Analysis of Feature Selection Algorithms on High Dimensional Data

Sowmya Sanagavarapu
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
sowmya.ruby7@gmail.com

Mariam Jamilah
Department of Civil Engineering
College of Engineering Guindy, Anna University
Chennai, India
mariamjamilah24@gmail.com

Barathkumar V
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
barathkumarv98@gmail.com

***Abstract*— Dimensionality of a dataset refers to the number of attributes present in the dataset. At times, the number of attributes is greater than the number of observations, this gives rise to high dimensional data. In high dimensional data, the dimensions are so high that calculations become extremely difficult and this in turn increases the processing and training time. Thus, it is vital to reduce the dimensionality of data [1]. Dimensionality reduction means to simplify the data without affecting data integrity. For this study, we have taken the Dorothea dataset [10] from UC Irvine Machine Learning Repository. Dorothea is a drug discovery dataset. Drugs are organic molecules that bind to a target on a receptor, they are classified as active or inactive based on their ability to bind. New drugs are formed usually by identifying and isolating the receptor to which the chemical compounds have to bind. Then many small molecules are tested for their ability to bind to this receptor. The class label shows whether the molecule will bind to the drug or not. In this paper, we investigate the dimensional reduction achieved by applying three Feature Selection algorithms [2]- Filter, Wrapper and Hybrid with no loss in the integrity of the dataset. We evaluated the accuracy of the obtained data using a C4.5 Classification algorithm [6]. It is used to predict categorical class label of the dataset after training it using the training dataset. The results of each algorithm [1] have been compared and analyzed in order to arrive at the best suited algorithm.**

***Keywords*— classifier, relief filter, hybrid, Las Vegas wrapper, test data, training data**

I. Introduction

A data is said to have high dimensionality when the number of attributes is greater than the number of observations in a dataset. Such data is hard to handle and hence it is important to reduce the dimensionality of the dataset without any drop in the accuracy with which the class label is predicted. If the reduction in the dimensionality of data results in decrease in accuracy, then the decreased accuracy counterweights the reduction. “Curse of dimensionality” [8] means that large data doesn’t necessarily have to have a good accuracy in the prediction of the class label of the dataset. The dataset used in the project is from the UC Irvine Dorothea drug testing dataset [10]. There were a significant number of missing values in the dataset.

A raw dataset has dirty data [3] -that is, the data contains missing values and noise- in it. These anomalies make the data inconsistent, harder to process and give us inaccurate results upon processing. To avoid this, data preprocessing is performed to obtain a clean dataset- a dataset with little noise in it. The obtained dataset with reduced number of features after feature selection gives more accurate results than the original dataset, consequently reducing the run time. On a clean dataset, the various feature selection algorithms are applied to obtain a dataset with reduced number of attributes. To this dimensionally reduced dataset, the classification algorithm is applied to check the accuracy with which the class label of the dataset is predicted.

The dataset used in the project is from the UC Irvine Dorothea drug testing dataset [10]. There were a significant number of missing values in the dataset. The preprocessing techniques were applied to this dataset.

Data processing [7] is first begun by removing the missing values from the dataset. There are a few methods through which this can be achieved. This fully filled dataset is subjected to further processing to smoothen out the noise in the data. Here, the binning method [8] has been employed to smoothen out the noise in the data. The data that results from the preprocessing stage is complete and consistent. Next, the Feature Selection is applied to obtain a reduced feature subset. Here the three methods of feature selection that have been applied are- the Filter Method, the Wrapper Method and the Hybrid Method ([1], [2], [8]).The preprocessing techniques were applied to this dataset. A Classification algorithm [6] is employed to find the accuracy with which the class label is predicted in the dataset. After obtaining each reduced dataset, the accuracy with which the class label of this dataset is predicted is calculated and compared with that of the original dataset.

A Classification algorithm [6] is employed to find the accuracy with which the class label is predicted in the dataset. After obtaining each reduced dataset, the accuracy with which the class label of this dataset is predicted is calculated and compared with that of the original dataset.

