International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.3, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

EFFICIENT FEATURE SUBSET SELECTION FOR HIGH DIMENSIONAL DATA

G.SORNALATHA[1], M.RETHINAKUMARI[2], S.KASTHOORI[3]

[1,3]PG Scholar, Department of CSE, Renganayagi Varadharaj College of Engineering, sivakasi, India
[2]Assistant Professor, Department of CSE, Renganayagi Varadharaj College of Engineering, sivakasi, India
[3]PG Scholar, Department of CSE, Renganayagi Varadharaj College of Engineering, sivakasi, India

Article Received:07/05/2015 | Article Revised on:15/05/2015 | Article Accepted on:19/05/2015

ABSTRACT

Feature selection is also known as attribute selection (or) variable selection. Feature selection technique is used to predict the relevant feature from the large collection of data in the database. From the database a particular dataset was selected and the dataset has more number of attributes. By using K-medoids algorithm data are selected randomly and a center data head was selected. Based on the head data selection Euclidean distance are computed by measuring the distance between the data and cluster the data. After clustering irrelevant and redundant data are removed using T-Relevance and F-correlation. T-Relevance are computed based on the relevance between feature and target .F-Correlation are estimated based on the correlation between two features .After removing irrelevant and redundant features minimum spanning tree(MST) is constructed and ant colony optimization algorithm is used to remove the unwanted edges from the tree. In turn it improves the attribute feature selection accuracy .Ant colony optimization algorithm is used to avoid the computational problem and it can be used to find optimal path in a graphs when MST problem occur. Ant Colony Optimization algorithm provides better result when compared to other algorithms such as Correlation based Feature Selection (CFS), Fast Correlation based Feature Selection (FCBS), RIPPER and fast clustering-based feature selection algorithm (FAST).

Key Words—*Cluster, FAST, High dimensional, MST, Correlation. Feature Selection*

I. INTRODUCTION

Data mining refers to "using a variety of techniques to identify nuggets of information or executive knowledge in bodies of data, and extract these in such a way that they can be put to use in the areas such as choice support, prediction, forecasting and evaluation. The data is frequently large, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is valuable". Data mining manage have to group a model from the proof, and in the case of supervised learning this requires the user to define one or more classes. The database contains one or more attributes that denote the class of a tuple and

these are known as predicted attributes whereas the remaining attributes are called predicting attributes. A grouping of values for the predicted attributes defines a class. When learning arrangement rules the system has to find the rules that predict the class from the predicting attributes so firstly the user has to define conditions for each class, the data mine system then constructs images for the classes. Basically the method should given a case or tuple with certain known attribute values be able to predict what class this case belongs to, once classes are defined the system should infer rules that govern the Classification therefore the system should be able to find the description of each class.

My main aim of this project is to select the relevant features from the large data sets. Now a day's feature selection is one of the major problems 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 (Fast Correlation Based Feature Selection) [3], Relief [4], Consistency based search for feature selection [5] etc. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be separated into four large categories: the Embedded, Wrapper, Filter, and Hybrid methods. The embedded methods include feature selection as a part of the training process and are usually specific to given learning algorithms, and as a result may be more efficient than the other three categories.

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

The survey for A Novel Relief Feature Selection Algorithm Based on Mean-Variance Model is done by Yuxuan SUN, Xiaojun LOU, Bisai BAO they proposed RELIEF Algorithm. RELIEF usually performs better than the other filter based approaches due to the feedback of the nearest-neighbor classifier. RELIEF algorithm is an effective, simple, and widely used approach to feature weight estimation. The weight for a feature of a measurement vector is defined in terms of feature relevance. These two probabilities are of the value of a feature being different conditioned on the given nearest miss and nearest hit, respectively. In addition, RELIEF is often more efficient than the wrapper approach because RELIEF determines the feature weights through solving a convex optimization problem. In addition, RELIEF is often more efficient than the wrapper approach because RELIEF determines the feature weights through solving a convex optimization problem. A novel feature selection algorithm based on a Mean-Variance model is proposed in this feature weight estimation method in RELIEF to perform a more accurate.

The Survey for Independent Feature Elimination in High Dimensional Data: Empirical Study by applying Learning Vector Quantization method Authors and Affiliations done by Dr. Babu Reddy M they proposed Learning Vector Quantization method.LVQ method in supervised classification environment has been studied with the original data set and with a reduced dataset in which few irrelevant and redundant attributes have been eliminated.

b) On Lung Cancer micro array data set, features with very low coefficient of dispersion were discarded from the further processing and results are tabulated and analyzed.LVQ has great

importance in Feature Selection and Classification tasks. The LVQ method has been applied on the benchmark dataset of Lung cancer Patients. And an attempt has been made to identify some of the insignificant/redundant attributes, by means of Class correlation(C-Correlation), inter-Feature correlation (F-Correlation) and Coefficient of Dispersion among all the instances of a specified attribute. This has lead to the betterment of classification efficiency in a supervised learning environment. A comparison has been made on the efficiency of classification by considering the original dataset and the corresponding preprocessed dataset with reduced set of attributes. Better performance has been noticed on the preprocessed dataset.

