Feature Selection Techniques in Data Mining: A Study

[1]K.Pavya, [2]Dr.B.Srinivasan

[1]Assistant Professor, [2]Associate Professor
Department of Computer Science,
Vellalar College for Women, Erode, India

***Abstract*- One of the major challenges these days is dealing with large amount of data extracted from the network that needs to be analyzed. Feature Selection plays the very important role in Intrusion Detection System. Feature Selection assists in selecting the minimum number of features from the number of features that need more computation time, large space, etc. Feature selection has become interest to many research areas which deal with machine learning and data mining, because it provides the classifiers to be fast, cost-effective, and more accurate.**

***Keywords*: Feature Selection, Data mining, Filter approach, Wrapper approach**

I. INTRODUCTION

Due to availability of large amounts of data from the last few decades, the analysis of data becomes more difficult manually. So the data analysis should be done computerized through Data Mining. Data Mining helps in fetching the hidden attributes on the basis of pattern, rules, so on. Data Mining is the only hope for clearing the confusion of patterns. Basically, the data gathered from the network are a raw data and contains large log files that need to be compressed. So the various feature selection techniques are used for eliminating the irrelevant or redundant features from the dataset. Feature selection [FS] is the processes that choose a subset of relevant features for building the model. Feature selection is one of the frequently used and most important techniques in data preprocessing for data mining [1].The goal of feature selection for classification task is to maximize classification accuracy [2].Feature selection is the process of removing redundant or irrelevant features from the original data set. So the carrying out time of the classifier that processes the data will decreases and also accuracy increases because irrelevant features can include noisy data affecting the classification accuracy negatively [3]. With feature selection the understandability can be improved and cost of data handling becomes smaller [4].

II. FEATURE SELECTION AND ITS METHODS

Data holds many features, but all the features may not be related so the feature selection is used so as to eliminate the unrelated features from the data without much loss of the information. Feature selection is also known as attributes selection or variable selection [5]. The feature selection is of three types:

- Filter approach
- Wrapper approach
- Embedded approach

2.1 Filter approach

Filter approach or Filter method shown in Figure 1. This method selects the feature without depending upon the type of classifier used. The advantage of this method is that, it is simple and independent of the type of classifier used so feature selection need to be done only once and drawback of this method is that it ignores the interaction with the classifier, ignores the feature dependencies, and lastly each feature considered separately.

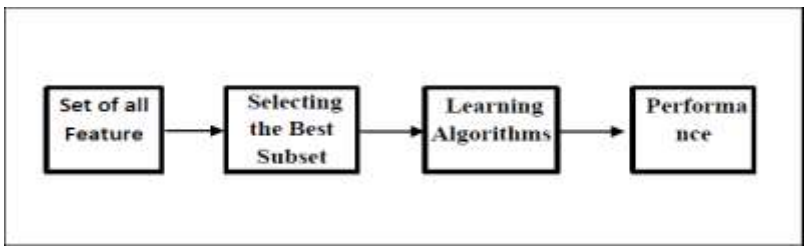

Fig 1: Filter Approach

2.2 Wrapper approach

Wrapper approach or Wrapper method is shown in Figure 2. In this method the feature is dependent upon the classifier used, i.e. it uses the result of the classifier to determine the goodness of the given feature or attribute. The advantage of this method is that it removes the drawback of the filter method, i.e. It includes the interaction with the classifier and also takes the

feature dependencies and drawback of this method is that it is slower than the filter method because it takes the dependencies also. The quality of the feature is directly measured by the performance of the classifier.