The paper is organized as follows. Section II, discusses the preprocessing steps performed on the dataset. In Section III, the Feature Selection algorithms are explained. Section IV, where the classification algorithm has been used. Section V displays the results obtained from each of the Feature Selection Algorithm on the dataset. Finally, we conclude the paper and discuss future work.

II. Data Preprocessing

The dataset chosen initially contained dirty data -meaning incomplete, noisy and inconsistent data. The dataset had missing values: lacking attribute values or certain attributes of interest or/and noisy and inconsistent data: containing errors, discrepancies or outliers.

Missing values maybe due to

- Equipment malfunctioning

- Inconsistent with other recorded data and thus deleted
- Data not entered due to misunderstanding
- Certain data may not be considered important at the time of entry

In case of missing values, we can do any one of the following

1) Manually enter the value

By manually entering the missing values in the data, we are giving way to data discrepancies in the data. This doesn't necessarily ensure a good accuracy of data.

2) Assign a constant value

The global constant of the attribute in the data is chosen to fill the missing values.

3) Eliminate rows with missing values

This method is not preferred as a high number of attributes may be lost thus resulting in a very much reduced feature subset.

4) Fill the missing values with the mean of the column

The mean value of the non-missing values of an attribute is calculated and the missing values are filled with the mean value of that attribute. To find the mean of the attribute, an inbuilt mean function is used.

5) Fill the missing values with the median of the column

The median value of the non-missing values of an attribute is calculated and the missing values are filled with the median value of that attribute. To find the median of the attribute, an inbuilt median function is used.

6) Fill the missing values with the mode of the column

The mode value of the non-missing values of an attribute is calculated and the missing values are filled with the mode value of that attribute. To find the mode of the attribute, a mode function was coded and used to find the mode value of an attribute.

Noisy and inconsistent data [9] maybe due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation

For reducing the noise in the dataset binning method [8] is used. In binning method, the data values are placed into equal width bins. The bins are set to accept data values between two limits. The limit is decided depending upon the range of the values in that attribute.

The limits of the bins are to be set such that the lowest and the highest limit are also taken into consideration. The noisy data is then smoothened by applying the mean, median or boundary of the bins to the data values in the bin.

1. Binning by mean

Here the mean of the bin limits is applied to the data values in the bin to smoothen out the noise.

2. Binning by median

Here the median of the bin limits is applied to the data values in the bin to smoothen out the noise.

3. Binning by boundary

Here the upper or the lower boundary of the bin limits is applied to the data values in the bin to smoothen out the noise.

The missing values in the Dorothea dataset were handled by filling them with binning by median value in the column. This ensured that the extreme noises observed in some features didn't decrease the integrity of the dataset, as it would have happened with using the mean of a feature.

III. Feature Selection Algorithms

Feature Selection [5] is the process of selecting a subset of relevant features for further processing. The main concern when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

A. Filter method

Filter feature selection methods apply a statistical measure to assign a score to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable.

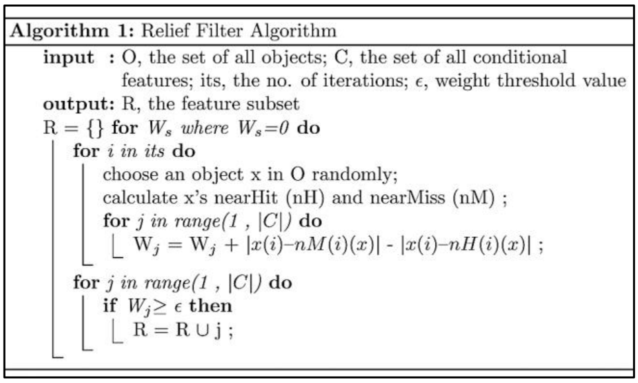
Here, the Relief Filter Algorithm [1] is applied to the dataset. Filter method selects the feature subset on the basis of the intrinsic characters of the data, independent of the machine learning algorithm. Among that, the Relief algorithm is considered one of the most successful algorithms used for assessing the quality of features due to its simplicity and effectiveness. The key idea of Relief Algorithm [1] is to iteratively estimate the feature weights according to their ability to discriminate between neighboring patterns. In this algorithm, a random object is selected from the dataset and the nearest neighboring sample with the same class label (nearHit) and different class label (nearMiss) are identified. The nearest neighboring sample of an object means the object with the greatest number of features having same value. The weight of each feature is updated by:

