Feature Selection Algorithms for Data Mining Classification: A Survey

N. Krishnaveni[1]* and V. Radha[2]

[1]Department of Information Technology, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore – 641043, Tamil Nadu, India; Krishnaveni.adu@gmail.com
[2]Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore – 641043, Tamil Nadu, India; radharesearch@yahoo.com

Abstract

Objectives: This study summarizes the feature selection process, its importance, different types of feature selection algorithms such as Filter, Wrapper and Hybrid. Moreover, it analyses some of the existing popular feature selection algorithms through a literature survey and also addresses the strengths and challenges of those algorithms. **Methods/ Statistical Analysis:** When there are many methods are in hand to be obtained, then Review of Literature is the best approach to learn about existing methods before going for a new model. **Findings**: Feature selection is a predominant preprocessing strategy in Data Mining, which helps in advancing the performance of mining, by selecting only the relevant features and avoiding the redundant features. There are plenty Feature Selection algorithms developed and used by most researchers. But still it is an emerging area in machine learning to be focused for data mining and analysis process for pattern recognition. Many feature selection algorithms confront severe challenges in terms of effectiveness and efficiency, because of recent increase in data variety and velocity. Different types of feature selection algorithms are available in literature such as Filter based, Wrapper based and Hybrid algorithms. Moreover, analyses some of the existing popular feature selection algorithms through a literature survey, also addresses the strengths and challenges of those algorithms. **Application/Improvements:** There is a need for an effective unified framework, which should provide feature selection for any size of dataset without noisy data, low computational complexity and highest accuracy.

Keywords: Classification, Data Mining, Feature Selection, Filter, Hybrid, Wrapper

1. Introduction

In recent years, data stored and collected for different purposes are very large. Such data set may consist of millions of records and each of which may be represented by hundreds or thousands of features. Nowadays, dataset became big data with extremely more number of features. When data mining and machine learning algorithms are applied on high-dimensional data, dimensionality is the critical issue that should be handled[1,2]. It refers to the phenomenon that the data become sparser in high-dimensional space, adversely affecting algorithms designed for low-dimensional space. Also, with a large number of features, learning models tend to over fit, this leads to performance degradation on unobserved data. Data of high dimensionality can significantly increase the memory storage requirements and computational costs for data analytics.

---

*Author for correspondence

Manual management of these datasets is impractical. Therefore, data mining and machine learning techniques were developed to automatically discover knowledge and recognize patterns from these data. However, generally more noise is associated with these collected data. There are many reasons causing noise in these data, among which imperfection in the technologies that collected the data and the source of the data itself are two major reasons.

The individual property for the data analysis, which is considered, is the feature. A set of features are used for performing classification in any machine learning strategies. Previously those applications were using hundreds or thousands of features for analysis process.

Many of the features in such data set contain useful information for understanding the data, relevant to the problem, but it also contains large amount of irrelevant features, and redundant features. This leads to reducing the learning performance and computational efficiency[1,3]. The person should have profound learning experience in the problem field to decide all those features to be utilized to develop a classifier from the existing huge number of variables.

The features, which are most applicable to the problem, can be selected automatically. The constructive information, which is needed, should not be vanished during subset selection. This process is called feature selection, which has other names such as variable selection and attributes selection.

This preprocessing step reduces the dimensionality of dataset before applying the data mining process[1]. It can be useful for any data mining process like classification, clustering, association rule mining. It can be a selection of attributes by selecting a subset of relevant features for using model construction automatically[2].

Dimensionality reduction is another most popular technique, which is also helpful in feature selection through noise removal. Noise means irrelevant or redundant data. Feature selection and dimensionality reduction are different in many aspects but both methods are tending to dropping irrelevant attributes in a dataset, feature selection process decides whether a particular attribute is to be used or not without modification, whereas dimensionality reduction checks the attributes by assessing different aggregations of the available list of attributes. Singular Value Decomposition, Principal Component Analysis and Sammon's Mapping are some of the example for Dimensionality reduction. In fact, feature selection process performs screening process among the available features through a filter, so that the unwanted features that are removed.

