International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Humming Bird (01st March 2014)

RESEARCH ARTICLE OPEN ACCESS

FAST: A New Novel Feature Selection Algorithm for Large Datasets

RethinRaveendran ,Bhuma V.R

1RethinRaveendran. Author is currently pursuing M.Tech(Information Technology) in Vinschristian college of Engineering. e-mail:rethinraveendran@gmail.com,

Assistant Professor, Department of Information Technology, Vins Christian College of Engineering

Abstract:
In the modern world, So many kinds of datasets are used in all industries. Dataset is a huge collection of different kinds of data in a database. Mining a meaningful feature from the dataset is very difficult, because the dataset contain large amount of data, so it is very difficult to access and a dataset contain relevant as well as irrelevant features so we cannot get the feature. To avoid these problems, implementing a new novel feature selection algorithm, namely FAST. Mainly FAST algorithm is works in some steps; in the first step selecting a dataset and converting the dataset into a table format. In the second step, calculate the T-Relevance value and calculate the threshold, by using the threshold value remove the irrelevant features. In the third step; after removing the irrelevant features form a minimum spanning tree. In the fourth step, By using the minimum spanning tree features are divided into clusters

***Keywords*—**feature selection, irrelevant feature removal, Relevance, Correlation

I. INTRODUCTION

My main aim of this project is to select the relevant features from the large data sets. Now a days feature selection is one of the major problem in the learning algorithm, decision tree making, etc. for the efficient selection of relevant or redundant feature is done by the FAST algorithm [1]. So many feature selection algorithms present, likely FCBF (Fast Correlation Based Feature Selection) [2], FCBF# (an extension to FCBF algorithm) [3], Relief [4], Consistency based search for feature selection [5] etc.

In all feature selection algorithms, feature selection method is mainly divided in to two types. The first type is filter method and second type is wrapper method [6]. Filter method is widely used in all algorithms. Fiter method is defined as "The filter model relies on general characteristics of the training data to select some features without involving any learning algorithm." Wrapper method is defined as"predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected".

In this paper implementing a new novel clustering based feature selection algorithm for large datasets. This paper mainly focused on the datasets and efficient feature selection from the larger dataset. A dataset contain large amount of data that is the main problem dealing with the data set. Dataset can take more time to load in a system. Second problem with dataset is, it contain relevant feature as well as the irrelevant feature so we cannot get the correct output from the dataset.

FAST algorithm is used to avoid the problems with the dataset and high dimensional data. The main goal of this project is to remove the irrelevant features and form a minimum spanning tree with most relevant or redundant feature from the dataset. Features that are strongly related to the target class is selected and get the output as clusters. Clusters contain the relevant features only and then we can classify the data on the basis of their priority.

In the construction of the minimum spanning tree PRIM's algorithm is used. PRIM's algorithm is more efficient and easy to understand. By using this algorithm we get the spanning tree with smaller weight and the graph must contain the shortest path between the nodes. Shortest path is more easy to find out the correlation between the adjacent nodes.

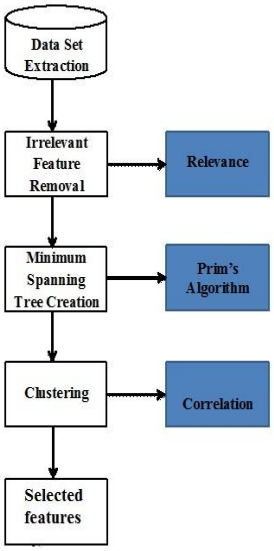
II. RELATED WORK

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because: (i) irrelevant features do not contribute to the predictive accuracy [7], and (ii) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

Feature selection algorithm is mostly used in the field of medical, education, defense, decision making, and learning purposes. In all kind of industries are using feature selection algorithms for their decision making purpose

III. FAST CLUSTERING BASED FEATURE SELECTION ALGORITHM

3.1 Architecture



The architecture shows the step by step process of FAST algorithm. Mainly FAST algorithm is works on four steps 1. Dataset Extraction 2.Irrelevant feature removal 3.Mininmum spanning tree creation 4.Clustering.

