ISSN:0975-9646

A New Clustering Based Algorithm for Feature Subset Selection

Nitin B Chopade [1], Beena S Khade [2]

[1]*Information Technology,Pune University*
DGOIFOE,Bhigwan,Pune-Maharastra,India

[2]*Information Technology,Pune University*
VIIT,Pune-Maharastra,India

***Abstract*— The major idea of feature selection is to choose a subset of key variables by eliminating features with modest or no predictive information. Feature selection can drastically improve the clarity of the consequential classifier models and often construct a model that simplifies better to hidden points. A feature Extraction algorithm may be estimated from efficiency as well as usefulness points of view.**
In the proposed work, a fast clustering-based feature selection algorithm is proposed based on these criteria. The algorithm has different steps. In the first step, features are separated into clusters by means of graph-theoretic clustering methods. In the next step, the most delegate feature that is robustly related to target classes is preferred from every cluster to make a subset of Features. Also, we use Prim's as well as Kruskal's algorithm for handling large data set with efficient time complexity. The proposed system also deals with the Feature interaction which is crucial for effective feature extraction. Most of the present algorithms only focus on handling irrelevant and redundant features.

***Keywords*— Clustering-Based, Feature selection, Feature Interaction**

I. INTRODUCTION

Feature selection facilitates us to focus the consideration of an induction algorithm in which features that are the most outstanding to visualize a target model. If the complete arithmetical distribution were known, using additional features could only get better results, in practical learning circumstances it may be better to use a compact set of features. A large number of features used as input for induction algorithms may cause the unique algorithm to be very incompetent as memory and time consumers, even turning them irrelevant. Feature extraction can concentrate the dimensionality of the data and may improve a learner either in terms of learning performance, generalization capacity or model simplicity. It further helps recognize and better understand the results obtained by the learner, lessen its degree of storage, decrease the noise generated by irrelevant or preventable features and eradicate these useless features. Therefore, many feature subset selection algorithms have been projected to handle the irrelevant features or unneeded features. Of these algorithms, some of them can effectively eliminate irrelevant features but can't discover redundant ones. we estimated the algorithm that removes the irrelevant as well as redundant data by means of different technique like feature interaction and existing base algorithms that will be helpful in the clustering methods.

A. Challenges

In feature extraction scenario, various algorithms have been proposed for feature selection In the existing FAST feature extraction algorithm, main focus is on both removing unnecessary as well as immaterial data that means it extracts only targeted features. Problem in existing algorithm is that when it removes irrelevant features, it considers only single feature and match up to that to the target feature. If that feature matches then it extracts that feature otherwise remove it. So in this existing method the challenge lies in the fact that if more than one feature are joint and they suit the target feature then it can be treated as relevant. Feature interface is the new challenge for identifying the applicable feature. In existing system where feature interface is not supported that proves to be the modification criteria for this work.

B. Purpose

Feature selection, also recognized as a changeable selection, attribute collection or variable extraction, is the process of selecting a compartment of applicable features for use in model creation. The central supposition, when using a feature selection method, is that the data contains many redundant or irrelevant features. Unneeded features are those which offer no extra information than the presently selected features, and irrelevant features supply no useful information in any context.

The main and important reason of the feature extraction algorithm is that they discover out or extract only targeted features out of many features. They don't measure the irrelevant and redundant data because irrelevant and redundant data affects the competence and effectiveness of the algorithm. In existing algorithm that uses a variety of different techniques to decide relevant features or removing irrelevant or redundant features , when it removes the irrelevant features it does not consider the interface of different features. Proposed algorithm not only removes the irrelevant features but also focus on interaction of the features.

C. *Objective of System*

Feature subset collection is a useful for reducing dimensionality, removing Irrelevent data, increasing learning accurateness, and improving the result clarity. In feature subset selection, we can shrink dimensionality of the data and improve the competence and effectiveness of the algorithm. In our projected algorithm, our center of attention is relevant features and to extract only those features that are more relevant and interactive with each other

II. RELATED WORK

Feature subset choice has been an active investigated topic since 1970s, and a large deal of research work has been published. Of the obtainable research work, most feature subset collection algorithms can effectively recognize the irrelevant features based on dissimilar assessment functions. Still very few of them can eliminate the redundant features and take the feature interface into consideration [2]. On the basis of whether they can deal with immaterial features, unnecessary features and the feature interaction, the obtainable feature subset collection algorithms can be grouped into three categories: (a) the algorithms that can only handle immaterial features; (b) the algorithms that can handle both unrelated and unneeded features; and (c) the algorithms that deals with unrelated and unneeded features while considering feature interface. Next, we give a brief analysis of the three categories correspondingly. Usually, the study work on feature subset selection has focused on search for relevant features. Feature weighting ranking algorithms [5] weigh features independently and level them based on their importance to the target concept.