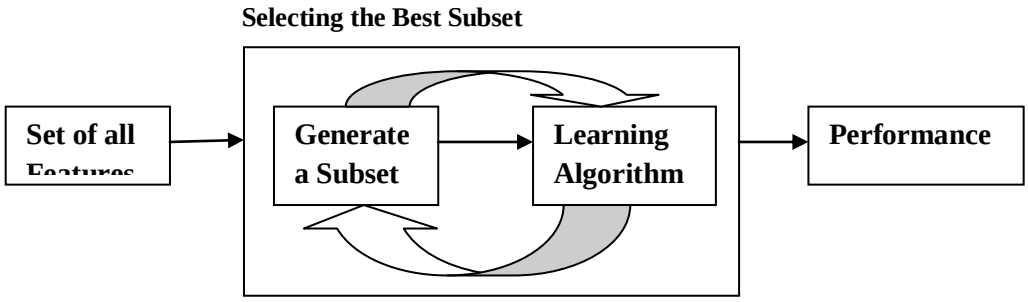

Fig 2: Wrapper Approach

2.3 Embedded approach

The embedded approach or embedded method is shown in Figure 3. It searches for an optimal subset of features that is built into the classifier construction. The advantage of this method is that it is less computationally intensive than a wrapper approach.

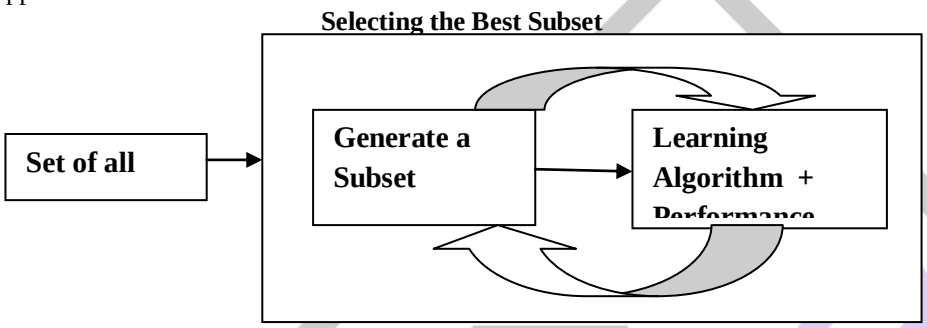

Fig 3: Embedded Approach

The accuracy of the classifier depends not only on the classification algorithm but also on the feature selection method used. Selection of irrelevant and inappropriate features may confuse the classifier and lead to incorrect results. The solution to this problem is Feature Selection i.e. feature selection is necessary in order to improve efficiency and accuracy of classifier. Feature selection selects subset of features from original set of features by removing the irrelevant and redundant features from the original dataset. It is also known as Attribute selection. Feature selection reduces the dimensionality of the dataset, increases the learning accuracy and improves result comprehensibility. The two search algorithms 'forward selection' and 'backward eliminations' are used to select and eliminate the appropriate feature. Feature selection is a three step process namely search, evaluate and stop.

Feature selection methods are also classified as attribute evaluation algorithms and subset evaluation algorithms. In first method, features are ranked individually and then a weight is assigned to each feature according to each feature's degree of relevance to the target feature. The second approach in contrast, selects feature subsets and then ranks them based on certain evaluation criteria. Attribute evaluation methods do not measure correlation between feature are hence likely to yield subsets with redundant features. Subset evaluation methods are more efficient in removing redundant features. Different types of feature selection algorithms have been proposed. The feature selection techniques are broadly categorized into three types: Filter methods, Wrapper methods, and Embedded methods. Every feature selection algorithm uses any one of the three feature selection techniques.

2.1 Filter methods

Ranking techniques are used as principle criteria in Filter method. The variables are assigned a score using a suitable ranking criterion and the variables having score below some threshold value are removed. These methods are computationally cheaper, avoids over fitting but Filter methods ignore dependencies between the features. Hence, the selected subset might not be optimal and a redundant subset might be obtained. The basic filter feature selection algorithms are as follows:

2.1.1 *Chi-square test*

The chi-squared filter method test checks the independence between two events. The two events X, Y are defined to be independent if P(XY) = P(X)P(Y) or equivalently P(X/Y) = P(X) and P(Y/X) = P(Y). More particularly in feature selection it is used to test whether the occurrence of a specific term and the occurrence of a specific class are independent. Thus the following quantity for each terms are estimated and rank them by their score: In equation: (1) High scores on $\chi 2$ indicate that the null hypothesis (H0) of independence should be eliminated and thus that the occurrence of the term and class are dependent.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}. \qquad (1)$$