$$W_j = W_j + |x(i) - nM(i)(x)| - |x(i) - nH(i)(x)| \qquad (1)$$

After running the for loop for the number of iterations, the weight threshold value is calculated by taking the mean of weight of all the features. Then select the features which are having the weight greater than the threshold value.

Clean dataset
Feature Selection
Learn algorithm

Fig. 1. Flow of the Relief Filter Algorithm



```
Algorithm 1: Relief Filter Algorithm
    input  : O, the set of all objects; C, the set of all conditional
             features; its, the no. of iterations; $\epsilon$, weight threshold value
    output: R, the feature subset
    R = {} for $W_s$ where $W_s$=0 do
        for i in its do
            choose an object x in O randomly;
            calculate x's nearHit (nH) and nearMiss (nM) ;
            for j in range(1 , |C|) do
                $W_j = W_j + |x(i)–nM(i)(x)| - |x(i)–nH(i)(x)|$ ;
        for j in range(1 , |C|) do
            if $W_j \geq \epsilon$ then
                R = R ∪ j ;
```

Features of Filter Method:

1) Relief method considers all attributes into consideration.

2) It doesn’t consider the relationship between two attributes.

3) Eliminates attributes by comparing with threshold value.

4) The subset selection is done only a single time.

B. Wrapper Method

Wrapper methods [2] consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model. At times, Filter methods have been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems. Here, the Las Vegas Wrapper Algorithm ([1], [2]) is applied to the dataset.

The Las Vegas Wrapper algorithm uses random subset creation that guarantees given enough time, the optimal solution will be found. It produces intermediate solutions while working towards better ones that result in a lower classification error.

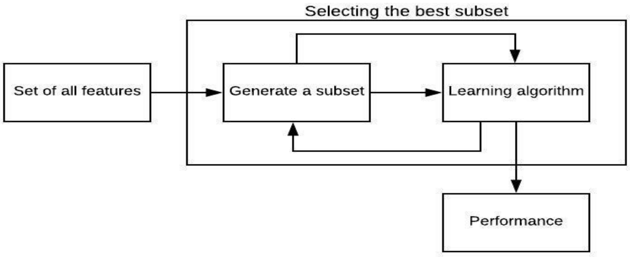
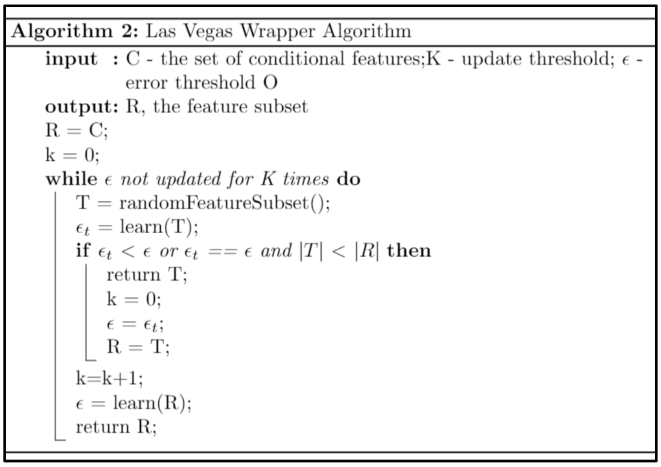