Feature subset selection is the procedure of identifying and deleting as many unrelated and unneeded features as possible. This is because :(a) unrelated features do not donate to the prognostic accuracy [4], and (b) unneeded features do not redound to getting a superior predictor for that they offer frequent information which is previously present in another feature. Of the many feature subset collection algorithms, some can effectively remove irrelevant features but unsuccessful to handle redundant features [2], [6], [1], yet some of the others can remove the irrelevant while taking care of the unneeded features [4], [8], [2], [3]. Our proposed optimized algorithm falls into the second group.

Newly, hierarchical clustering has been adopted in word collection in the context of text categorization (e.g., [7], [2], and [4]). Distributional clustering has been used to organize words into groups based either on their involvement in accurate grammatical associations with other words by Pereira et al. [2] or on the division of class labels concerned with each word by Baker and McCallum [10]. As distributional clustering of words are collective in nature, and outcome in sub-optimal word clusters and high computational cost, Dillon et al. [6] designed a novel information-theoretic disruptive algorithm for word clustering and applied it to text classification. Butterworth et al. [4] intended to cluster features using a particular metric of distance, and then builds use of the dendrogram of the consequential cluster hierarchy to prefer the most appropriate attributes. Unfortunately, the cluster assessment measure based on Barthelemy-Montjardet distance does not recognize a feature subset that permits the classifiers to recover their original performance accuracy. Moreover, when compared with other feature selection methods, the obtained accuracy is lesser. A Features Set estimate hinged on Relief It used six in practice dataset from the UCI repository has been used. Three of them have classification crisis with discrete features, the subsequent two categorizations with discrete and incessant features, and the final one is approximation dilemma.

The learning algorithm is used to authenticate the superiority of feature selected is a classification and deterioration tree layer with trimming [9]. This procedure is implemented by the orange data mining scheme. On the whole, the non-parametric tests, specifically the Wilcox on and Friedman test are suitable for our dilemma. They are suitable since they suppose some, but incomplete consistency. They are more secure than parametric tests since they do not suppose regular distributions or homogeneity of difference. There is an another view among statisticians that outcome tests should not be performed at all since they are often imprecise, either due to confusion or by putting too much pressure on their results. The main drawback of the system is it calculates to low accuracy of the search process. On Feature Selection through Clustering introduces an algorithm for feature extraction that clusters attributes using a specific metric and, then utilizes a hierarchical clustering for feature subset selection. Hierarchical algorithms produce clusters that are placed in a cluster tree, which is known as a dendrogram. Clustering's are gained by extracting those clusters that are located at a suitable height in this tree. It uses more than a few data sets from the UCI dataset repository and, because of space limitations we discuss only the results obtained with the votes and zoo datasets, Bayes algorithms of the WEKA package were used for applying classifiers on data sets attained by projecting the first data sets on the sets of delegate attributes. Approach to characteristic selection is the chance of supervision of the process permitting the user to choose between quasi-equivalent attributes it faces classification problems that connect thousands of features and moderately a few numbers of examples are considered. We try to relate our techniques to this type of data.

III. IMPLEMENTATION DETAILS

The technologically driven world in which we live in has enlarged the need for human interface with the system, mainly with computer-based system that are used to accomplish a vast selection of tasks with the aim of helping the user in achieving goal.

In feature selection process the main focus is on the different features with having irrelevant and redundant contain so algorithm work on that features and get only relevant features that is helpful in reducing the dimensionality of the data set. The algorithm works in different steps like data set selection, Graph and MST tree construction, partition of tree in to different clusters and finally get the result as the final selected features.