3.2 Algorithm and Analysis

FAST algorithm is works on the given steps,

1. Dataset Extraction

A dataset (or data set) is a collection of data. Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where each column of the table represents a particular variable, and each row corresponds to a given member of the dataset in question. The dataset lists values for each of the variables, such as height and weight of an object, for each member of the dataset. Each value is known as a datum. The dataset may comprise data for one or more members, corresponding to the number of rows. A dataset contain large amount of data. The data is in the form of image, text, array, string and characters. In our system we cannot handle the datasets directly, so datasets are converted into some tables assigned in the system by the help of SQL database queries. The table contains attributes and all values presented on the dataset. Conversion from dataset to the table is very fast and easily understandable.

2. Irrelevant feature removal

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Irrelevant feature removal module is mainly composed of two connected components;Irrelevant feature removal and Redundant feature elimination Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification by a number of researchers.

The symmetric uncertainty is defined as follows,

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \text{ ..... (1)}$$

Where,

1. (X) is the entropy of a discrete random variable X. Suppose $p(x)$ is the prior probabilities for all values of X, $H(X)$ is defined by,

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \text{ ..... (2)}$$

2. IG (X|Y) is the amount by which the entropy of Y decreases. It reflects the additional information about Y provided by X and is called the information gain which is given by,

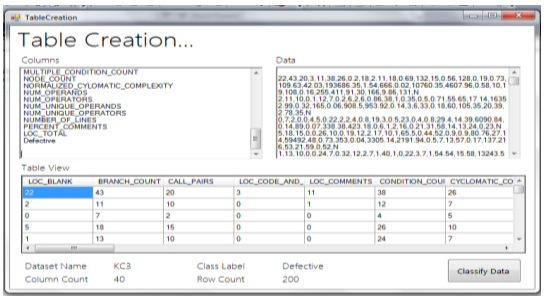
$$IG(X|Y) = H(X) - H(X|Y) \text{ ..... (3)}$$

3. Where (X|) is the conditional entropy which quantifies the remaining entropy (i.e. uncertainty) of a random variable X given that the value of another random variable Y is known. Suppose $p(x)$ is the prior probabilities for all values of X and $p(x|y)$ is the posterior probabilities of X given the values of Y ,$H(X|Y)$ is defined by,

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \text{ ... (4)}$$

Calculate the mean value of the symmetric uncertainty and set the mean value as the threshold value. If the symmetric uncertainty or T-Relevance

value is less than the threshold value, the data will be an irrelevant feature. Irrelevant features are removed in this step.



3. Minimum spanning tree creation

Minimum spanning tree is created based on the T-Relevance value gets in the second module. So many nodes are deleted in the form of irrelevant feature, so we can easily create a minimum spanning tree. This module is works in two steps,Minimum spanning tree is created by using PRIM's algorithm and F-Correlation is calculated for the adjacent nodes in the minimum spanning tree. Prim's algorithm is a greedy algorithm that finds a minimum spanning tree for a connected weighted undirected graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized.

Correlation or dependence refers to any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence.

4. Clustering

In the clustering module, features are divided into clusters based on the correlation value. Node will be eliminated when the correlation between 2 adjacent nodes on the minimum spanning tree is less than that of the T-Relevance value of the node. After completing this module we get the clusters of the selected features.

inputs: D($F1$, $F2$, ..., Fm, C) - the given data set
θ- the T-Relevance threshold.
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 >θthen
S = S ∪ {$Fi4$ };
Part 2 : Minimum Spanning Tree Construction

5 G = NULL; //G is a complete graph
6 for each pair of features {Fi, }⊂ S do
7 F-Correlation = SU (,Fj)
8 $AddF'iand/orF'jtoGwith$F-Correlation
$astheweightofthecorrespondingedge$;
9 minSpanTree = Prim (G); //Using Prim Algorithm to
generate theminimum spanning tree