Fig. 2. Flow of the Las Vegas Wrapper Algorithm



```
Algorithm 2: Las Vegas Wrapper Algorithm
    input  : C - the set of conditional features;K - update threshold; $\epsilon$ -
             error threshold O
    output: R, the feature subset
    R = C;
    k = 0;
    while $\epsilon$ not updated for K times do
        T = randomFeatureSubset();
        $\epsilon_t$ = learn(T);
        if $\epsilon_t < \epsilon$ or $\epsilon_t == \epsilon$ and $|T| < |R|$ then
            return T;
            k = 0;
            $\epsilon = \epsilon_t$;
            R = T;
        k=k+1;
        $\epsilon$ = learn(R);
        return R;
```

In this algorithm, initially the full set of conditional features is taken as the best subset. The algorithm then generates a random feature subset and evaluates the error threshold εt using an inductive learning algorithm, here C4.5 [6] is used. It compares εt with ε,

- if $\varepsilon_t < \varepsilon$ or $\varepsilon_t == \varepsilon$ and $|T| < |R|$, then εt becomes the new ε, T becomes the new R and k becomes 0 and then the algorithm continues to generate random subsets.

- if $\varepsilon_t > \varepsilon$ or $\varepsilon_t == \varepsilon$ and $|T| > |R|$, the algorithm continues to generate random subsets until K times.

This algorithm requires two threshold values to be supplied: ε, the classification error threshold and the value K, used to determine when to exit the algorithm due to there being no recent updates to the best subset encountered so far.

Features of Wrapper Method:

1) Not all attributes are considered.

2) A varying number of attributes are randomly chosen and the mean absolute error is compared for the different subsets.

3) The relationship between attributes is considered here.

4) The subset is selected a number of times as per requirement.

C. Hybrid Method

The Hybrid Algorithm [1] is defined as a combination of both Iterative Relief Filter method [2] and Las Vegas Wrapper method [2]. The complete and consistent dataset is considered and the Iterative filter method algorithm is applied to it. With the feature subset obtained thus, we apply the Las Vegas Wrapper method [2] to it to get a further reduced feature subset. Thus, with the now doubly reduced feature set, we test its accuracy by applying the C4.5 Classification algorithm [6] and the results are recorded.

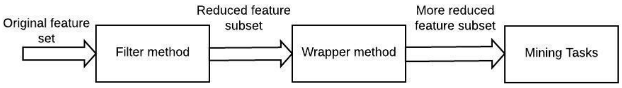


Fig. 3. Flow of the Hybrid Algorithm

IV. Classification Algorithm

A Classifier [6] is a tool in Data Mining that takes a bunch of data representing things we want to classify and attempts to predict which Class Label the test data belongs to. Classification is used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constraints. Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. Several major kinds of classification algorithms include C4.5, ID3, K-nearest neighbor classifier, Naive Bayes, SVM, ANN etc. Here, C4.5 [6] was used as the Classification algorithm. C4.5 is used because of its quick classification and high precision. C4.5 is a Classification algorithm that is used to produce a decision tree which is an expansion of prior ID3 calculation. It enhances the ID3 algorithm. C4.5 creates decision trees from a set of a training data same way as an ID3 algorithm. Decision tree learning creates something similar to a flowchart to classify new data. C4.5 uses greedy (non-backtracking) approach in which decision trees are constructed in top – down recursive divide and conquer manner. C4.5 algorithm is a supervised learning algorithm as it cannot learn on its own. For this it was trained by using the Training dataset. The algorithm analyzes the training set and builds a decision tree and now it uses the decision tree to classify. As the decision tree is being built, the goal of each node is to decide the split attribute (feature) and the split point that best divides the training instances belonging to that leaf.

V. Analysis

The accuracy of the preprocessed high-dimensional dataset was observed to be at 94.25% with 6061 features. The feature algorithms were applied on this dataset and results are analysed.

