International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

A Swift Clustering based Algorithm to Explore Different Correlation Measures

G. Julie Priyadharsana[1]
[1]*Computer Science and Engineering, Anna University*

Abstract: ***Clustering analysis technique is the statistical data analysis involves for identifying relevant feature. Feature selection identify a subset of features which to produces the result same as the target features. The novel Swift Clustering based algorithm removes both irrelevant and redundant features. The feature subset selection algorithm is to takes minimum time to find most useful features but provide quality one. Swift Clustering Based Feature Subset Selection(CBFS) algorithm has two Graph-theoretic methods are Minimum Spanning Tree (MST) and Tree Partitioning contained features are divided into clusters using the Spanning Tree Construction process and then the Cluster representatives are selected to form the effective feature set. The Minimum Spanning Tree (MST) eliminates redundancy using kruskal's algorithm. Instead of T-Relevance, the proposed works finds the supervised similarity between the attributes also takes into account the unsupervised similarity between two attributes. Finally in Tree Partitioning the features are divided into clusters according to the representatives used to get minimized amount of related features. SWIFT Clustering algorithm works more efficient compared with FAST Clustering based feature selection algorithms produced most representative feature as a result have best agreement with human performance.***

Keywords: ***Swift CBFS, standard deviation, T-Relevance, F- Correlation, MST, Tree Partition, kruskal's algorithm, clustering***

I. INTRODUCTION

The attributes in most real world data are redundant and simply irrelevant to the purposes of discovering interesting patterns. Feature selection selects relevant features in the dataset prior to performing data mining important for accuracy of further analysis as well as for performance. Because the redundant and irrelevant attributes could mislead the analysis, including all of the attributes in the data mining procedures not only increases the complexity of the analysis, but also degrades the accuracy of the result [8].

For instance, clustering techniques [10], which partition entities into groups with a maximum level of homogeneity within a cluster, may produce inaccurate results. In particular, because the clusters might not be strong when the population is spread over the irrelevant dimensions, the clustering techniques may produce results with data in a higher dimensional space including irrelevant attributes. Feature selection process improves the performance of data mining techniques by reducing dimensions so that data mining procedures process data with a reduced number of attributes. Feature selection and reduction aim at choosing a small sub-set of attributes that is sufficient to describe the data set identify and removes as much as possible irrelevant and redundant information [8].

The intrinsic dimensionality of data is the minimum number of parameters needed to account the observed properties of the data. For many learning algorithms, machine learning the training and/or the classification time increases directly with the number of features. Sophisticated attribute selection methods have been developed to tackle three problems: reduce classifier cost and complexity, improves model accuracy (attribute selection), and also improves the visualization and comprehensibility of induced concepts. Feature selection serves two main purposes. First purpose is makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Second purpose of feature selection is increases classification accuracy by eliminating noise features [9].

The Clustering analysis technique involves for the statistical data analysis. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other to those in other groups. There are two categories in clustering analysis they are distance and conceptual based clustering [10]. Using the Minimum Spanning Tree and the Tree Partitioning performing both Distance and Conceptual based Clustering. Based on the MST method propose swift clustering based feature Selection algorithm (CBFS). MST is constructed for the features received then using hierarchal clustering technique with similarity measure. The representative features are taken from each cluster so that the getting target made easy, thereby reducing the feature set and reducing the computational time to build the predictive model [4].

II. RELATED WORK

Some feature subset selection method used in machine learning applications could be trained from the data they are Wrapper, Filter and Hybrid approach. The Wrapper method predicts accuracy from results that already predicted or predetermined [1], [11], [15]. Filter feature selection method is for finding data from large sets. If newly added more data's in dataset means filter method is a good choice for finding relevant data [1], [11]. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper with the similar time complexity of filter methods [1]. Then some of the Traditional feature selection algorithms are used they are FAST [2], [5], Distributed clustering [10]. Distributed clustering cluster the words based on grammatical relations with other words learning.

The FAST clustering-based feature subset selection algorithm [5] has Minimum Spanning Tree using PRIMS Algorithm. Prim's Algorithm deals with node. For Prims's algorithm is necessary to have an edge of minimum weight to be an adjacent node.