The survey for High Dimensional Unsupervised Clustering based Feature Selection Algorithm done by Ms.Barkha Malay Joshi they proposed FSFC Algorithm. Feature selection through feature clustering algorithm is reduced the more attributes than the standard feature selection algorithm like relief and fisher filter. Feature selection is the pre-processing step which reduces the feature subset and thus supports high dimensional dataset. Many feature selection algorithm along with the Feature selection through feature clustering algorithm. The available algorithms like relief, fisher filter and proposed algorithm Feature selection through feature clustering are applied on cancer data and proved that Feature selection through feature clustering reduce more number of attributes compared to relief filter and fisher filter .The relief and fisher filter generate the less number of attributes but this algorithms does not remove the redundant data. Fast Clustering based feature selection algorithm remove the redundancy from the attributes and also provide the reduced or required attributes from the original attribute set Clustering is the technique that put the object of high similarity into one cluster and objects having less similarity into different cluster. So that the same cluster data objects are most similar than the different clusters .Many feature selection algorithm along with the Feature selection through feature clustering algorithm.

The survey for Feature Subset Selection Problem using Wrapper Approach in Supervised Learning done by Asha Gowda Karegowda. Feature selection is a process that selects a subset of original features. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. The goal of feature selection for unsupervised learning is to find the smallest feature subset that best uncovers clusters form data according to the preferred criterion .Feature selection in unsupervised learning is much harder problem, due to the absence of class labels. Feature election for clustering is the task of selecting important features for the underlying clusters Feature selection for unsupervised learning can be subdivided in filter methods and wrapper methods. Wrapper model approach uses the method of classification itself to measure the importance of features set; hence the feature selected depends on the classifier model used. Wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, wrapper methods are too expensive for large dimensional database in terms of computational complexity and time since each feature set considered must be evaluated with the classifier algorithm used.

III. SYSTEM DESIGN

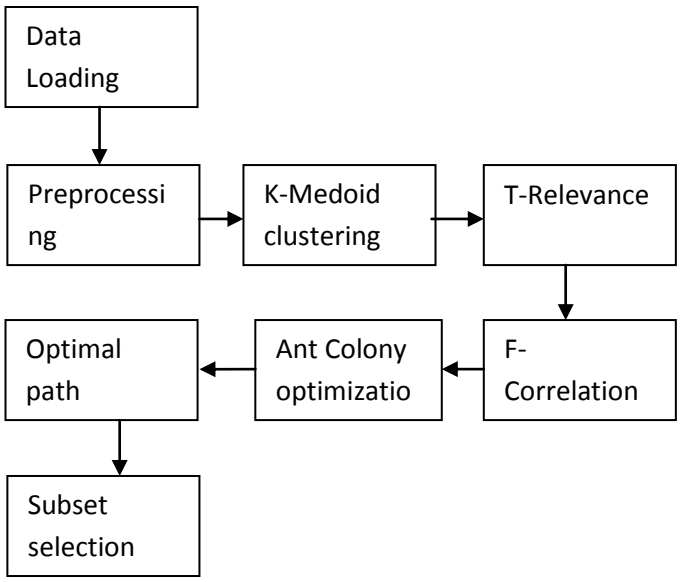

Fig1:System Design

IV. PROPOSED WORK

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. This algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods and 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. Pearson correlation is used for improving the efficiency and accuracy.

Advantages:

- Low Time Consuming process
- Effective search is achieved based on feature search.
- There should be no outliers in the data.
- Easy to cluster the values.

The proposed system has mainly six modules namely Load Data and Classify, Information Gain Computation, T-Relevance Calculation, Pearson-Correlation Calculation, MST Construction, Cluster Formation. The data has to be preprocessed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format which is the standard format for WEKA toolkit. From the arff format, only the attributes and the values are extracted and stored into the database.

DESIGN MODULES:

The Proposed System consists of following modules

- Data Loading and Preprocessing
- K-Medoid Clustering
- T-Relevance and F-Correlation Computation
- Find Optimal path
- Subset Selection

A. Data Loading and Preprocessing:

The dataset is selected and Loaded into the data base.In preprocessing step garbage value should be removed from documents. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.Data with different representations are put together and conflicts within the data are resolved.

B.K-Medoid Clustering:

The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster.

K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. A useful tool for determining k is the silhouette. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances .A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. i.e. it is a most centrally located point in the cluster.

