection for high-dimensional data: a fastcorrelation-based filter solution, in Proceedings of 20th International Conferenceon Machine Leaning, 20(2), pp 856-863, 2003.

[13] Kohavi R. and John G.H., Wrappers for feature subsetselection, Artif.Intell., 97(1-2), pp 273-324, 1997.

[14] Fleuret F., Fast binary feature selection with conditional mutual Information, Journal of Machine Learning Research, 5, pp 1531-1555, 2004.

[15] Yu L. and Liu H., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research, 10(5), pp 1205-1224,2004.

[16] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.

[17] Baker L.D. and Mc Callum 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.

[18] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3,pp 1265-1287, 2003.

[19] 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.
[20] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004).

[21] Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining ,pp 306-313, 2002.

[22] Mitchell T.M., Generalization as Search, Artificial Intelligence, 18(2), pp203-226, 1982.

[23] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517, 1992.

[24] Prim R.C., Shortest connection networks and some generalizations, Bell System Technical Journal, 36, pp 1389-1401, 1957.