The SWIFT Clustering-based feature subset selection algorithm has two Graph-theoretic methods consist of Minimum Spanning Tree [4] and Tree Partitioning [6]. SWIFT Clustering algorithm has MST using kruskal's Algorithm deal with edges considering minimum edges to span [3]. In the MST with kruskal's algorithm there is no necessary to move with adjacent vertex. Various correlation measures takes place because of newly added data's to eliminate redundant feature [8].

A. SWIFT CBFS subset selection algorithm and analysis

A Swift CBFS algorithm effectively and efficiently deals with both relevant and redundant features. It can be achieve through a feature selection framework. Framework includes irrelevant and redundant elimination it described in steps

1) *Step 1:* From dataset the user searches for some concept as they required. It shows some relevant data is a part of distributed clustering.
2) *Step 2:* The constructing Minimum Spanning Tree from a weighted complete graph. Condition for MST is need to visit each point for one time, no closed loop occur, the tree need to be connected. If cycle forms there having redundancy. Distance between two points is a weight of an edge.
3) *Step 3:* Tree Partition provide partitioning of MST after all nodes get connected once.
4) *Step 4:* Step 2 and Step 3 for removing the redundant feature. Finally, collect the most representative feature from the partition it is a cluster formation to get target feature.

Instead of the T-Relevance concept the Swift CBFS works finds the Supervised Similarity between the attributes. This new quantitative measure efficiently removes the redundancy among the attributes. It also takes into account the unsupervised similarity between two attributes.

The MST construction, Kruskal's Algorithm [3] is replaced instead of Prim's Algorithm. Relevant features have strong correlation with target concept so always necessary for a best subset of feature, while redundant features are not because their values are completely correlated with each other.

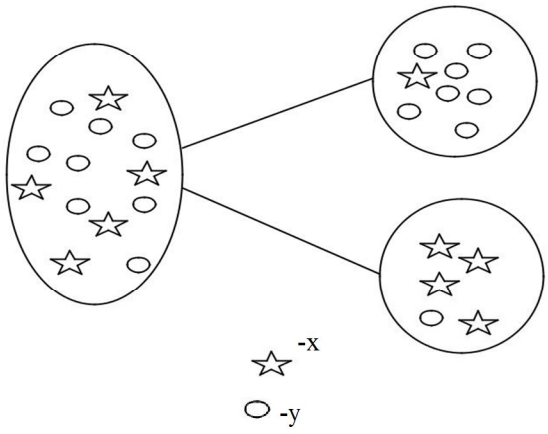

Fig. 1 Separation of Attributes Based on features class label

Thus notions of the feature redundancy and the feature relevance are normally in terms of the feature correlation and the feature-target concept correlation. To find the relevance of each attribute with the class label, Information gain is computed. This is also said to be Mutual Information measure [12]. Mutual information measures how much the distribution of feature values and target

classes differ from statistical independence. This is the needed nonlinear estimation of correlation between feature values and target classes. The Symmetric Value (SV) is derived from 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.

To calculate the Standard deviation formula is shown below used find the neighbourhood distance among attributes contain various feature sets

$$\sqrt{1/N \sum_{i=1}^{N} (x_i - \mu)^2}$$

Fig 1. Shows the separation of attributes based on class label [7]. After making the separation to calculate Information Gain formula is defined as follows

$$(X|Y) = H(Z) - H(X|Y)$$
$$= H(Z) - H(Y|X)$$

To calculate gain, we need to find the entropy and conditional entropy values. The equations for that are given below

$$-\sum_{x \in X} p(x) \log_2 p(x) \quad (3)$$
$$-\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \quad (4)$$

Where p(x) is the probability density function and p(x|y) is th conditional probability density function[13].

B. T-Relevance Computation

The relevance between the feature RFi € F and the target concept TC is referred to as the T-Relevance of RFi and TC, and denoted by SV(RFi,TC). If SV(RFi,TC) is greater than a threshold , then it is said that RFi is a strong T-Relevance feature. After finding the relevance value, the redundant attributes will be removed with respect to the threshold value.

$$SU(X, Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}$$

C. F-Correlation Computation

The correlation between any pair of features PFi and PFj (PFi,PFj € F ^ i ≠ j) is called the F-Correlation of PFi and PFj, denoted by SV(PFi, PFj). The equation Symmetric Value which is used for finding the relevance between the attribute and the class is again applied to find the similarity between two attributes with respect to each label.