Part 3 : Tree Partition and Representative Feature Selection
10 Forest = minSpanTree
11 for each edge Eij11 ∈Forest do
12 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 Ti15 ∈Forest do
16 FjR= argmax$F'k$∈TiSU($F'k$,C)
17 S = S ∪ {FjR};
18 returnS

FAST Algorithm

IV. RESULTS AND ANALYSIS

A. Dataset selection

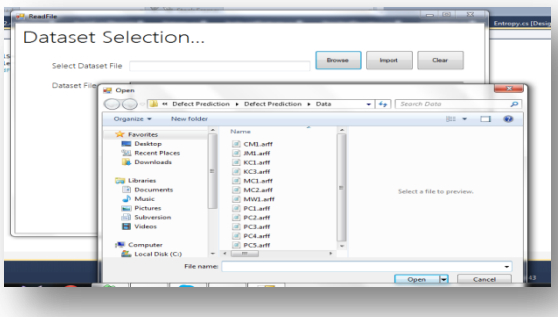
Selecting the data set from a large database. Dataset is collection of different data like text, image, microarray data etc. Dataset is the input of this module. In this project mainly text dataset is used as an input

B. Dataset Processing:

In the Data processing step, select the dataset file and differentiate the attributes and variables from the dataset for creating a new table for the data present in the dataset.

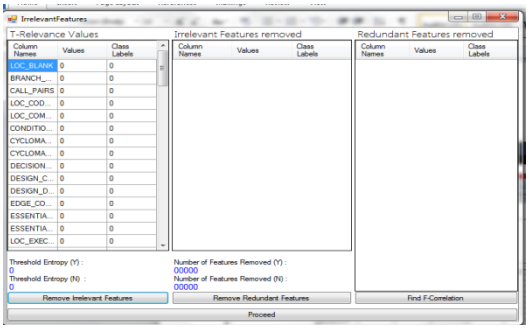
C. Table creation:

In the table creation stage, the selected datasets are converted into table based on the attributes present. Tables are created by the SQL database. By using tables i can easily calculate the correlation and redundancy.



D. Irrelevant feature removal:

Irrelevant feature removal is based on the T-Relevance value and Correlation value. A threshold will be calculated from the T-Relevance, if the T-Relevance value is less than that of the threshold then that node is an irrelevant node and it can be terminated.



V. CONCLUSION

Main purpose of implementing a novel clustering based feature selection for high dimensional data is to select most relevant features or the features that are strongly related to the target class is selected. The algorithm involves

(i) Removing irrelevant features.
(ii) Constructing a minimum spanning tree from relative ones.
(iii) Partitioning the MST and selecting representative features.

In my third semester I completed 2 modules of the project. These modules have dataset creation and irrelevant feature removal. Irrelevant feature removal is one of the important processes in the feature selection algorithm. Output of the second module is redundant features.

In the future work I have to complete 2 modules namely, minimum spanning tree creation and clustering. After completing these modules only we will get the selected features from a large dataset. PRIM's algorithm is used to creating the minimum spanning tree from the minimum spanning tree, calculate the correlation between the adjacent nodes of a tree and compare the correlation value and the T-Relevance value. If the correlation is less than the T-Relevance then that node will be deleted and form a cluster including the other sets. These clusters gives the selected features.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1 YEAR 2013.
[2] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
[3] BarisSenliol,GokhanGulgezen, Lei Yu, ZehraCataltepe, "Fast Correlation Based Filter (FCBF) with a Different Search Strategy ".
[4] AntonioArauzo-Azofra, Jose manuel Benitez, and Juan Luiz Castro "A feature set measure based on Relief".
[5] Manoranjan Dash, Huan Liu, Hiroshi Motoda, "Consistency Based Feature Selection".
[6] Sanmay Das, "Filters, Wrappers and a Boosting-Based Hybrid for FeatureSelection"
[7] 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.

RethinRaveendran received the B.Tech in Information Technology from Bethlahem Institute of Engineering in2012 and M.Tech. Degree in Information Technology from VINS Christian College of engineering and Technology2014, respectively.His Area of Interest includes feature selection in data mining.