Transformation is not applied during feature selection; meanwhile the original feature set is maintained without changing its meaning. Hence feature selection enhances the readability. This property has its significance in many practical applications such as finding relevant genes to a specific disease and building a sentiment lexicon for sentiment analysis. For the classification problem, feature selection aims to select subset of highly discriminated features. In other words, features are selected that are accomplished to selective samples from classes of diverse. For the problem of feature selection for classification, due to the availability of label information, the relevance of features is assessed as the capability.

The following issues in classification are solved by Feature Selection process.

- Predictive models should be accurate and feature selection methods have to help in this aspect. If a small number and only the feature which is required for predictions are provided to the model, then good accuracy can be expected.

- Accuracy in the classifier is the most important aspect. Feature Selection contributes a vital role in accuracy through identifying and removing irrelevant and redundant features.

- Simple models are easy to adopt and explain and if the numbers of features are less, then it can be adoptable in an uncomplicated way.

- Actually, a lot of feature selection methods and algorithms are available in literature. Datasets also

contain different kinds of features with high or very high dimension variables.

- If the number of variables is reduced and redundant or irrelevant variables are removed, then the computational time is reduced and the prediction performance is enhanced.

- In Pattern recognition or machine learning applications, the feature selection algorithms can provide a refined inception on the data.

This study tries to sketch out the critique feature selection practices from the composition of several approaches. A large range of machine learning applications can be used with different types of feature selection algorithms, which includes Filter based methods, Wrapper Based Methods and Hybrid or Embedded methods. The goal of this study is to provide a comprehensive idea with regards to feature selection.

The reasons to use feature selection are:

- If the model is provided with the right set of variables, it will improve the accuracy.

- The performance of machine learning algorithms in the classifier can be faster.

- The model complexity is reduced and interpretation is also easy.

- Over fitting can be reduced.

2. Survey of Literature

Methods for analyzing the redundancy and relevance of features as unsupervised and multivariate filter- based feature selection methods were proposed. The features are estimated using ant colony optimization algorithm. The accuracy of the methods is measured with the novel heuristic information measure by considering the similarity between subsets of features[3].

A recommender system for gait biometric representation used Robita Gait system. A new feature selection algorithm called Incremental Feature Selection (IFS) with Analysis of Variance (ANOVA) was proposed. Statistical significance is increased when applied with a classifier fusion model[4].

The challenges of feature selection for big data analytics are comprehended. As the size of the data grows rapidly, the feature selection algorithm also needs to be technically improved for reducing redundant data[5].

A comparative study on four different types of feature selection algorithms were provided[6]. Decision trees, entropy measure for ranking features, estimation of distribution algorithms, and the bootstrapping algorithm were compared and found each algorithm has its own merits and demerits. Also proved that the elimination of noise is the most important consideration in classification process.

A fuzzy rough dependency is used as a criterion for feature selection and introduced a new fuzzy rough set model to guarantee that the membership degree of a fragment of the same type influences the highest amount[7]. It also effectively prevents samples from being misclassified. A greedy forward algorithm for feature selection is also used.

A fast sequential feature selection algorithm is proposed using affinity propagation clustering[8]. This algorithm divides the dataset into many clusters and then sequential feature selection is applied to each cluster separately. All the results are collected together for the feature selection process. This algorithm works faster and provides high accuracy.

A novel hybrid feature selection algorithm is proposed[9] using filter based rough conditional mutual information and wrapper based naïve Bayesian classifier. This reduced the computing complexity and number of irrelevant features also reduced.

A new feature selection algorithm called class dependent density based feature elimination for binary datasets using feature ranking approach called diff-criterion is proposed[10]. It outperforms in many aspects such as dimensionality and computational complexity reduction.

A feature selection algorithm called hybridization of Genetic Algorithm and Particle Swarm Optimization through integrating the velocity and update rules such as selection rules, crossover and mutation. It outperforms with limited sample size and automatically selects the features. It is dataset distribution independent[11].

