



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Wrapper based Feature Selection and Classification for Real Time Dataset

S.Visalakshi¹, V.Radha²

¹Research Scholar, ²Professor

^{1,2}Department of Computer Science,

Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamil Nadu, INDIA

Abstract: Detection and identification of natural or intentional contamination have become very critical in drinking water. Feature selection plays an important role in both detection and classification and has become an active research area in pattern recognition, statistics and data mining communities. As the dimensionality enlarges, the feature set increases in size. Feature selection is a pre-processing technique commonly used in all high dimensional data. The focus of feature selection is to select the subset of features (variables) from the input, which can efficiently describe the input data, by identifying irrelevant data or noise to produce high-quality results of prediction. A feature is one of the measurable properties of the process being observed. In this research, an attempt has been made to investigate the pre-processing framework for feature selection in drinking water dataset, using Back Propagation Neural Network (BPNN) and Support Vector Machine (SVM). The main purpose of this work is to select a subset of features or list of attributes or variables which help to describe data for constructing models. Selection of features and training of the model are performed simultaneously in this work. The wrapper based method is used to select the subset of features, and the model is trained with SVM classifier. SVM and BPNN classifiers are used to train drinking water dataset for two different seasons, and comparisons are made between two classifiers, to conclude that the performance of SVM is better when compared to BPNN.

Keywords: subset generation, subset evaluation, embedded method, feature subset selection, feature relevance, feature selection.

I. INTRODUCTION

Feature selection is a vigorous and productive field of research area in machine learning, pattern recognition, data mining, statistics communities, text categorization, image retrieval, customer relationship management, intrusion detection, automatic scene annotation, gene function prediction, etc. Feature selection is a process of selecting a subset of related features which are used for model construction. The optimal feature selection is the process of finding a subset of $X_{opt} = \{x_{1,opt}, x_{2,opt}, \dots, x_{m,opt}\}$ which contains $m \leq n$ features from the original set of features $X_{opt} \subseteq X_{all}$. In order to use data mining tools effectively, researchers and practitioners realize that data preprocessing is important for a successful data mining process. The main purpose of feature selection is that, reduces in dimensionality, reduces the number of redundant features, removes irrelevant or noisy data, reduces the amount of data which is needed for learning processes, improves the accuracy of the result, speeds up the process of data mining algorithm and improves the mining process.

Feature selection improves the performance of classifier models and works effectively for high dimensional data [9,18, 8, 20]. Feature selection is also known as feature subset selection. Generally, the feature selection can be viewed as a search problem, where each state of search space specifies a subset of the possible features. For example, the dataset has three different features, namely, $\{f_1, f_2, f_3\}$, and when selecting the features, the presence of the feature will be coded in binary form, i.e., 0 implies features is not present and 1 denotes the feature is present in the feature set. Then, there will be a total of 2^3 reduced features subset which will be coded as $\{0,0,0\}$, $\{0,0,1\}$, $\{0,1,0\}$, $\{0,1,1\}$, $\{1,0,0\}$, $\{1,0,1\}$, $\{1,1,0\}$ and $\{1,1,1\}$. The problem of feature selection depends upon the size of data, i.e., when the search space is small, then searching of a feature in any order will be over within a short period, but when the search space is large, then it will be 2^N where the number of dimension N in data mining application is large. Feature selection is evaluated using some criteria such as error probability, distance measures, dependence measures, entropy measures and interclass distance measures [12].

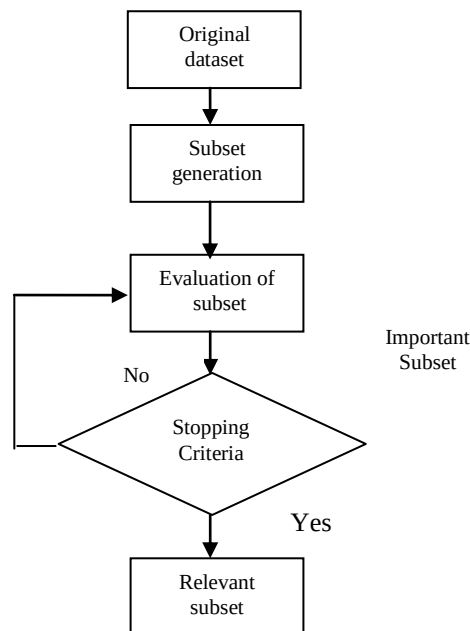
In data mining, extracting features from the dataset are carried out via different stages such as clustering, classification, association rules and regression. The feature selection can be classified mainly in two ways: selecting a subset of original feature set and extract less number of features [16]. The classification accuracy will decrease when there is presence of relevant data or highly interrelated features present in the process of feature selection. In feature selection, it must satisfy all criteria such as selective power, performance of model or subset cardinality [19, 7]. The main objective of feature selection is to select a subset of features. Feature