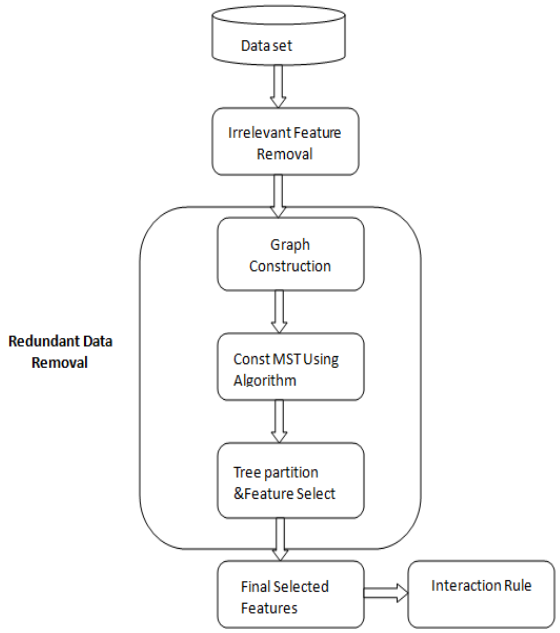

Figure 1: Framework of proposed feature selection algorithm

A. Feature selection algorithm architecture

The predictable system develops a novel algorithm which can competently and successfully deal with both unrelated and unneeded features. We achieve high-quality feature subset through a new feature selection.

Framework (shown in Fig.1) which is made up of the two linked components of unsuitable feature removal and redundant feature removal. The earlier obtains feature applicable to the target concept by eliminating inappropriate features, and the next eliminates redundant features from related ones via choosing representatives from dissimilar feature clusters, and hence constructs the last subset. The irrelevant feature elimination is easy once the right relevance is decide and defined or selected, while the redundant feature removal is a bit of complicated. In our proposed algorithm, it absorbs (a) the formation of the minimum spanning tree (MST) through a weighted complete graph; (b) the partitioning of the MST into a suitable forest with each tree indicating a cluster; and (c) the choice of representative features from the clusters.

B.Distributed clustering

Based on their donation in particular grammatical relations with other words, the Distributed clustering has been used to cluster words into groups. As distributional clustering of words is cooperative in nature, and effect in suboptimal word clusters and larger computational cost anticipated a novel information-theoretic disruptive algorithm for word clustering and useful it to text classification projected to cluster features by means of a specific metric of distance, and then it uses the resultant cluster hierarchy to settle on the most relevant attribute. Unfortunately, the cluster evaluation based on distance does not recognize a feature subset that permits the classifiers to recover their original performance accurateness. Furthermore, even compared with other feature selection techniques, the obtained accuracy is lower

C.Subset selection algorithm

The inappropriate features, along with preventable features, severely change the accuracy of the learning machines. Thus, feature subset collection should be able to distinguish and remove as much of the irrelevant and unneeded information as possible. Moreover, good feature subsets contain features tremendously interrelated with (predictive of) the class, yet unassociated with each other. Keeping these in mind, we build up a novel algorithm which can professionally and efficiently handles both irrelevant and redundant features, and attain a good quality feature subset.

D.Algorithm

Inputs of algorithm:-D $(F1, F2, \ldots, Fm, C)$ -Data set

θ - the T-Relevance threshold.

Output of algorithm:- S - selected feature

Eliminate outstanding irrelevant features

1 for i = 1 to m do

2 T-Relevance = SU (Fi, C)

3 if T-Relevance > θ then

4 S = S $\cup \{Fi\}$;

5 G = NULL; // G is a complete graph

6 for each couple of features $\{Fi, F6\}$ S do

7 F-Correlation = SU $(F, Fj7)$

8 Add Fi and/or $F8j$ too G with F-correlation as the weight of the corresponding edge;

9 minSpanTree = Prim (G);

OR // It decided using type of graph

MinSpanTree = Kruskal (G); //Using Prim's OR Kruskal's Algorithm to generate the minimum spanning tree

10 Forest = minSpanTree

11 for each edge $Eij \in$ Forest do

if SU$(Fi, F) <$ SU(Fi,C) $\wedge$ SU(Fi,Fj) $<$ U (Fj,C) then

13 Forest = Forest $- Eij$

14 $S = \phi$

15 For each tree $Ti \in$ Forest do ///// Tree Partition and Representative Feature Selection

16 $Fj R =$ argmax $\in SU(F, C)$

17 S = S $\cup \{Fj, R\}$;

18 return S

19 Lastly we apply feature Interaction rule and get interactive features.

E.Time complexity

The huge amount of work for Algorithm absorbs the calculation of SU values for TR relevance and F-Correlation, having linear complexity in terms of the number of occurrences in a given data set. The initial portion of the algorithm has a linear time complexity in terms of the amount of features m. considering features is selected as appropriate ones in the first part, when k only one feature is selected

IV. RESULTS AND DISCUSSION

A. Data set

In order to perform and implement our work various types of data sets are used i.e. Microarray, Image, and text data having different features. That is already downloaded from particular websites. The process for generating our experimental data sets is as follows: Firstly WEKA tool browse dataset from particular system location, after browsing that dataset it is converted into suitable system format. After this step, we got the actual input for our algorithm.