2.1.2 ***Euclidean Distance***

In this feature selection technique, the correlation between features is calculated in terms of Euclidean distance. If sample feature say 'a' contains 'n', then these 'n' number of features are compared with other 'n-1' features by calculating the distance between them using the following equation: (2) The distance between features remains unaffected even after addition of new features.

$$d(a,b) = \{\sum_i (a_i - b_i)^2\}^{1/2} \quad (2)$$

2.1.3 ***Correlation criteria***

Pearson correlation coefficient is simplest criteria and is defined by the following equation: (3) Where, xi is i_{th} variable, Y is the output class, var() is the variance and cov() denotes covariance. The disadvantage is that correlation ranking can only detect linear dependencies between variable and target.

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}} \quad (3)$$

2.1.4 ***Information Gain***

Information gain tells us how important a given attribute of the feature vectors is. IG feature selection method selects the terms having the highest information gain scores. Information gain measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature) defined as:

$$Entropy = \sum_{i=1}^{n} - p_i log_2 p_i \quad (4)$$

Where _n' is the number of classes, and the Pi is the probability of S belongs to class 'i'. The gain of A and S is calculated as:

$$Gain(A) = Entropy(S) - \sum_{k=1}^{m} \frac{S_k}{S} * \mathbf{Entropy(Sk)} \quad (5)$$

Where, Sk is the subset of S.

2.1.5 ***Mutual Information***

Information theoretic ranking criteria [ ] uses the measure of dependency between two variables. To describe MI we must start with Shannon's definition for entropy given as:

$$\mathbf{H(X)} = -\sum_i \mathbf{P(y) \log P(y)} \quad (6)$$

Above equation represents the uncertainty (information content) in output Y. Suppose we observe a variable X then the conditional entropy is given by:

$$H(Y/X) = -\Sigma_i \sum_y P(x,y) \log P(y/x) \quad (7)$$

Above equation implies that by observing a variable X, the uncertainty in the output Y is reduced. The decrease in uncertainty is given as:

$$I(Y, X) = H(Y) - H(Y \mid X) \quad (8)$$

This gives the MI between Y and X meaning that if X and Y are independent then MI will be zero and greater than zero if they are dependent. This implies that one variable can provide information about the other thus proving dependency. The definitions provided above are given for discrete variables and the same can be obtained for continuous variables by replacing the summations with integrations.

2.1.6 ***Correlation based Feature Selection (CFS)***

Correlation-based Feature Selection algorithm selects attributes by using a heuristic which measures the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The highly correlated and irrelevant features are avoided. The equation used to filter out the irrelevant, redundant feature which leads the poor prediction of the class is defined as:

$$F_s = \frac{N * r_a}{N + N(N-1) r_n} \quad (9)$$

2.1.7 ***Fast Correlation based Feature Selection***

FCBF (Fast Correlation Based Filter) [4] is a multivariate feature selection method which starts with full set of features, uses symmetrical uncertainty to calculate dependences of features and finds finest subset using backward selection technique with sequential search strategy. The FCBF algorithm consists of two stages: the first one is a relevance analysis that orders the input variables depending on a relevance score, which is computed as the symmetric uncertainty with respect to the target output. This stage is also used to discard irrelevant variables, whose ranking score is below a predefined threshold. The second stage is a

redundancy analysis, which selects predominant features from the relevant set obtained in the first stage. This selection is an iterative process that removes those variables which form an approximate Markov blanket. Symmetrical Uncertainty (SU) is a normalized information theoretic measure which uses entropy and conditional entropy values to calculate dependencies of features. In Symmetrical Uncertainty the value 0 indicates that two features are totally independent and value of 1 indicates that using one feature other feature's value can be totally predicted.