An improved Ant Colony Optimization algorithm using wrapper based method for classification is proposed. It outperforms in accuracy when used with a biomedical dataset. The alpha band was found that it has more features than beta band[12].

A new algorithm using a wrapper based feature selection method by using a hybrid search method and particle swarm optimization and local search is proposed. It works with chaos interia weight and local search to search among 2^d possible cases[13].

Elitism based Multi-Objective Differential Evolution algorithm for feature selection (FAEMODE) based on Filter approach is proposed. A contemporary objective formula is generated by considering the linear and nonlinear dependencies for feature selection process. It gives a powerful result when compared with seven filter approaches[14].

A wrapper based feature selection algorithm using Harmony Search for Holistic Bangla word recognition is proposed. Multi-Layer Perception was used with HS to improve the accuracy. This algorithm is compared with Particle Swarm Optimization and Genetic Algorithms to ensure classifier accuracy[15].

A hybrid filter-wrapper feature selection algorithm for short term load forecasting is proposed. Partial mutual information from the filter approach and a wrapper based firefly algorithm were jointly used to extract relevant features. This algorithm is implemented in well-established support vector regression. It performs more efficiently compared with other algorithms[16].

A combination of EMD–LDA–PNN–SFAM (Empirical mode Decomposition-Linear Discriminant Analysis-Probabilistic Neural Network-Simplified Fuzzy Adaptive Resonance Theory Map) algorithm for feature selection process is proposed and it outperforms in accuracy. J-Measure is calculated. Feature reduction is done using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The algorithm is tested with online diagnosis dataset[17].

An optimization technique for ensemble systems using filter based approach is proposed. It is applied with mono and bi-objective versions using Particle Swarm Optimization, Ant Colony Optimization, and Genetic Algorithms. Bi-Objective version outperforms in this method[18].

3. Feature Selection Process

Feature selection process involves four important steps such as feature subset generation, subset evaluation, stopping criterion and result validation. The feature subset generation helps in the candidate selection subset for
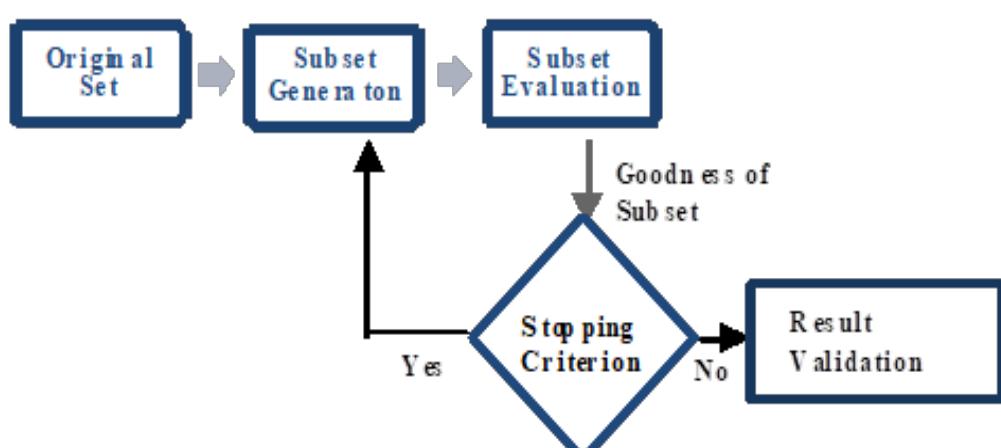

Figure 1. General framework for feature selection.

evaluation. Actually it follows a heuristic approach. The searching approaches it follows to produce subsets are progressive, comprehensive and random search of features. The quality of the subset generated is assessed with an evaluation criterion. The new subset is compared with the previous subset and found the best one. The first-rated subset is further used for next comparison. This comparison process is repeated till the stopping criterion is reached and best subset is generated. The final best subset is further validated by different tests or with a prior knowledge[19]. Figure 1 illustrates the feature selection process.

