International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 599 - 602

FAST Selection in Clusters for High Dimensional Data

Priyanka Patil [#1],Swati Hojage[#2],Harshada Sonkamble[#3]
patilpriyanka561@gmail.com, swati.hojage93@gmail.com , sonkambleharshada16@gmail.com

***Abstract*—** Process of selecting relevant features from available dataset is known as features selection. Feature selection is use to remove or reduce redundant and irrelevant features. Various feature selection algorithms such as CFS (correlation feature selection), FCBF (Fast Correlation Based Filter) and CMIM (Conditional Mutual Information Maximization) are used to remove redundant and irrelevant features. To determine efficiency and effectiveness is the aim of feature selection algorithm. Time factor is denoted by efficiency and quality factor is denoted by effectiveness of subset of features. Problem of feature selection algorithm is accuracy is not guaranteed, computational complexity is large, ineffective at removing redundant features. To overcome these problems Fast Clustering based feature selection algorithm (FAST) is used. Removal of irrelevant features, construction of MST (Minimum Spanning Tree) from relative one and partition of MST and selecting representative features using kruskal's method are the three steps used by FAST algorithm.

***Keywords*—** *Feature subset selection, graph theoretic clustering, FAST*

*****

I. INTRODUCTION

Highlight subset choice can be seen as the strategy for distinguishing and evacuating a considerable measure of disconnected and superfluous components as likely in light of the fact that (i) inconsequential elements don't give the prescient accuracy (ii) pointless elements don't redound to getting an unrivaled indicator for that they give principle information which is already there in extra elements.

There are various element subset choice calculations, a couple can effectively expel irrelevant components yet not succeed to hold pointless element however a couple of others can evacuate the unimportant while taking worry of the superfluous elements [1],[2],[3],[4]. Proposed FAST calculation course into the resulting bunch. Typically include subset choice study has been caution on scanning for vital components. Concerning the objective ideas, the point of selecting a subset of good components is for decreasing dimensionality and expelling irrelevant information. Subset determination is thinking about as a viable way which can expanding learning exactness, and enhancing result clarity [5], [6]. Highlight determination calculations can be separated into four general classifications: they are Embedded, Wrapper, Filter, and Hybrid methodologies. The implanted strategies incorporate element choice as a part of the preparation process; case of installed methodologies are conventional machine learning calculations like choice trees or simulated neural systems [7].The wrapper strategy utilize the prescient exactness of a foreordained learning calculation to decide the decency of chose subsets. The channel techniques are self-overseeing of learning calculations and point of confinement inquiry space, with great disentanglement. By joining channel and wrapper techniques the half and half strategy happened. Proposed FAST calculation utilizes channel technique and it works in two stages. In the initial step, elements are parcel into groups by utilizing chart theoretic bunching strategies. In the second step, the most illustrative component that is emphatically identified with target classes is chosen from every group to shape a subset of elements. Highlights in various bunches are generally free; the grouping based methodology of FAST has a high likelihood of deciding a subset of valuable and autonomous elements. To recognize the effectiveness of FAST, we embrace the base spreading over tree strategy.

In this paper, presenting incomplete aftereffects of this work. In next segment II, introducing the writing overview over the different techniques exhibited for highlight subset determination. In area III, the proposed methodology and its proposed framework square graph is portrayed. In area IV, introducing the calculation utilized. At last conclusion and future work is anticipated in area V.

II. LITERATURE SURVEY

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. Of the many feature subset selection algorithms, some can effectively remove irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features [8]. Proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has concentrated on searching for relevant features. A well-known example is Relief [7].The key idea of Relief is to estimate the quality of features according to how well their value distinguish between instance that are nearer to each searches the data set from its two nearest neighbors: one from its own class, called nearer Hit H and the other from the different class, called nearer miss M. It updates the quality estimation W $|A_i$ for all the features A_i based on the values of different function diff ( ) about X, H and M. The process is repeated m times where m is a user defined parameter. For instance $X_{1i,}$ X_{2i} , diff($A_{i,}X_{1i}$,X_{2i}) calculate the difference between values x_{1i} and x_{2i} of feature $A_{i..}$

$$\text{diff}(A_{i,}x_{1i},x_{2i}) = \begin{cases} |x_{1i}-x_{2i}| & \text{if } A_i \text{ is numeric} \\ 0 & \text{if } A_i \text{ is nominal and } x_{1i} = x_{2i} \\ 1 & \text{if } A_i \text{ is nominal and } x_{1i} \neq x_{2i} \end{cases}$$

However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. However, along with irrelevant features, redundant features also hamper the speed and accuracy of learning algorithms, and thus should be eliminated as well. The purpose of Relief-F is to select a subset of features from the feature space which is good enough regarding its ability to describe the training dataset

and to predict for future cases. CFS [8], FCBF [10] and CMIM [11] are examples that take into consideration the redundant features.

Equation 1 (Ghiselli 1964) formalises the heuristic

$$Merit_s = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \quad (1)$$

Where $Merit_s$ is the heuristic merit of a feature subset S containing k features
r_{cf} is the mean feature class correlation (f £S) and
r_{ff} is the average feature inter correlation

Equation 1 is in fact Pearson's correlation where all variables have been standardized. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. The heuristic determine irrelevant features as they will be poor predictors of the class Redundant attributes are discriminated against as they will be highly correlated with one or more of the other features. CFS [8] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, but not correlated with each other.