To calculate F- Correlation, the Pearson Correlation Coefficient technique is used formula shown below,

$$\sum ((X - M_y)(Y - M_x)) / \sqrt{\left((SS_x)(SS_y)\right)}$$

In Fig. 2 Datasets get loaded to find T-Relevance value it removes the irrelevant feature. The MST and Tree Partitioning are used to remove redundant feature. Then finally from clustering get a target concept.

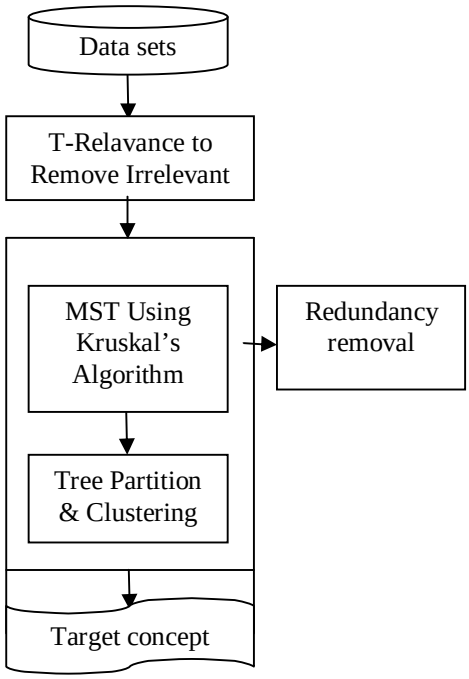

Fig. 2 Shows the simplified architecture of the Swift CBFS

D. MST Construction

By using the F-Correlation value computed above, the Minimum Spanning Tree is constructed. Kruskal's algorithm is used for constructing Minimum Spanning Tree effectively. Kruskal's algorithm is a greedy algorithm in graph theory that finds a Minimum Spanning Tree for a connected weighted graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest.

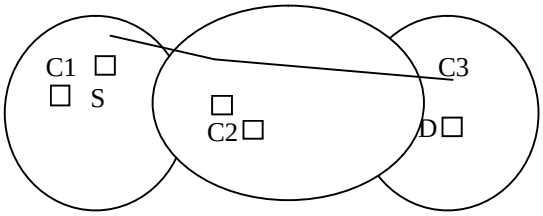

Fig. 3 Example of MST Construction and Cluster Formation

In Fig. 3 Three clusters are formed among the features. After forming the Minimum Spanning Tree from the source (S) to destination (D) the unnecessary edges get removed using kruskal's algorithm [3] by visiting minimum weighted edges. Then the cluster C1, C2, C3 gets formed with the minimum amount of required feature.

E. Cluster Formation

After building the MST, the edges whose weights are smaller than both of the T-Relevance SV(RFi`, TC) and SV(RFj`, TC) are removed from the MST. After removing all the unnecessary edges, a cluster is obtained. Each tree Tj € cluster represents a grouping of related feature that is denoted as VS(Tj), which is the vertex set of Tj as well. The features in each cluster are redundant, so for each cluster VS(Tj) we choose a representative feature RFj RP whose T-Relevance SV(RFj RP,TC) is the greatest.

F. Results and Analysis

The dataset is given as the input it should be in .csv format. During the process attribute values are separated using ',' symbol. Once probabilities are found, reduction would be done based on the threshold value. Then selected features are outputted in .arff format.

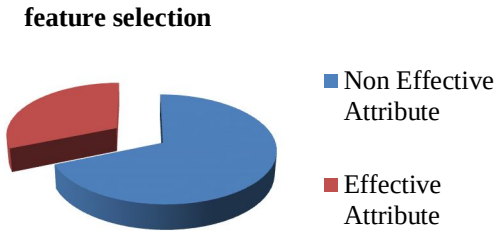

Fig. 4 The Pie Chat shows effective and non effective attributes rate during Feature Selection

The techniques like T- Relevance and F-Correlation values used to form a tree using Minimum Spanning Tree by that unnecessary edge gets removed. Edges removal made because of MST not forms looping and visit nodes where are all having minimum edges first.

In Fig. 4 Pie Chart shows the selection of attributes that related to the target concept as a reduced amount from the whole datasets [14], [18]. The dataset contain data may be in a high dimension it is not a problem for a Swift Clustering algorithm. It can easily separate the effective attribute within a minimum amount of time by exploring correlation of the feature from the high dimensional data.