A. *Iterative Relief Filter Algorithm*

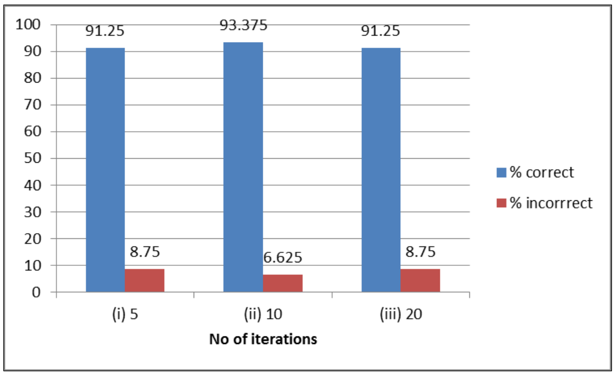


Fig. 4. Iterative Relief Filter Algorithm for varying "its"

From the Fig. 4. - (i) we infer that, after applying the Relief Filter algorithm where the number of iterations is 5, the number of attributes has reduced from 6061 to 852 and the percentage of correctly classified attributes thus obtained is 91.25%. Therefore with 852 attributes we can obtain an accuracy of 91.25%. Hence dimensionality reduction is achieved.

From the Fig. 4. - (ii) we infer that, after applying the Relief Filter algorithm where the number of iterations is 10, the number of attributes has reduced from 6061 to 1112 and the percentage of correctly classified attributes thus obtained is 93.375%. Therefore with 1112 attributes we can obtain an accuracy of 93.375%. Hence dimensionality reduction is achieved.

From the Fig. 4. - (iii) we infer that, after applying the Relief Filter algorithm where the number of iterations is 20, the number of attributes has reduced from 6061 to 1436 and the percentage of correctly classified attributes thus obtained is 91.25%. Therefore with 1436 attributes we can obtain an accuracy of 91.25%. Hence dimensionality reduction is achieved.

B. *Las Vegas Wrapper Algorithm*

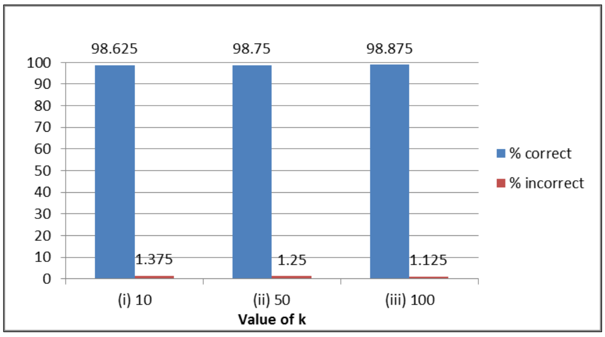


Fig. 5. Las Vegas Wrapper Algorithm for varying "k"

From the Fig. 5. - (i) we infer that, after applying the Las Vegas Wrapper algorithm where K=10, the percentage of correctly classified attributes has increased from 94.25% to 98.625% and the number of attributes has reduced from 6061 to 3034. Therefore with 3034 attributes we can obtain an accuracy of 98.625%which is higher than the accuracy of the clean dataset. Hence dimensionality reduction is achieved.

From the Fig. 5. - (ii) we infer that, after applying the Las Vegas Wrapper algorithm where K=50, the percentage of correctly classified attributes has increased from 94.25% to 98.75% and the number of attributes has reduced from 6061 to 4349. Therefore with 4349 attributes we can obtain an accuracy of 98.75% which is higher than the accuracy of the clean dataset. Hence dimensionality reduction is achieved.

From the Fig. 5. - (iii) we infer that, after applying the Las Vegas Wrapper algorithm where K=100, the percentage of correctly classified attributes has increased from 94.25% to 98.875% and the number of attributes has reduced from 6061 to 3427.Therefore with 3427 attributes we can obtain an accuracy of 98.875% which is higher than the accuracy of the clean dataset. Hence dimensionality reduction is achieved.

C. *Hybrid Algorithm*

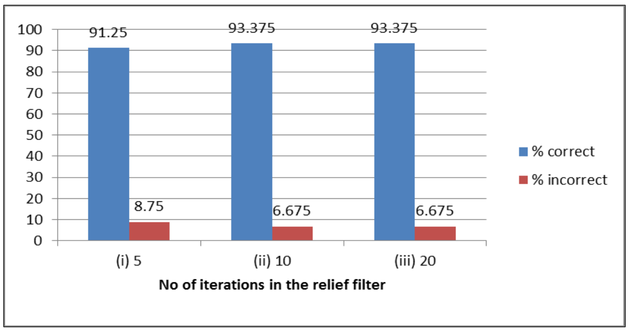


Fig. 6. Hybrid Algorithm for k=10, and varying "its"

From the Fig. 6 - (i) we infer that, after applying the Hybrid algorithm where the number of iterations is 5 and K= 5, the number of attributes has reduced from 6061 to 132 and the percentage of correctly classified attributes thus obtained is 91.25%. Therefore with 132 attributes we can obtain an accuracy of 91.25%. Hence dimensionality reduction is achieved.

From the Fig. 6. - (ii) we infer that, after applying the Hybrid algorithm where the number of iterations is 10 and K= 5, the number of attributes has reduced from 6061 to 1102 and the percentage of correctly classified attributes thus obtained is 93.375%. Therefore with 1102 attributes we can obtain an accuracy of 93.375%. Hence dimensionality reduction is achieved.

From the Fig. 6 - (iii) we infer that, after applying the Hybrid algorithm where the number of iterations is 20 and K= 5, the number of attributes has reduced from 6061 to 1436 and the percentage of correctly classified attributes thus obtained is 95.375%. Therefore with 1436 attributes we can obtain an accuracy of 95.375%. Hence dimensionality reduction is achieved.

VI. Conclusion

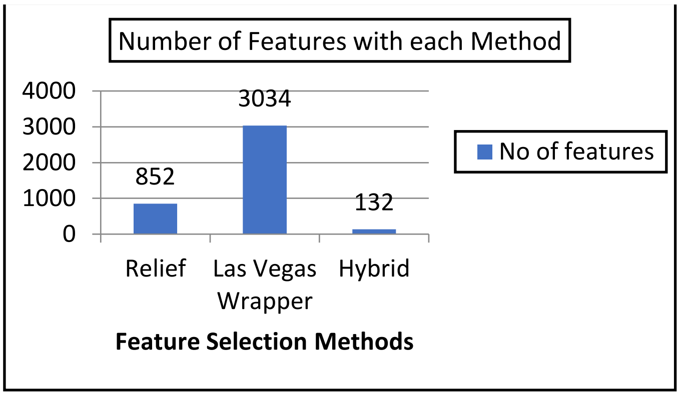


Fig.7. Comparing no. of reduced features obtained after applying the algorithms

From the Fig. 7. it is observed that although the Iterative Relief Algorithm achieves a high reduction in the number of features considered from 6061 to 852, it also results in a slight drop in the accuracy of the reduced dataset to 91.25% from 94.25%. The Hybrid algorithm achieves dimensionality reduction from 6061 features at accuracy of 94.25% to 136 with accuracy of 91.25%. The Las Vegas Wrapper Algorithm performs by reducing the number of the features to almost half with an increase in the accuracy of the dataset from 94.25% to 98.625%.

VII. Future Works

The algorithms involved in this research and the results obtained can be used for various applications. These algorithms can be applied to reduce the dimensionality of dynamic data, consequently reducing the processing/run time. They can also aid in the reduction of bands in satellite imagery obtained through multispectral or hyperspectral sensors. In combination with these algorithms, different classifiers can be used to test the accuracy for different applications. The number of iterations for each algorithm can be varied and based on the resulting accuracy further analysis can be made.

Acknowledgement

This research was supported by College of Engineering, Guindy, Anna University, India. We are grateful to Dr. T. V. Geetha, Dean of College of Engineering Guindy for giving us this golden opportunity. Our wholehearted gratitude to Dr. S. Valli, Head of Department of Computer Science, and Engineering, College of Engineering Guindy for providing us with the lab facilities and the necessary essentials. We thank our mentor Dr. A P Shanthi, Professor, Department of Computer Science and Engineering, College of Engineering Guindy who provided insight and expertise that greatly assisted the research. We thank Dr. T. Raghuveera, Associate Professor, Department of Computer Science and Engineering, College of Engineering Guindy and Dr. D Manjula for their undying support and encouragement throughout the project, and Ms. Susi E, Research Scholar, Department of Computer Science and Engineering, College of Engineering Guindy for her guidance and constant supervision in completing the project successfully. We take full responsibility in case of any errors and this should not tarnish the reputation of these esteemed persons.

References

[1] K.Sutha and Dr.J.Jebamalar Tamilselvi, " A Review of Feature Selection Algorithms for Data Mining Techniques " Vol. 7 No.6 Jun 2015

[2] Richard Jensen and Qiang Shen, “Computational Intelligence and Feature Selection Rough and Fuzzy Approaches” Aberystwyth University

[3] Preeti Patidar and Anshu Tiwari, “Handling Missing Value in Decision Tree Algorithm” International Journal of Computer Applications (0975 – 8887) Volume 70– No.13, May 2013

[4] WANG Hongwei, “A Method of Feature Selection for Continuous Features base on Similarity Degrees of Interval Numbers” College of Information Science and Technology, Bohai University, Jinzhou, China

[5] S. Visalakshi and V. Radha, “A Literature Review of Feature Selection Techniques and Applications Review of Feature Selection in Data Mining” IEEE International Conference on Computational Intelligence and Computing Research 2014

[6] Hehui Qian, Zhiwei Qiu, “Feature Selection using C4.5 Algorithm for Electricity Price Prediction” International Conference on Machine Learning and Cybernetics, Lanzhou 2014

[7] Surekha Samsani, “An RST based Efficient Preprocessing Technique for Handling Inconsistent Data” IEEE International Conference on Computational Intelligence and Computing Research

[8] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques” Morgan Kaufmann Publishers, Third Edition, 2011

[9] Qing Ang, Weidong Wang, Zhiwen Liu and Kaiyuan Li, “Explored Research on Data Preprocessing and Mining Technology for Clinical Data Applications” 2nd IEEE International Conference on Information Management and Engineering 2010, Chengdu, China

[10] UC Irvine Machine Learning, Center for Machine Learning and Intelligent Systems. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/dorothea

[11] K. Naidu, A. Dhenge and K. Wankhade, "Feature Selection Algorithm for Improving the Performance of Classification: A Survey," 2014 Fourth International Conference on Communication Systems and Network Technologies, Bhopal, 2014, pp. 468-471.

[12] A. S. Abdullah, C. Ramya, V. Priyadharsini, C. Reshma and S. Selvakumar, "A survey on evolutionary techniques for feature selection," 2017 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2017, pp. 58-62.

[13] B. C. Santos, M. W. Rodrigues, L. N. Cristiano Pinto, C. N. Nobre and L. E. Zárate, "Feature selection with genetic algorithm for protein function prediction," 2019 IEEE International

Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 2434-2439.

[14] B. C. Santos, M. W. Rodrigues, L. N. Cristiano Pinto, C. N. Nobre and L. E. Zárate, "Feature selection with genetic algorithm for protein function prediction," 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 2434-2439.

[15] W. Pearson, C. T. Tran, M. Zhang and B. Xue, "Multi-Round Random Subspace Feature Selection for Incomplete Gene Expression Data," 2019 IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand, 2019, pp. 2544-2551.

[16] T. Chandak, C. Ghorpade and S. Shukla, "Effective Analysis of Feature Selection Algorithms for Network based Intrusion Detection System," 2019 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2019, pp. 1-5.

[17] F. Koumi, M. Aldasht and H. Tamimi, "Efficient Feature Selection using Particle Swarm Optimization: A hybrid filters-wrapper Approach," 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2019, pp. 122-127.