The approximation method for relevance and redundancy analysis presented before can be realized by using FCBF (Fast Correlation-Based Filter) [10]. It involves two connected steps:

1. Determining a subset of relevant features, and
2. Selecting predominant features from relevant ones.

For a data set S with N features and class C, the method finds a set of predominant features S_{best}, calculates the SU value for each feature, selects relevant features into S'_{list} based on a predefined threshold value δ, and arrange them in a descending order according to their SU values. In the second step, it further processes the ordered list S'_{list} to select predominant features. A feature Fj that has already been identified to be a predominant feature can always be used to filter out other features for which Fj forms an approximate Markov blanket. Since the feature with the maximum C-correlation does not have any approximate Markov blanket, it must be one of the predominant features. So the iteration starts from the first element in S'_{list} and continues as follows. For all the remaining features (from the one right next to Fj to the last one in S'_{list}), if Fj happens to form an approximate Markov blanket for Fi, Fi will be taken out from S'_{list}. After first round of filtering features based on Fj, the algorithm will consider the remaining feature right next to Fj as the new reference to repeat the filtering process. The method stops until no more predominant features can be selected. Figure 1 illustrates how predominant features are selected with the rest features removed as redundant ones. In Figure 1, six features are selected as relevant ones and ranked according to their C-correlation values, with F1 being the most relevant one. In the first round, F1 is determined as a predominant feature, and F2 and F4 are removed based on F1. In the second round, F3 is determined, and F6 is removed based on F3. In the last round, F5 is determined.

Where, S (F1,F2,...,FN,C) - a training data set ,δ - a predefined threshold, S_{best} - a selected subset

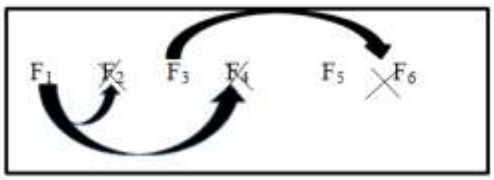

Fig. 1 Selection of Predominant Features

CMIM[7] (Conditional Mutual Information Maximization) iteretively picks features which maximize their mutual informatin with the class to predict,conditionally to the response of any feature already picked. Different from these algorithms, proposed FAST algorithm employs clustering based method to choose features.

A bunching calculation is connected preceding a classifier to lessen highlight dimensionality by gathering together "comparable" components into a little number of highlight groups, i.e. groups can be utilized as components for the ordering different undertaking supplanting the first element space. The primary grouping strategies have been connected in the writing: data bottleneck, distributional bunching, divisive bunching, and progressive bunching. The fundamental point of the IB strategy is to remove the data from one variable that is pertinent for the forecast of another variable. Distributional bunching as an element grouping technique for content arrangement. The comparability between two components, and is measured as the similitude between the class variable. Divisive bunching, and is relevant to the content arrangement. The technique determines a worldwide target work that expressly catches the optimality of highlight bunches. Progressive bunching has been utilized to choose highlights on unearthly information. The similarity between two features is estimated using the absolute value of the correlation:

Quite different from these above clustering based algorithms, proposed FAST algorithm make use of minimum spanning tree based method to cluster features. Moreover, proposed FAST does not limit to some specific types of data.

III. PROPOSED APPROACH AND BLOCK DIAGRAM

3.1 Problem statement

Selection of subset of useful features from large high dimensional data is difficult. Existing features selection techniques have not been so effective in minimizing the search space and selecting useful features that produces compatible results. Various features selection techniques fail to remove both irrelevant and unrelated data therefore decreasing result comprehensibility. Hence, there exists no solution to limit search space and to remove both irrelevant and unrelated data from high dimensional data.

In proposed system, the fast clustering-based feature selection algorithm (FAST) 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. 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. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. The proposed FAST algorithm sensibly consists of three steps:

(i) removing immaterial features,
(ii) MST is constructed from relative ones, and
(iii) MST is portioned and then selecting representative features.

3.2 Block Diagram

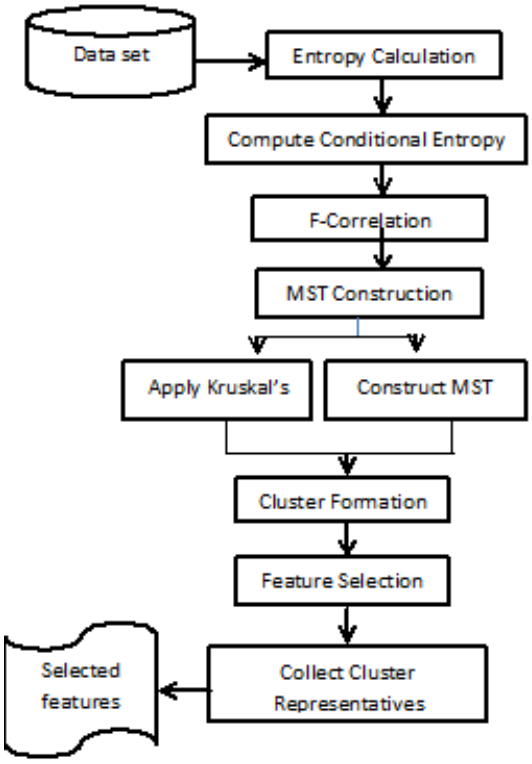

Fig 2: Framework of the proposed feature subset selection algorithm

3.3 System modules:

1. Removal of Irrelevant features
2. T-Relevance, F-correlation calculation
3. MST construction
4. Relevant feature calculation