III. CONCLUSIONS

The A novel Feature selection algorithm using Swift Clustering Based Feature Selection is consists of two steps. In first step, the high dimensional data taken, and then the entropy and conditional entropy measures are applied on its features to find the T-Relevance correlation in order to remove the irrelevant features. The selected relevant features by first step are given into the second step. In second step, the Minimum Spanning Tree is constructed for the features using hierarchal clustering technique with similarity measure and various clusters are formed upon this features based on their similarity. The representative features are taken from each cluster so that the redundant features are eliminated, thereby reducing the feature set to improve the classification accuracy and reducing the computational time to build predictive model. The performance of the Swift CBFS was evaluated and compared with the performance of the FAST feature selection methods. It is observed that, the performance of the Swift CBFS is much better than other feature selection methods. The application of CBFS algorithm can be extended to the pattern recognition also. Experiment can conduct with other well known classifiers.

REFERENCES

[1] Amin Jashki, Majid Makki, Ebrahim Bagheri, and Ali A. Ghorbani. (2009). An Iterative Hybrid Filter-Wrapper Approach to Feature Selection for Document Clustering. Canadian AI '09 Proceedings of 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence. Doi:10.1007/978-3-642-01818-3_10. 74-85.

[2] Lei Yu and Huan Liu. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Department of Computer Science & Engineering Arizona State University. Tempe, AZ 85287-5406. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997

[3] Mario Ruthmair and Gunther R. Raidl. (2009). A Kruskal-Based Heuristic for the Rooted Delay-Constrained Minimum Spanning Tree Problem. 12th Intenational conference on Computer Aided Systems Theory. doi:10.1007/978-3-642-04772-5_92. 713-720.

[4] Prim,R.C. (2004). Shortest Connection Networks and Some Generalizations. The Bell System Technical Journal. doi:bstj36-6-1389. 1389-1401.

[5] Qinbao Song, Jingjie Ni, and Guangtao Wang. (2013). A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data. IEEE Transactions on Knowledge and Data Engineering. doi: 0.1109/TKDE.2011.181. 1041-4347

[6] Shakil Ahmed, Frans Coenen, and Paul Leng. (2006). Tree-based Partitioning of Data for Association Rule Mining. Journal of Knowledge and Information Systems. doi:10.1007/s10115-006-0010-1. 315 – 331

[7] Sunita Beniwal, Jitender Arora. (2012). Classification and Feature Selection Techniques in Data Mining. International Journal of Engineing Research & Technology. ISSN: 2278-018.

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

[9] Guyon I. and Elisseff. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research. 3, 1157-1182

[10] Butterworth R., Shapiro G. and Simovici D.A. (2005). On Feature Selection through Clustering. In Proceedings of the Fifth IEEE international conference on Data Mining. doi:10.1109/ICDM.2005.106. 581-584

[11] Hall M.A and Smith L.A. (1999). Feature Selection for Machine Learning: Comparing a correlation- Based Filter Approach to the Wrapper. In Proceeding of the Twelfth international Florida Artificial intelligence Research Society conference. ISBN:1-57735-080-4. 235-239
[12] Battiti R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks. doi:10.1109/72.298224. 5(4), 537-550
[13] Fleuret F. (2004). Fast binary feature selection with conditional mutual Information. Journal of Machine Learning Research. doi:10.1.1.60.8398 5. 1531-1555.
[14] Andrew Y.Ng. (1998). On feature selection learning with exponentially many irrelevant features as training examples. In Proceedings of the Fifteenth International Conference on Machine Learning. ISBN:1-55860-556-8 404-412. 404-412.
[15] Dijk G. and van Hulle M.M. (2006). Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis. International Conference on Artificial Neural Networks. doi:10.1007/11840817_4. 31-40.
[16] Jiawei Han and Micheline Kamber. (2006). Data Mining: Concepts and Techniques. (2nd ed.). University of Illinois at Urbana-Champaign Morgan Kaufmann Publishers
[17] Puntambekar A.A. (2007). Data Structures & Algorithms. (2nd ed.). Technical Publication Pune, Second Revised Edition, 2007.
[18] Ryan Eshleman and Rahul Singh. (2017). Reconstructing the Temporal Progression of Biological Data using Cluster Spanning Trees. IEEE Transactions on NanoBioscience. doi:10.1109/TNB.2017.2667402.