3.1 Feature Selection Algorithms

Feature selection algorithms are broadly classified as Filter based Feature Selection, Wrapper Based Feature Selection and Hybrid Feature Selection methods[19,20]. Any how it is also categorized into four main groups: similarity-based, information-theoretical-based, sparse-learning-based, and statistical-based methods when considering the type of data[21].

3.1.1 Filter Methods

In general, the choice of features is sovereign of any machine learning algorithms. Different types of statistical tests are carried out and the scores are generated. The correlation between these scores forms a groundwork for Filter based feature selection. The correlation is a subjective term here. The filter methods do not remove multicollinearity. That is two or more predictors are highly correlated, which leads to statistical inference.

A statistical measure is applied to allocate a scoring to every feature. Either the selected feature should be kept or removed will be decided through this ranking. The methods are often referring to a single characteristic or attribute independently even if the variable is dependent on each other.

The above Figure 2 depicts the filter based feature selection algorithm steps.

The coefficients such as Pearson's Correlation, Mutual Information, Kendall Correlation, Spearman Correlation, Linear Discriminant Analysis, Chi-Square test, Fisher Score, Count based and ANOVA (Analysis of Variance) are some of the methods used in filter based approach.

Pearson Correlation: Pearson's correlation is statistic or coefficient used to find out the strength of the correlation between two variables.

Mutual Information: It helps to reduce uncertainty about the value of another variable. Different dimensions of dataset the reciprocal data in datasets are maximized between the targeted variables and joint distribution.

Kendall Correlation: It is a grading technique used to find out the association. The ranking for ordinal variables are calculated such as different rankings and ranking of different variables are considered for finding relationships.

Spearman Correlation: The rate of constant relationship among two variables is represented using Spearman Correlation coefficient.
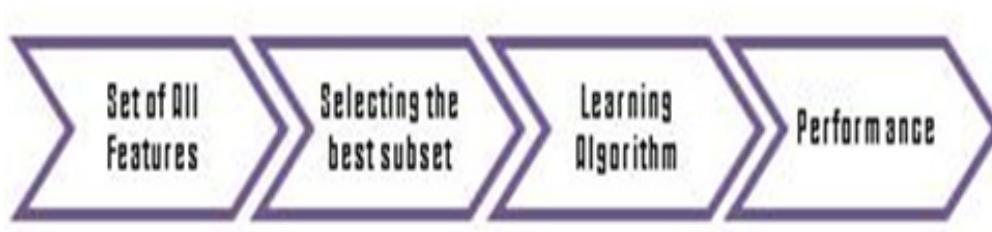
Figure 2. Filter based feature selection process.

Linear Discriminant Analysis (LDA): Closely related to ANOVA and Regression Analysis. It works in Linear model and more suitable for the classification classes more than two.

Chi Square Test: The distance between the actual results and expected results are compared with a statistical technique called chi-square test.

Fisher Score: The differences between the expected and observed values are found through fisher score. The information is maximized when the difference is minimized.

Count Based: The most significant information is not presented in all columns of data. The weight of the values from each column is counted to get an idea about the data.

ANOVA: Analysis of variance (ANOVA) is a group of statistical models to test the significance between means.

3.1.2 Wrapper Methods

In wrapper methods, there are subset of features with different combinations are developed and used in a model for training. Based on the inferences that drawn from the previous model, it will be decided whether to add or remove features from the selected subset through evaluation. The wrapper based feature selection methods are very expensive computationally. The basic idea is like a searching strategy. Model accuracy is evaluated with the combination of features with score assigned through a predictive model. Figure 3 explains about the wrapper based feature selection process.

<insert figure 3 here>

There are many wrappers based feature selection methods are used widely. Forward feature selection algorithm, Backward Feature Selection Algorithm and Recursive feature selection algorithm are some of the common examples.

Forward Feature Selection Algorithm: It is basically an iterative method started with zero features. In every iteration, a new feature is added to the model and substantiated to verify whether it provides improvement in the performance of model.