3.3.1 Modules Description:

Step 1: Removal of Irrelevant features

Removal of irrelevant features helps to improve result, remove irrelevant and unnecessary data, reduce dimensionality and increase learning accuracy. Machine learning applications use many feature selection methods. If we consider a Dataset 'D' with m features F= {F1, F2, …. Fn } and class C, features are available with target relevant feature. The generality of the selected features is limited and computational complexity is large. Filter and wrapper methods are combine to form hybrid method, therefore reduces search space that will be considered by the subsequent wrapper. Method to reduce search space that will be considered by the subsequent wrapper.

Step 2: T-Relevance, F-correlation calculation

T-Relevance is calculated by using feature and the target concept C, F-Correlation is calculated by using a pair of features, feature redundancy i.e F-Redundancy and representative feature i.e

R-Feature of a feature cluster can be defined. According to definitions, which are specified above feature subset selection process identifies strong T-Relevance and selects R-Features from features clusters. The reason behind this is as follows

1. Target concept provides weak correlation to irrelevant features.
2. Representative feature can be identified from group of cluster, which is formed by redundant features..

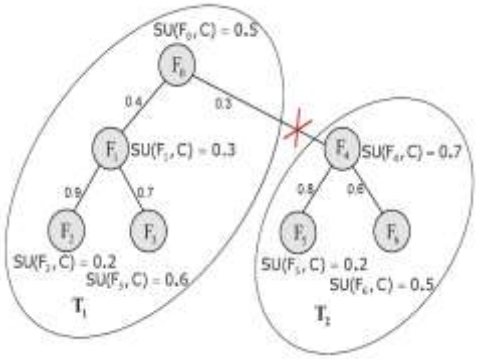

Fig 3: Example of Clustering

Step 3: MST construction:

Minimum-spanning tree (MST) clustering method is used to ensure the efficiency and effectiveness of FAST. Comparison between FAST and several representatives feature selection algorithm is carried out by using various experiments. Minimum spanning tree is constructed with weights, which connects all vertices such that the sum of weights of edges is minimum, using well-known Kruskal's algorithm.

Step 4: Relevant feature calculation:

After partition of tree and removal of unnecessary and irrelevant results in two disconnected trees. Removal of unnecessary edges, a forest is obtained. Each tree acts as a cluster. Finally, feature subset is generated, and then calculates the accurate or relevant feature.

IV.ALGORITHM

FAST Algorithm :

inputs: D($F1$, $F2$, ..., Fm, C) - the given data set

θ - the T-Relevance threshold. output:

Output :S - selected feature subset .

//==== Part 1 : Irrelevant Feature Removal ====

1. for i = 1 to m do
2. T-Relevance = SU (Fi, C)
3. if T-Relevance >θ
4. then S = S ∪{Fi}

//==Part 2:Minimum Spanning Tree Construction =

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

6. for each pair of features {$F'i$,$F'j$}⊂S do

7. F-Correlation = SU ($F'i$, $F'j$)

8. $Add\ F'i\ and/or\ F'j\ to\ G\ with$ F-Correlation $as\ the\ weight\ of the\ corresponding\ \ edge$;

9. minSpanTree = Krushal's (G); //Using Krushal's Algorithm to generate the minimum spanning tree

//== Part 3: Tree Partition and Representative Feature Selection ===

10. Forest = minSpanTree
11. for each edge Eij ∈ Forest
12.do if SU($F'i$,$F'j$) < SU($F'i$,C)∧SU($F'i$,$F'j$) < SU($F'j$,C) then
13. Forest = Forest – Eij
14. S = ϕ
15. for each tree Ti ∈ Forest
16. do $Fj\ R$ = argmax$F'k$∈Ti SU($F'k$,C)
17. S = S ∪{$Fj\ R$};
18. return S

V.CONCLUSION AND FUTURE WORK

The existing feature selection algorithms cannot eliminate both unrelated and unwanted features. Hence, it becomes difficult to select relevant features from entire set of features and thereby reduces efficiency and effectiveness of relevant features. So, in order to overcome the difficulties of irrelevant and unwanted feature removal, a new algorithm called "FAST Algorithm" has been implemented. It works in two steps. In first step, graph-theoretic clustering methods are used to partition features into clusters. In second step, subset of features is formed by combining representative features from target class. This has improved the performance of search algorithm on large scale.

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] Mitchell T.M, Generalization as search, Artificial Intelligences 18(2),pp203-226,1982.

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

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

[5] Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.

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

[7] Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.

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

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

[10] 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.

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