We extracted and processed those datasets using our algorithm for finding out relevant feature subset.

B.Pperformance measurement

In the existing feature subset selection process, various algorithms and techniques have been proposed, but our proposed algorithm is somewhat different as compared to existing algorithm. This proposed feature selection algorithm increase efficiency and reduces the time complexity of the algorithm we can compare this algorithm to another algorithm as FCBF, CFS, ReliefF, Consist and FOCUS-SF. While evaluating the performance of our algorithm different factors are considered,

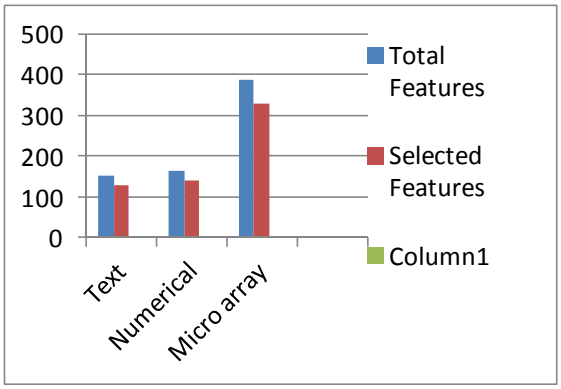

Figure 2: Feature selected in different data sets

1] Ratio of the number of feature selected by the algorithm
2] Time required to finding out the features
3] Accuracy of the selected features and
4] Performance level of algorithm i.e. better, equal or worst.

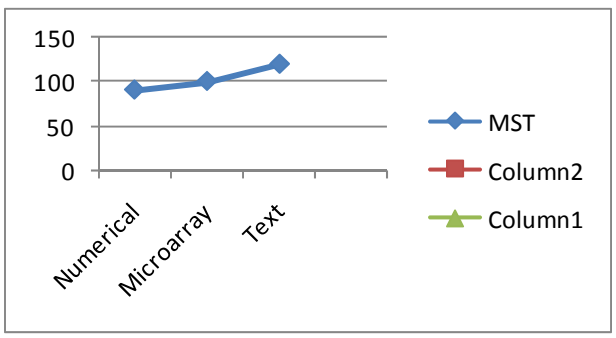

Figure 3: Time Required for MST

V. CONCLUSION AND FUTURE WORK

In this paper we represent an efficient clustering-based algorithm for feature extraction. The proposed algorithm works on a massive quantity of data to mine most relevant set of features. In the clustering based algorithm the features are divided into different cluster so that it can reduce the unnecessary data from data set.

In this paper, we also initiate feature interaction rule that is useful to removing completely unrelated data set but not the incomplete irrelevant. In this case the data items are irrelevant when considered independently but make a relevant feature when united together. Paper introduces one more method in the construction of MST i.e. we use Prim's as well as Kruskal's algorithm based on the type of graph i.e. Dense graph or Sparse graph. If the graph is Dense graph use Prim's algorithm otherwise use use Kruskal's algorithm for constructing the MST the type of graph affects the time complexity of the algorithm.

In the future scope we can enhance our work by extending the symmetric uncertainty for extracting the feature subset selection

REFERENCES

[1] Souza J., Feature selection with a general hybrid algorithm, Ph.D, University of Ottawa, Ottawa, Ontario, Canada, 2004

[2] 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.

[3]L.C.Molina,L.Belanche,Nebot,Featureselectionalgorithms:asurveyandexperimentalvaluation,in:ProceedingsofIEEEInternationalConferenceonDataMining,IEEEComputerSociety,2002,pp.306–313.

[4] Q. Song, J. Ni, G. Wang, A fast clustering-based feature subset selection algorithm for high dimensional data, IEEE Transactions on Knowledge and Data Engineering (99) (2011) 1–14

[5] Z. Zhao, H. Liu, Searching for interacting features in subset selection, Intelligent Data Analysis 13 (2) (2009) 207–228.

[6] J. Xie, J. Wu, Q. Qian, Feature selection algorithm based on association rules mining method, in: 2009 Eighth IEEE/ACIS International Conference on Computer and Information Science, IEEE, 2009, pp. 357–362.

[7] P. Chanda, Y.R. Cho, A. Zhang, M. Ramanathan, Mining of attribute interactions using information theoretic metrics, in: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, IEEE Computer Society, 2009, pp. 350–355.

[8] 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.

[9] Sha C., Qiu X. and Zhou A., Feature Selection Based on a New Dependency Measure, 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 1, pp 266-270, 2008.

[10] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transaction on Knowledge and Data, Engineering, Vol. 25, No. 1, January 2013.