Backward Feature Elimination Algorithm: It is a reverse model of Forward Feature Selection. This model starts with all the features. It is also an iterative model and removes the least significant feature in each iteration. The performance of the model is measured and features are removed until no improvement is observed.
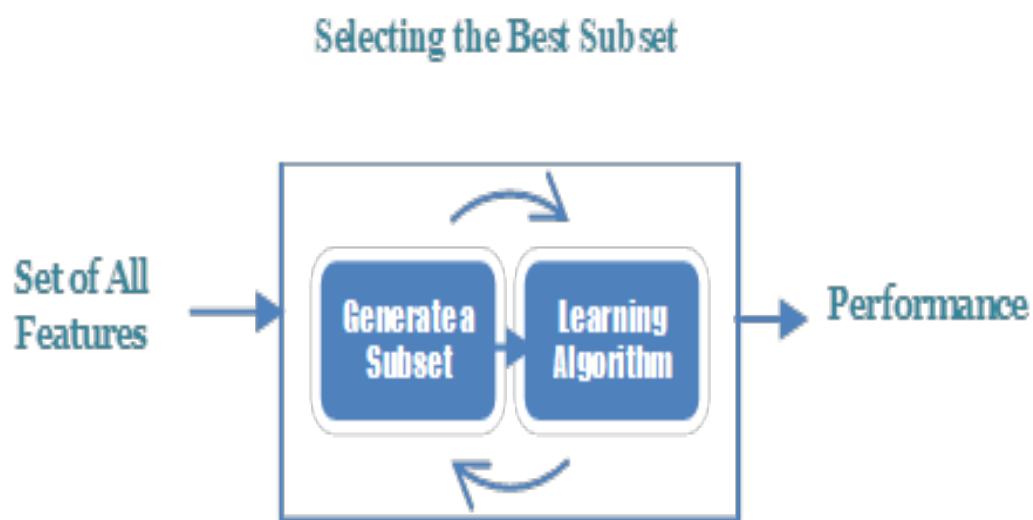

Figure 3. Wrapper based feature selection process.

Recursive Feature elimination: It is a type of greedy optimization technique. The most useful feature subset is found through this method. At each iteration, a new model with different feature subset is created and found the best and worst performing features. The model construction process is continued until all features are run through. The elimination of features is done based on its ranking.

Adding and removing features is done through forward or backward process. The heuristic searching techniques are adopted for searching.

Difference between Filter and Wrapper methods

Following are some of the differences between the filter and wrapper feature selection methods.

- The significance of features is measured by correlation variable in filter methods. Wrapper methods measure the significance of a feature through training a model.

- Wrapper methods are slower than filter methods because it performs model training for feature measurement. So that the computational cost of wrapper method is high compared to Filter methods.

- The subsets of features are weighed by statistical tools in filter methods. Cross validation methods are used in wrapper methods.

- The wrapper based methods rely on finding best subsets than filter methods.

- The wrapper methods providing the models with more over fitting compared to filter models.

3.1.3 Embedded or Hybrid Methods

The best characteristics of both the filter and wrapper based methods are combined to form the embedded or hybrid models. Its own built-in feature selection methods are used for implementation of algorithms. Figure 4 depicts the process of hybrid feature selection process.