C.T-Relevance Calculation:

The relevance between the feature Fi € F and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by SU(Fi,C). If SU(Fi,C) is greater than a predetermined threshold, we say that Fi is a strong T-Relevance feature.

$$S(F,C)= \frac{2*Gain(Fi \div c)}{H(Fi)+H(c)}$$

After finding the relevance value, the redundant attributes will be removed with respect to the threshold value. Irrelevant features are removed using T-relevance.

F-Correlation Calculation:

The correlation between any pair of features Fi and Fj (Fi,Fj € ^ F ^ i ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj). The equation symmetric uncertainty 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.

$$S(Fi,Fj) = \frac{2*Gain(Fi \div Fj)}{H(Fi)+H(Fj)}$$

Feature Selection:

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.

D. Find optimal path:

The Ant colony optimization algorithm (ACO) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs .Ant colony optimization algorithms have been applied to many combinatorial optimization problems, ranging from quadratic assignment to protein folding or routing vehicles and a lot of derived methods have been adapted to dynamic problems in real variables, stochastic problems, multi-targets and parallel implementations. It has also been used to produce near-optimal solutions to the travelling salesman problem. An ant is a simple computational agent in the ant colony optimization algorithm. It iteratively constructs a solution for the problem at hand. The intermediate solutions are referred to as solution states. At each iteration of the algorithm, each ant moves from a state x to state y, corresponding to a more complete intermediate solution.

E. Subset selection:

Relevant features are grouped into clusters and a representative of each cluster is retrieved to get a required feature without redundancy. The 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. Moreover, good feature subsets selection methods must be used to obtain features that are highly correlated with the class, yet uncorrelated with each other. A novel method is proposed which can efficiently and effectively deal with both irrelevant and redundant features, and obtains a good feature subset .Finally compare the features using existing algorithms.ant colony optimization algorithm is the best one to find the optimal path. so accuracy will be increased compare to existing algorithm.

V. EXPERIMENTAL RESULTS

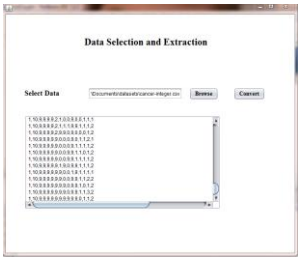

Fig2: Dataset Loading

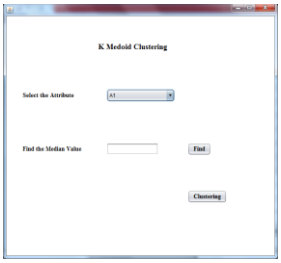

Fig3: K-medoids clustering

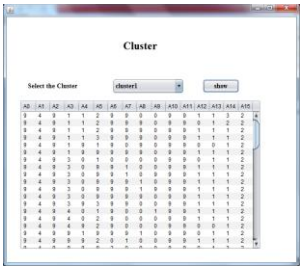

Fig4:Cluster Formation

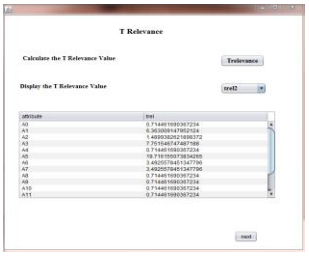

Fig5: T-Relevance Calculation

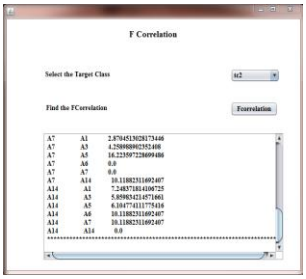

Fig6: F-Correlation Calculation

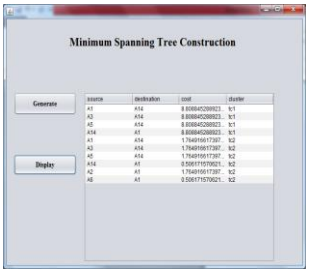

Fig7: Tree construction

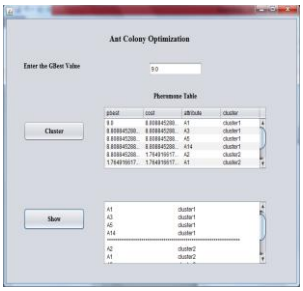

Fig8:Ant colony optimization

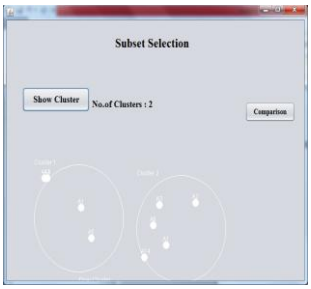

Fig9:Subset selection

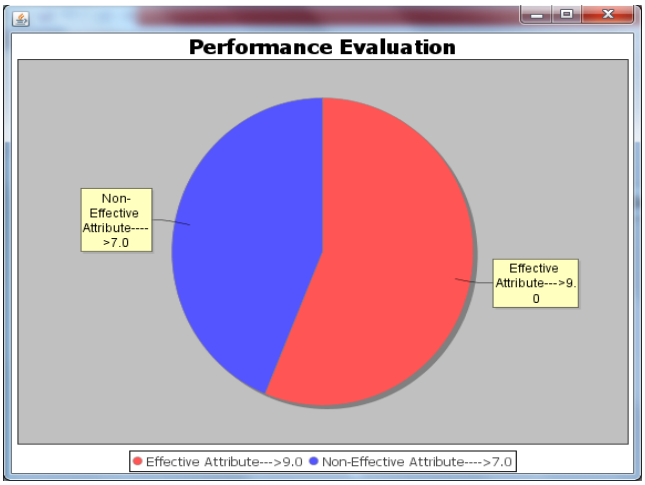

Fig7: Attribute selection

TABLE 1:Attribute count

Attribute Selection	Attribute count
Effective Attribute	9.0
Ineffective Attribute	7.0

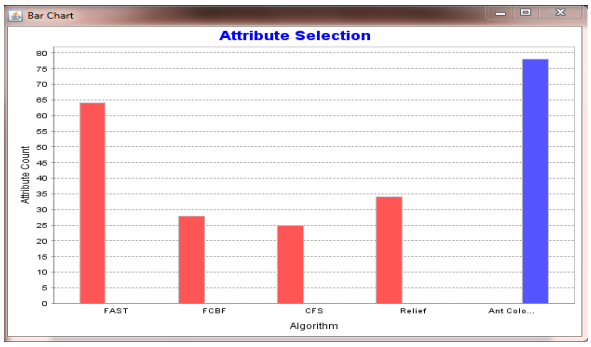

Fig8: Comparison Chart

Table2: Accuracy comparison for different algorithms

Algorithm	Accuracy
Ant colony	78
FAST	64
FCBF	28
CFS	25
Relief	34

VI CONCLUSION&FUTURE WORK

In this Project present Efficient Future subset selection for high dimensional data. This algorithm involves 1) removing irrelevant data using T-relevance and removing redundant data using F-Correlation.2)Then constructing a minimum spanning tree using Ant colony optimization algorithm from comparative ones, and 3) partitioning the Minimum spanning tree and selecting representative features. In this algorithm, a cluster consists of attributes. Each and every cluster is treated as a single feature and thus dimensionality is drastically reduced. The text data from the four dissimilar aspects of the section of chosen features, runtime, classification exactness of a given classifier. For the future work, we plan to explore different types of correlation measures, algorithms and study some formal properties of feature space to improve the accuracy and reduce the computational complexity.

VII. REFERENCES

[1] Asha Gowda Karegowda, M.A.Jayaram.A.S. Manjunath" Feature Subset Selection Problem using Wrapper Approach in Supervised Learning"International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7 13,2010.

[2] A.Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[3] A Krishna Mohan ,MHM Krishna Prasad," A Novel Feature Clustering Algorithm for Evaluation of Descriptive Type Examination", International Journal of Computer Applications (0975 – 8887) Volume 98– No.9, July 2014.

[4] Dr. Babu Reddy M "Independent Feature Elimination in High Dimensional Data : Empirical Study by applying Learning Vector Quantization method" International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 10, October 2013 ISSN 2319 – 4847.

[5] K. Jaganath1, Mr. P. Sasikumar2 "Graph Clustering and Feature Selection for High Dimensional Data" International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1, March 2014.

[6] Ms.Barkha Malay Joshi "High dimensional unsupervised clustering based feature selection algorithm" Ms.Barkha Malay Joshi et al. / International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.05 May 2012.

[7] Qinbao song, jingjie Ni, and Guangtao Wang,"A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data"IEEE Transation on Knowedge and Engineering,vol25,no1,january 2013.

[8] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.

[9] S.Vanaja, K.Ramesh kumar"Analysis of Feature Selection Algorithms on Classification: A Survey" International Journal of Computer Applications (0975 – 8887) Volume 96– No.17, June 2014.

[10] Yuxuan SUN_, Xiaojun LOU, Bisai BAO" A Novel Relief Feature Selection Algorithm Based on Mean-Variance Model" Journal of Information & Computational Science 8: 16(3921–3929) Journal of Information & Computational Science 8: 16 (2011) 3921–3929, 2011.

[11] Z. Zhao and H. Liu, "Searching for Interacting Features in Subset Selection," J. Intelligent Data Analysis, vol. 13, no. 2, pp. 207-228, 2009.

[12] Z. Zhao and H. Liu, "Searching for Interacting Features," Proc. 20th Int'l Joint Conf. Artificial Intelligence, 2007.