2.2 Wrapper methods

Wrapper methods are better in defining optimal features rather than simply relevant features. They do this by using heuristics of the learning algorithm and the training set. Backward elimination is used by the wrapper method to remove the insignificant features from the subset. The SVM-RFE is one of the feature selection algorithms which use the Wrapper method. The Wrapper method needs some predefined learning algorithm to identify the relevant feature. It has interaction with classification algorithm. The over fitting of feature is avoided using the cross validation. Though wrapper methods are computationally expensive and take more time compared to the filter method, they give more accurate results than filter model. In filter model, optimal features can be obtained rather than simply relevant features. Another advantage is it maintains dependencies between features and feature subsets. Wrapper methods are broadly classified as sequential selection algorithms and heuristic search algorithms as follows:

2.2.1 *Sequential Selection Algorithms*

The Sequential Feature Selection (SFS) [7][8][9] algorithm starts with an empty set and adds one feature for the first step which gives the highest value for the objective function. After the first step, the remaining features are added individually to the current subset and the new subset is evaluated. The individual features that give maximum classification accuracy are permanently included in the subset. The process is repeated until we get required number of features. This algorithm is called a naive SFS algorithm since the dependency between the features is not taken into consideration.

A Sequential Backward Selection (SBS)[10][11] algorithm is exactly reverse of SFS algorithm. Initially, the algorithm starts from the entire set of variables and removes one irrelevant feature at a time whose removal gives the lowest decrease in predictor performance. The Sequential Floating Forward Selection (SFFS) algorithm is more flexible than the naive SFS because it introduces an additional backtracking step. The algorithm starts same as the SFS algorithm which adds one feature at a time based on the objective function. SFFS algorithm then applies one step of SBS algorithm which excludes one feature at a time from the subset obtained in the first step and evaluates the new subsets. If excluding a feature increases the value of the objective function then that feature is removed and algorithm switches back to the first step with the new reduced subset or else the algorithm is repeated from the top. The entire process is repeated until the required numbers of features are obtained or required performance is reached. SFS and SFFS produce nested subsets since forward inclusion was unconditional.

2.2.2 *Heuristic Search Algorithms*

Heuristic search algorithms include Genetic algorithms (GA)[12], Ant Colony Optimization(ACO)[13], Particle Swarm Optimization(PSO)[14],etc. A genetic algorithm is a search technique used in computing to find true or approximate solution to optimization and search problems. Genetic algorithms are based on the Darwinian principle of survival of the fittest theory. ACO is based on the shortest paths found by real ants in their search for food sources. ACO approaches suffer from inadequate rules of pheromone update and heuristic information. They do not consider random phenomenon of ants during subset formations. PSO approach does not employ crossover and mutation operators, hence is efficient over GA but requires several mathematical operators. Such mathematical operations require various user-specified parameters and dealing with these parameters, deciding their optimal values might be difficult for users. Although these ACO and PSO algorithms execute almost identically to GA, GA has received much attention due to its simplicity and powerful search capability upon the exponential search spaces.

2.3 Embedded methods

In embedded method [15], a feature selection method is incorporated into a learning algorithm and optimized for it. It is also called the hybrid model which is combination of filter and wrapper method. Embedded methods [16] reduce the computation time taken up for reclassifying different subsets which is done in wrapper methods. The KP-SVM is the example for embedded method. The problem of nesting effect of SFS and SFFS was overcome by developing an adaptive version of SFFS called Adaptive Sequential Forward floating Selection (ASFFS) algorithm. In ASFFS algorithm, two parameters ‘r‘ and ‘o‘ are used where ‘r‘ specifies number of features to be added while parameter ‘o‘ specifies number of features to be excluded from the set so as to obtain less redundant subset than the SFFS algorithm. The Plus L is a generalization of SFS and the Minus R is the generalization of SBE algorithm. If L>R then the algorithm start with SFS algorithm i.e. start from empty set and add the necessary features to the resultant set, else the algorithm start with the SBE algorithm i.e. start from entire set and start eliminating the irrelevant features and produce the resultant set. The Plus-L-Minus-r search method also tries to avoid nesting. In this method, the parameters L and r have to be chosen arbitrarily. It consumes less time than wrapper method but gives less accurate results than wrapper model as some of the important features may be lost by the filter model.