The learning process in embedded models enables to find the best accuracy level during feature selection. The regularization method is one of the common embedded type feature selection. The other name of regularization methods are penalization methods. Additional con-
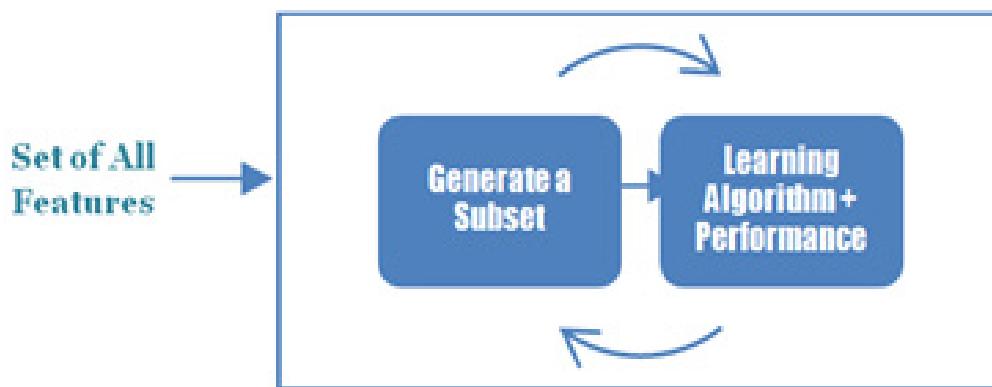
Figure 4. Hybrid feature selection process.

Table 1.   Comparison of feature selection algorithms

Algorithm	Type	Factors/ Approaches Used	Result/Inference	Limitation/s
Unsupervised and multivariate filter-based feature selection method[3]	Filter Based	Ant Colony Optimization	The performance of the algorithm is improved.	New State Transmission Rule to control the randomness can be developed.
Incremental Feature Selection(IFS) with Analysis of Variance(ANOVA)	Filter	ANOVA	Statistical Significance is increased	Other Validations can be done
Affinity Propagation-Sequential Feature Selection Algorithm[8]	Wrapper Based	Cluster Based	Faster for high dimensional data	Accuracy is comparable
Fuzzy Rough Set Feature selection algorithm	Filter	Fuzzy Based\ Greedy Forward Algorithm	Works better in large degree of overlapping datasets	Does not work for small stack datasets
Novel Hybrid Feature Selection Algorithm	Hybrid	Rough Conditional Mutual Information. Bayesian Classifier	Computational complexity is reduced Irrelevant Features are reduced. Improves prediction accuracy	Accuracy can be improved
Class dependent density based feature elimination	Filter	Feature Ranking Feature Elimination Selection	Works better for High dimensional binary data. Works along with the classifier.	Other data types can be verified
Hybridization of Genetic Algorithm and Particle Swarm Optimization	Hybrid	Genetic Algorithm Particle Swarm Optimization	Automatic Feature Selection with High Accuracy with small number of Samples in High Dimensional Dataset.	SVM can be improved Verified with parameter initialization
Improved Ant Colony Optimization-SVM[12]	Wrapper	Ant Colony Optimization. Support VectorMachines	Accuracy of FS is improved. Found that more relevant features are in alpha band	Data Scalability is not verified. Might consider Beta band also.

Table 1 Continued

Choas Binary Particle Swarm Optimization with Local Search[13]	Wrapper	Particle Swarm Optimization. Local Search	works with chaos interia weight. Searches among 2nd possible cases with local search	Filter based approach Using Global Searches
Filter Approach using Elitism based Multi-objective Differential Evolution algorithm for feature selection (FAEM ODE) [14]	Filter Based	Differential Evolution. Multi objective Optimization	Linear and Nonlinear dependency were considered	Other parameters also to be considered
Harmony Search(HS) for Word Recognition[15]	Wrapper Based	Harmony Search. Multi Layer Perception Classifier	Classifier accuracy is good compared with PSO and GA. Both local and global search were used	Automatic feature selection using stopping criteria. Scheme can be formulated to dynamically identify HS
Hybrid Filter-Wrapper feature selection for short term load forecasting[16]	Hybrid	Filter based Partial Mutual Information. Wrapper based firefly algorithm	Reduced the redundant features without degrading the forecasting accuracy.	Invest some of the exogenious variables and lagged variables. More extensive comparison
Combination of EMD–LDA–PNN–SFAM[17]	Filter	Empirical mode Decomposition Linear Discriminate Analysis Probabilistic Neural Network Simplified Fuzzy Adaptive Resonance Theory Map	J-Measure is improved. Real data set is used and high classification accuracy is attained. Optimal separation of features in different classes. Better categorization	Variable operating conditions are of speed and charge can be considered.
Optimization techniques for ensemble systems[18]	Filter	Particle Swarm Optimization(PSO) Ant Colony Optimization Genetic Algorithms	Compared with Mono and Bi-Objective versions PSO provides better accuracy in both. Found that Bi-Objective works better.	Other Optimization techniques Evaluation criteria can be improved.

straints like regression algorithm are introduced into the optimization of a predictive algorithm in order to create a model with fewer coefficients to achieve lower complexity. LASSO regression and RIDGE regression are some of the famous regression methods which reduces over fitting through inherent correction. Regularized trees, Random multinomial logit and Memetic algorithm are some of the other examples.

4. Comparison of Feature Selection Algorithms

The high computation efficacy and generality are the benefits of Filter based methods. Wrapper based method guarantees better results, but it is computationally expensive for large dataset.

The pros of both the methods are obtained through embedded or hybrid methods. Anyhow all these methods have been widely used by many researchers for the classification problems. If the dimensionality of a dataset is different, same feature selection algorithm may not be suited. So, new approaches of Feature Selection Algorithms are always in need. Table 1 summarizes some of the feature selection algorithms with all the three types such as Filter based, Wrapper based and Hybrid. Each algorithm has its own merits and demerits.

5. Conclusion

There are many feature selection algorithms. Each algorithm selects only the features without considering computational redundancy. The performance and accuracy is not considered in some algorithms. The presence of noisy data is not taken into account when selecting features in some algorithms. The computational time is increased and learning process will become insignificant. Filter based method practices the entire training data when creating a subset. Filter methods can be applied to large datasets with voluminous features as it works faster. But it does not reflect in better accuracy. Wrapper methods select best features with high accuracy. But, the computational cost is large. Some hybrid methods tried

to solve the issues that both the methods have. From this survey it is clear that there is a need for an effective unified framework, which should provide feature selection for any size of dataset without noisy data, low computational complexity and highest accuracy.

6. References

1. Chandrashekar G, Sahin F. A survey on feature selection methods. Computers and Electrical Engineering. 2014; 40(1):16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
2. Jantawan B, Tsai CF. A comparison of filter and wrapper approaches with data mining techniques for categorical variables selection. International Journal of Innovative Research in Computer and Communication Engineering. 2014; 2(6):4501–8.
3. Tabakhi S, Moradi P. Relevance–redundancy feature selection based on ant colony optimization. Pattern Recognition. 2015; 48(9):2798–811. https://doi.org/10.1016/j.patcog.2015.03.020.
4. Semwal VB, Singha J, Sharma PK, Chauhan A, Behera B. An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification. Multimedia tools and applications. 2017; 76(22):24457–75. https://doi.org/10.1007/s11042-016-4110-y.
5. Li J, Liu H. Challenges of feature selection for big data analytics. IEEE Intelligent Systems. 2017; 32(2):9–15. https://doi.org/10.1109/MIS.2017.38.
6. Garcia VH, Rodriguez JR, Usaquén MA. A comparative study between feature selection algorithms. International Conference on Data Mining and Big Data; 2018. p. 65–76. https://doi.org/10.1007/978-3-319-93803-5_7.
7. Wang C, Qi Y, Shao M, Hu Q, Chen D, Qian Y, Lin Y. A fitting model for feature selection with fuzzy rough sets. IEEE Transactions on Fuzzy Systems. 2017; 25(4):741–53. https://doi.org/10.1109/TFUZZ.2016.2574918.
8. Zhu K, Yang J. A cluster-based sequential feature selection algorithm. Ninth International Conference on Natural Computation (ICNC); 2013. p. 848–52. https://doi.org/10.1109/ICNC.2013.6818094. PMid:23763950 .
9. Zeng Z, Zhang H, Zhang R, Zhang Y. A hybrid feature selection method based on rough conditional mutual information and Naive Bayesian classifier. ISRN Applied Mathematics; 2014. p. 1–11.
10. