A SURVEY ON FEATURE SUBSET SELECTION ALGORITHMS USING FILTER METHOD

S. Kowsalya

ABSTRACT

Data Mining is at present an extended field of research that digs out new patterns used in decision making applications. It is used to crack problems of an explicit realm. Due to amplified data density, the magnitude of data to be mined also has increased. It involves a pre-processing step called Feature Subset Selection to act upon the Knowledge Discovery Process. The insignificant attributes irrelevant to the class of study must be removed from the existing databases. This filtering process is called Feature Subset Selection. This study involves the various methods used for the dimensionality reduction process. This survey takes into account the diverse kinds of Feature Subset Selection methods used to discover the best features and remove statistically insignificant values. This survey also provides a subsidiary idea for future enhancement in this field.

Keywords : *Data Mining, Feature Subset Selection, Feature Selection Algorithms, Filter methods, Graphical clustering.*

I. INTRODUCTION

As a consequence of technology step up there is breakneck accumulation of data in large volumes with high dimensions. Data mining is a process that converts these data clusters into nuggets. It is a method used to discover new patterns and knowledge from an extended hefty volume of data. Feature selection is a process that selects a subset among the original large volume of features. An optimal feature subset is to be always evaluated based on some criteria. Due to the technology enhancements the dimensionality of the domain inflates. The number of features to describe those domains also increases on hand. It becomes imperative to find an optimal feature subset. The main objective of feature selection is to avoid inappropriateness and improve the performance of decision making models to provide faster and more cost-effective solutions [1]. The selection of optimal features adds an increased burden in the modelling.

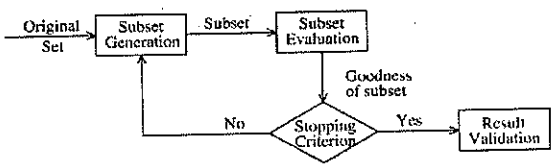

Figure 1 : Feature Selection Process

In 2005 Huan Liu and Lei Yu initiated a new feature selection process [2] that involved four steps to select the relevant subset data. This subset is to be derived from the given sample training data that is of high

Lecturer in the Department of Information Technology, Karpagam University, Coimbatore.
E-mail : kowsalya_1981@yahoo.co.in

dimensionality. The feature selection process is evaluated using vast extent of filter selection methods available. First step called Subset generation process is done in an adhoc manner using heuristic search techniques. Each state in the search space specifies a candidate subset for evaluation. This process assumes two set of predictions to be done in advance. Initially the user must decide the initial point which in turn influences the search direction. Therefore, different strategies have been explored for this process. They are complete, sequential and random search techniques. The second step is called Subset Evaluation. Each newly generated subset needs to be evaluated by an evaluation criterion. An evaluation criterion can be broadly categorized into two groups based on their dependency [3] on mining algorithms that will finally be applied on the selected feature subset. The evaluation criteria may be either Independent or Dependent Criteria. The third step Stopping Criteria determines when the feature selection process should stop. There are many frequently used stopping criteria. The user may follow the appropriate set of criteria based on the application developed. The final step in subset selection is the Result Validation. A straightforward way for result validation is to directly measure the result using prior knowledge about the data. If the relevant features are predicted in advance using historic data, it can be made in use in validating a resultant subset selected by any of the models.

Attribute selection process is divided into four categories based on the approach used by different methods. They are Embedded, Wrapper, Filter and Hybrid methods [4]. In Embedded methods the process of feature selection is embedded inside the training process itself. Traditional learning algorithms like Decision trees and Artificial Neural Networks use this type of approach. The second method named Wrapper defines a possible feature subset in the target space. The results produced by mining approach are used to obtain the relevance of attributes. The intrinsic properties of training data are only taken into consideration in Filter method. A score providing the feature relevance will be calculated. The attributes with low score of relevance will be removed. Classification algorithm gets the relevant feature subset list and reduces the dimensionality of the databases by removing the redundant features. A mixture of Filter and Wrapper method forms the Hybrid method. The Hybrid method uses both the features of Filter method and the Wrapper method to select the relevant attributes using the predictive accuracy. The Filter method is considered to be an efficient thing since a multidimensional dataset is reduced to a simple, fast and independent attributes list. Though Data mining process provides an immense key on useful knowledge extraction our survey focuses on the Filter method of attribute selection.

II. Related Work

In 1992 Kenji Kira and Larry A.Rendell proposed a Sequential and Distance based algorithms called RELIEF [5]. This algorithm selects relevant features using a feature weight selection method. It is an accurate algorithm even if the sample data is accumulated with noise. The time complexity depends on number of sample training data set and the number of features given. It does not depend on how complex the target class is. The RELIEF method is supplied with a training data set S, sample size m and a threshold value. The training data set S is subdivided into positive and negative instances. Each time a random positive and negative instance is picked up and its Near Hit or Near Miss instance is calculated using Euclid

Most of the feature selection approaches that use Filter method make use of the ranking process. This ranking determines the features ranked based on some order to reveal its predictability property. The ranking property is assumed to be made effective by the use of different types of scoring methods. This ranking criterion does not always ensure to eliminate redundant values. So there occurs the need for correlation based approaches. A binary feature selection process is proposed by Francois Fleuret [12] in 2004 that acts as a new method of feature selection using conditional mutual information. He proposed the conditional entropy H(U/V) based intuitive tool to pick features. If the two variables U and V are independent, no information can be gained from each other. So the value of conditional entropy H(U/V) is equal to the entropy itself. If they are dependent and deterministic then conditional entropy is zero as no new information is needed from U if V is known. If this conditional mutation is done sequentially then the computational argument generation will be in some case impossible. To maximize the conditional mutation information a random sampling method is followed. Although its performance is made better it does not fully avoid redundant data. The reason is that random selection may lead to redundant subset selection because all attributes is likely to have maximum information gain. This approach is applied for two types of datasets one with image data to find edges of face and the other with active molecule of drug design dataset. For both these methods binary selection of computation is underwent and the experiments show that mutation method experiences greater robustness when training set is accumulated with noisy data.

Most Data mining algorithms does not achieve good decisions if it works with the maximum number of attributes. So the importance of feature subset selection increased and the need to activate good selection techniques also drastically augmented. Some kind of feature selection process needs to be applied before the knowledge discovery process. An appropriate attribute set must be provided to the learning algorithms[13] . This necessitates the use of heuristic or probabilistic approaches to select relevant features. Most heuristic algorithms follow an exponential computation that extracts low order approximate relevance of features. Though this method enhances the reasonableness of measures, it may miss some features with high correlations.

The pre-processing step of feature selection may also lead to confusions thus moving way for false predictions and decision making. It may also some times memory inefficient and time consumers. So a good feature selection method must be followed. Yuxuan SUN et.al. in 2011 proposed another Relief feature selection method based on Mean-Variance model [14]. The RELIEF model proposed earlier computes feature weight measure for each attribute present in the training data set. This weight is calculated based on the difference in the weight of subsequent features. The features may be selected based on discrimination of same class and with different classes. The mean of all the weights selected is considered to be the final relevance measure. It is a time consuming process if for each feature the attribute measure is calculated. To solve this problem the mean-variance model of feature selection is proposed. This model obtains feature weight estimation based on the mean and variance. The most relevant feature W[F] is obtained that is a reasonable weight estimation W of feature F leads to minimal variance value. Using Lagrange Objective function a final weight measure problem is solved. This makes the result more

stable and accurate. If the sample data obtained from training set is random then the frequency of instance sampling is uncertain[15]. It leads to the fluctuation of the instances set.

Qinbao Song et.al. proposed a new clustering tree based feature subset selection method for high dimensional data. This algorithm FAST [16] was proposed in the year 2013. It evaluates to its result in two steps. In the first step, the relevance information of the features is calculated using Entropy method that supports correlation analysis. These features selected are used to construct a tree. The tree is divided into clusters [12] by using graph-theoretic clustering methods. In the second step, from the forest of set of clusters in the form of trees the most relevant feature is selected. This relevant feature is assumed to be the representative of the features in each cluster. It is the feature strongly related to target classes. Features in different clusters are relatively independent. From the given initial high dimensional data set the irrelevant features are removed first. The relevant features obtained as a result have a strong correlation with the target concept. This step is done by calculating the Symmetric Uncertainty (SU) value. This SU value represents the correlation between two features or between a feature and a concept. Using this, the irrelevant values are removed. In the next step the redundant features are eliminated. From the first step, the resultant data set without any irrelevant feature is acquired. This is taken as input for the next step. A minimum spanning tree structure is constructed that passes through all the possible edges. From this tree the partition is done to form a forest. Each forest represents the nodes with similar features. All the nodes of the representative forest seem to be similar and they have analogous characteristics. So each representative attribute is selected from every forest to form a resultant subset.

III. FINDINGS

This survey has a deep study on the accuracy of various algorithms applied to different classifiers. These algorithms when used in classifiers enhance their prediction capability and increases search performance. This study has analysed the classification accuracy of the four algorithms FAST, FCBF CFS and RELIEF with Naive Bayesian Classifier and C4.5 Classifier. The result of the study is illustrated with a chart. The **Table-1** shows the accuracy matrix of three types of data for four feature selection algorithms with Naïve Bayesian Classifier.

Table 1 : Accuracy Matrix for Naive Bayes

Data / Algorithm	Image	Microarray	Text
FAST	85.49	91.38	82.62
FCBF	87.32	84.30	66.83
CFS	86.58	87.22	70.12
Relief-F	84.97	79.22	61.43

The below chart depicts the representation of Table-1.

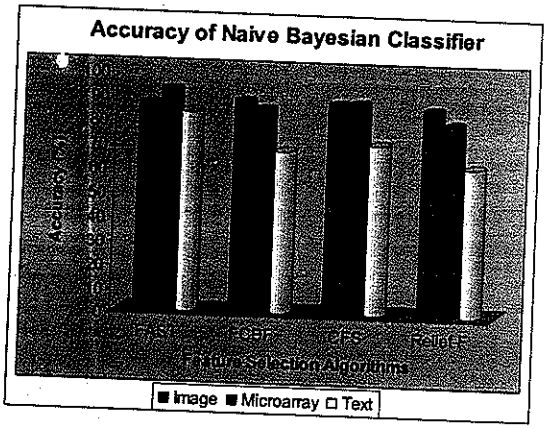

Figure 2 : Comparison of Accuracy of Naive Bayesian Classifier using four feature selection algorithms

Language Processing and Web Information Technology, pp 123-128, 2007.

[16] 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 Vol: 25 No: 1 Year 2013.

Author Biography

Mrs. S. Kowsalya has completed M.C.A. M.Phil., from Bharathiar University and pursuing M.E. (CSE) from Karpagam University. She is currently working as Lecturer in the Dept. of IT at Karpagam University. She has seven years experience in teaching. Her area of research includes Data mining. Se has presented three papers in National conferences and a paper in International conference.