III. NEED FOR FEATURE SELECTION

- Reduces the size of the problem.
- Reduce the requirement of computer storage.
- Reduce the computation time.
- Reduction in features to improve the quality of prediction.
- To improve the classifier by removing the irrelevant features and noise.

- To identify the relevant features for any specific problem.
- To improve the performance of learning algorithm.

IV. CONCLUSION

Feature selection is an important issue in classification, because it may have a considerable effect on accuracy of the classifier. It reduces the number of dimensions of the dataset, so the processor and memory usage reduce; the data becomes more comprehensible and easier to study on. In this study, various feature selection techniques have been discussed and among the three approaches to feature selection method, filter methods should be used to get results in lesser time and for large datasets. If the results to be accurate and optimal, wrapper method like GA should be used.

REFERENCES

[1] Asha Gowda Karegowda, M.A.Jayaram and A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications, Vol. 1, No. 7, pp. 0975–8887, 2010.
[2] Ron Kohavi, George H. John, "Wrappers for feature subset Selection", Artificial Intelligence, Vol. 97, No. 1-2. pp. 273-324, 1997.
[3] S. Doraisami, S. Golzari, A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, Content-Based Retrieval, Categorization and Similarity, 2008
[4] A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benítez, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications, 38 (2011) 8170-8177.
[5] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International Journal of Engineering Research & Technology (IJERT), 1(6).
[6] Uysal, A. K., & Gunal, S., ―A novel probabilistic feature selection method for text classification,‖ Knowledge-Based Systems, 36, 226–235, 2012.
[7] S. Guan, J. Liu, Y. Qi, ―An incremental approach to contribution-based feature selection,‖ Journal of Intelligence Systems 13 (1), 2004.
[8] M.M. Kabir, M.M. Islam, K. Murase, ―A new wrapper feature selection approach using neural network,‖ in: Proceedings of the Joint Fourth International Conference on Soft Computing and Intelligent Systems and Ninth International Symposium on Advanced Intelligent Systems (SCIS&ISIS2008), Japan, pp. 1953–1958, 2008.
[9] M.M. Kabir, M.M. Islam, K. Murase, ―A new wrapper feature selection approach using neural network,‖ Neurocomputing 73, 3273–3283, May 2010.
[10] E. Gasca, J. Sanchez, R. Alonso, ―Eliminating redundancy and irrelevance using a new MLP-based feature selection method,‖ Pattern Recognition 39, 313–315, 2006.
[11] C. Hsu, H. Huang, D. Schuschel, ―The ANNIGMA-wrapper approach to fast feature selection for neural nets,‖ IEEE Transaction son Systems, Man, and Cybernetics—Part B:Cybernetics32(2)207–212, April 2002.
[12] A. Ghareb , A. Bakar, A. Hamdan, ―Hybrid feature selection based on enhanced genetic algorithm for text categorization,‖ Expert SystemsWith Applications, Elsevier, 2015.
[13] R.K. Sivagaminathan, S. Ramakrishnan, ―A hybrid approach for feature subset selection using neural networks and ant colony optimization,‖ Expert Systems with Applications 33, 49–60, 2007.
[14] X. Wang, J. Yang, X. Teng, W. Xia, R. Jensen, ―Feature selection based on rough sets and particle swarm optimization,‖ Pattern Recognition Letters 28 (4), 459–471, November 2006.
[15] M.M. Kabir, M.M. Islam, K. Murase, ―A new local search based hybrid genetic algorithm for feature selection,‖ Neurocomputing 74, 2194-2928, May 2011.
[16] M. Zhu, J. Song, ―An Embedded Backward Feature Selection Method for MCLP Classification Algorithm,‖ Information Technology and Quantitative Management, Elsevier, 2013.