selection is an efficient technique for dimension reduction and its success in all kinds of data mining applications. There are three main objectives of feature selection: Improving the performance of the data mining model, Providing faster and most cost effective learning process and providing better understanding about the data present in dataset. The feature selection will select the features which are most informative and measure the features using some constraints such as information gain, minimum description, etc. There are many advantages in feature selection:

- Facilitate the visualization of data and understanding the data,
- Reduce in measurement and storage requirement,
- Reduce in training and utilization times,
- Decrease the problem of curse of dimensionality,
- Improve the prediction performance,
- Speed up the processing of algorithms,
- Enhance in data quality, and
- Increase in prediction accuracy of an algorithm [5].

The main goal of feature selection task is to see that the selected features achieve intended objectives. Phase of assessment includes a number of feature selection, time, scalability and performance of learning model. No universal algorithm is superior to another and each feature selection algorithm will possess special edges for various applications [3]. The algorithm of feature selection helps to evaluate the relevant features in two dissimilar traditions: Individual Evaluation and Subset Evaluation. Individual Evaluation is less in computation. This approach reviews individual feature and assign weights to each feature (ranks) according to the degree of class prediction. On the other hand, the individual feature evaluation is unable to detect redundant features because of similar rankings. Subset evaluation is to handle both feature relevance and feature redundancy. Compared to individual evaluation, this approach is capable of evaluating the features against the subset of features. Thus, computational cost is high [6]. Generally, feature selection consists of four basic steps and the process is depicted in the following Figure 1.

Figure 1: Process of Feature Selection



Subset generation: The subset generation procedure basically generates subsets of features for evaluation. The subset generation will start with no features, with all features and with a random subset of features. In the two cases, such as with no features and with all features, features are further added or removed iteratively, and in random subset of features, the features can be either added or removed or even produce random features iteratively.

Subset evaluation: The evaluation function helps to check the decency of subset produced by subset generation procedure and the new subset value is compared with the preceding best feature. If it is found better than the preceding feature is replaced with current feature.

Stop criterion: The stop criterion is based on generation procedure and the necessary features are selected until the iteration is reached.

Validating the result: It assess the validity of the selected subset by carrying out different tests and compare it with the previous results using real-world dataset or bench mark datasets [15, 13].

Each and every subset candidate is evaluated and compared with the preceding subset according to evaluation criteria. While validating, if the new generated subset is better, immediately it replaces the preceding best subset with new subset. In the process of feature selection, subset generation and evaluation will continue until the stopping criterion is reached which is portraying in below Figure 1.

The paper is organized as follows. Section 2 presents the literature review. Section 3 deals with the feature selection approaches. Section 4 outlines the methodology of work. Section 5 shows the experimental results and Section 6 concludes the paper.

II. LITERATURE REVIEW

Feature selection becomes a thrust area of research in data mining. Here, the overviews of different feature selection methods are presented along with classifiers.

According to Cheng-San Yang, Li-Yeh Chuang, Chao-Hsuan Ke and Cheng-Hong Yang [4], Gene expression profile represent the state of a cell at a molecular level, which is having great potential in medical diagnostic tool. To improve the classification accuracy, Correlation-Based Feature Selection (CFS) and Taguchi Genetic Algorithm (TGA) are combined into a hybrid method and KNN is used as a classifier.

Ozden Gur Ali and Kubra Yaman [11], propose a new algorithm of Row and Column Selection Algorithm (ROCSA), to select a small but informative dataset, to train a model using support vector regression with standard SVM tools. ROCSA uses ϵ -SVR (Support Vector Regression) models for the row and column selection steps. To evaluate this algorithm, six grocery datasets are used, and these dataset are trained with the standard SVM tool. Result states that, the data trained with ROCSA shows a 39% improvement in test accuracy vs. the full dataset with all rows and columns.

Adnan Idris, Muhammad Rizwan and Asifullah Khan [10] states that the telecommunication sector faces a ferocious competition to retain customers, and therefore they need an efficient churn the prediction model to monitor the behaviour of customers. The dataset of telecommunication is high, and hence becomes a challenge for churn prediction. The Particle Swarm Optimization (PSO) under sampling method is employed, to handle the imbalance data distribution, in collaboration with different feature reduction techniques such as Principal Component Analysis (PCA), Fisher's ratio, F -score and Minimum Redundancy and Maximum Relevance (mRMR) are used for dimensionality reduction of dataset. k-Nearest Neighbour (kNN) and Random Forest (RF) classifiers are used to evaluate the performance of optimally sampled and reduced feature set. The prediction performance is evaluated using sensitivity, specificity and Area Under the Curve (AUC). In the proposed approach Chr-PrRF, samplings are done based on PSO. Feature is selected using mRMR based, and RF classifier effectively handles the problem of churn prediction in telecommunication. The paper concludes that the proposed approach can be used in telecommunication industry, even if the size of data is large and involves high computation cost for churn prediction.

As per Sven F. Crone and Nikolaos Kourentzes [4], time series prediction faces multiple challenges. The greatest challenge in time series is specifying network architecture. There is need of universal methodology for heterogeneous datasets of time series. This paper proposes data driven methodology which combines both filter and wrapper approach for selecting features, feature evaluation, construction and transformation. The proposed methodology discovers time series creating patterns and transforming variables and state multilayer perceptions for time series data. This section concludes some work progressed already in feature selection.

III. FEATURE SELECTION METHODS/APPROACHES

Generally, the feature selection approaches are divided into three different types. They are Filter, Wrapper and Embedded Methods. The working process of these approaches differ in three ways: Search strategies, Defining the evaluation criterion and Estimation of evaluation criterion.

Filter Method: The features are selected and evaluation made before the data mining algorithm is applied to the dataset. Ranking techniques are used as the principal criteria for variable selection. Due to the simplicity of ranking methods, success rate is reported high for sensible applications. Ranking criterion is used to score the variables and threshold helps to remove a variable which is below the threshold value. Ranking methods are applied before classification to filter out irrelevant variables. The basic property of each feature must contain valid information on the different classes in the data. This property is defined as feature relevance. The various authors presented various definitions and measurement for the relevance of a variable [2, 14, 17]. When the

feature has no influence with class labels, those labels can be discarded. The advantage of feature ranking is low in computation cost and avoids over fitting. The filter method relies on feature relevance criteria, but the wrapper method relies on the classification method to obtain a feature subset. Some filter methods are correlation based information, t-test, information gain, mutual information and entropy based methods.

Wrapper Method: This method requires a learning algorithm, and performances are evaluated to find out the goodness of features. It searches the better features which will suit the learning algorithm and aim to improve the performance of mining. It is expensive compared to the filter method. In this approach, classification algorithm is used to evaluate the goodness of the selected feature subset. This method mainly focuses on improving the accuracy of classification, pattern identifying and perform better than filter method. In this method, to perform feature selection for the purpose of training and testing, genetic algorithm, branch and bound algorithm, tabu search, etc, are used.

Embedded Method: This method selects the features, based on criteria that are generated during the learning process of a classifier. The features selected in this method are sensitive to the structures of fundamental classifier. This reason leads to a minor problem, i.e., the feature selected by one embedded method might not be suitable to other feature [1]. Generally, embedded method uses inductive algorithm, where it will represent feature selector and classifier [10].

Thus, filter will evaluate each feature separately from the classifier, features will be ranked after the evaluation process is over and the superior ones will be considered. Wrapper takes subset of features and estimates the classification performance on each feature subset, and another subset will be evaluated in the classifier. The good subset feature will be selected for further process. Embedded method will perform feature selection during the learning process [4]. The above section presents various methods for feature selection.

IV. METHODOLOGY

Feature selection for SVMs

According to the literature study feature selection are categorized in three methods which are discussed in the above section. Before applying the classification algorithm, filter out surplus feature from the dataset. Based on the literature study, in this paper, the relevant feature set is given as input to wrapper part and model is trained by popular methods such SVM and BPNN. Generally, the wrapper method investigates the whole set of variables in datasets. The algorithm of wrapper method is computationally challenging but it gives accurate results compare to filter method. In the first phase of two stage design, a classifier independent feature selection method is used to remove most of the irrelevant features.

4.1. Initial Feature Selection Method

Steps for selecting features:

1. Consider F is having a feature set i.e. $F = (f_1, f_2, \dots, f_n)$.

- Calculate score for each feature f_i .
- To evaluate the relevant features, standard deviation and mean are used. The formula used to select relevant features are given below:

$$D_i = \frac{|m_1^i - m_2^i|}{std_1^i + std_2^i} \quad (1)$$

- Correlation between f_i and f_j =

$$\rho_{i,j} = \prod_{c=1}^2 \frac{cov(S_c(f_i), S_c(f_j))}{\sqrt{var(S_c(f_i)) \cdot var(S_c(f_j))}} \quad (2)$$

where S_c is the training vector for feature f_i which is labeled as class C. A High score indicates important feature.

2. Score value is calculated using the following formula:

$$R_i, F_s = \frac{D_i}{\max\{D_i\}} - |\rho_i F_s| \quad (3)$$

The features are selected using mean and standard deviation (f_i is considered). To find out the correlation between f_i and f_j , the correlation coefficient is used and finally high value indicates the relevant features for the further research are discussed in this section.

4.2. Wrapper based feature selection:

In the second stage of the wrapper based method is used in this research for further refine the feature set from first stage. Wrapper method will search for relevant feature subset in combination with a fixed classification method. The internal validation method helps to identify the goodness of feature subset. Different optimization frameworks are applied to search the best feature subset. They are, forward selection, backward elimination, hill climbing, beam search and randomized algorithm [11]. In this paper, forward feature selection algorithm is used

for feature selection, and classifier is used to train the dataset. Feature pruning is applied for increasing the generalization ability of classifier. The process of feature selection will start with empty set, i.e., $S=NULL$. The features will be added to S until SVM optimization is minimized. The resultant feature set will be given to wrapper stage, and resultant optimal feature will be trained with BPNN and SVM. To evaluate the performance of feature selection, Classification Accuracy, Speed and Normalized Root Mean Square Error (NRMSE) are used. The result shows that SVM performs better when compared to BPNN.

V. EXPERIMENTAL RESULTS

From high dimensional dataset, lots of features are extracted. The hybrid approach is adopted in the case of detection contamination in drinking water. The initial number of features is first reduced using filter method and again the wrapper method is applied to reduce the feature set. The dataset is collected from the Pillur, TWAD board, Coimbatore district for a period of two years (2012 & 2013) i.e. summer and winter seasons. Approximately 500+ data are used for experimentation. The classification accuracy, speed and NRMSE are evaluated for both BPNN and SVM classifiers.

TABLE 1: CLASSIFICATION ACCURACY

Datasets	Features (%)	Wrapper + BPNN	Wrapper + SVM
Summer	0	84.36	87.23
	10	90.03	92.92
	20	86.54	88.53
	30	85.06	89.03
	40	81.24	87.55
Winter	0	83.36	87.02
	10	90.07	94.27
	20	87.35	88.27
	30	86.64	87.42
	40	86.45	88.56

The above Table 1 presents the classification accuracy of wrapper based feature selection with BPNN and wrapper based feature selection with SVM for two different seasons. It is evident that the classification accuracy of SVM is increased when compared with BPNN.

TABLE 2: EXECUTION SPEED

Datasets	Features (%)	Wrapper + BPNN	Wrapper + SVM
Summer	0	5.14	5.01
	10	5.19	5.09
	20	5.39	5.3
	30	6.22	6.18
	40	7.11	7.02
Winter	0	6.36	6.28
	10	6.34	6.3
	20	6.88	6.78
	30	8.24	8.15
	40	8.91	8.86

The execution speed of BPNN and SVM with wrapper based feature selection is listed in the below Table 2 for summer and winter seasons, and it is clear that the execution speed is reduced in the case of SVM. Thus, the experimental results show that SVM classifier works effectively for identifying features in drinking water dataset.

The NRMSE is calculated for both the classifiers such as BPNN and SVM and presented in the below Table 3 for two different seasons, and it is evident that the NRMSE values of SVM is minimized compared with BPNN classifier.

TABLE 3: NRMSE

Datasets	Features (%)	Wrapper + BPNN	Wrapper + SVM
Summer	0	0.443	0.4306
	10	0.397	0.3914
	20	0.402	0.3953
	30	0.419	0.4132
	40	0.418	0.4114
Winter	0	0.448	0.459
	10	0.433	0.426
	20	0.452	0.4438
	30	0.436	0.4276
	40	0.449	0.4409

VI. CONCLUSION

This paper focuses on detecting important features from the dataset. The mean and standard deviation are used to select initial features from the dataset. Each feature is assigned to set S until the optimization of SVM is minimized in nature. The relevant features set is taken to wrapper stage for selecting optimal features and classifiers are used to train the dataset. The tentative result proves that, compared to BPNN the SVM classifier performs well in terms of classification accuracy, speed and NRMSE in the drinking water dataset. In the future, identifying the type of contamination and alerting the systems to take necessary action before reaching the public for usage can be considered.

VII. References

- [1] Adnan Idris, Muhammad Rizwan and Asifullah Khan, Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies, *Computers and Electrical Engineering*, 2012, vol.38, pp.1808-1819.
- [2] Cheng-San Yang, Li-Yeh Chuang, Chao Hsuan Ke, and Cheng-Hong Yang, A Hybrid Feature Selection Method for Microarray classification, *IAENG International Journal of Computer Science*, 2008, vol.35, no.3.
- [3] M. Dash and H. Liu, Feature Selection for Classification, *Elsevier, Intelligent Data Analysis*, 1997, pp.131-156.
- [4] Esra Mahsereci Karabulut, Slema Ayse Ozel and Turgay Ibrici, A comparative study on the effect of feature selection on classification accuracy, *SciVerse Science Direct, Elsevier*, 2012, vol.1, pp.323-327.
- [5] Haun Liu and Hiroshi Motoda, Computational Methods of Feature Selection, *Chapman & Hall/CRC*, 2008, pp.4-6.
- [6] Huan Liu, Hiroshi Motoda, Rudy Setiono and Zheng Zhao, Feature Selection: An Ever Evolving Frontier in Data Mining, *JMLR: The fourth Workshop on Feature Selection in Data Mining*, 2010, vol.10, pp.4-13.
- [7] Isabelle Guyon and Andr e Elisseeff, An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 2003, vol.3, pp.1157-1182.
- [8] Jillang Tang and Huan Liu, An Unsupervised Feature Selection Framework for Social Media Data, *IEEE Transactions and Knowledge and Data Engineering*, 2012, pp.1-14.
- [9] Krzysztof J. Cios, Witold Pedrys, Roman W. Swintars and Lukaz AS. Kurgan, Feature Extraction and Selection, *Springer*, 2007, pp.207-210.
- [10] Li-Yeh Chuang, Cheng-HueiYang, Kuo ChuanWu and Cheng-HongYang, A hybrid feature selection method for DNA microarray data, *Elsevier, Computers in Biology and Medicine*, 2011, vol.41, pp.228-237.
- [11] Milos Hauskrecht, Richard Pelikan, Michal Valko, and James Lyons-Weiler, Feature Selection and Dimensionality Reduction in Genomics and Proteomics, *Fundamental of Data Mining in Genomics and Proteomics, Springer US*, ISBN.978-0-387-47509-7, 2007, pp.149-172.
- [12] Monami Banerjee and Nikhil R. Pal, Feature selection with SVD entropy: Some modification and extension, *Infor Science, Elsevier*, 2014, vol.264, pp:118-134.
- [13] Newton Spolaor, Everton Alvares Cherman, Maria Carolina Monard and Huei Diana Lee, A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach, *Elsevier*, 2013, vol.292, pp.135-151.
- [14] Ozden Gur Ali and Kubra Yaman, Selecting rows and columns for training support vector regression models with large retail datasets, *European Journal of Operational Research, Elsevier*, 2013, vol.226, pp.471-480.
- [15] M . Ramaswami and R . Bhaskaran, A Study on Feature Selection Techniques in Educational Data Mining, *Journal of Computing*, 2009, vol.1, no.1, pp.7-11.
- [16] Robert Nisbet, John Elder IV and Gary Miner, Chap: Feature Selection, *Handbook of Statistical Analysis and Data Mining Application, Academic Press*, 2009, pp.77- 79.
- [17] Ron Kohavi and George H.John, Wrappers for feature subset selection, *Artificial Intelligence, Elsevier*, 1996, vol.97, pp.273-324.
- [18] Subramanian Appavu Alias Balamurugan and Ramasamy Rajaram, Effective and Efficient Feature Selection for Large-Scale Data Using Bayes Theorem, *International Journal of Automation and Computing*, 2009, vol.06, no.1, pp.62-71.
- [19] Susana M.Vieira, Jo oM.C.Sousa and UzayKaymak Fuzzy criteria for feature Selection, *Elsevier*, 2011, vol.189, pp.1-18.
- [20] Yong Seog Kim, W.Nick Street and Filippo Menczer, Feature Selection in Data Mining, *Data Mining: Opportunities and Challenges*, 2003, pp.80-102.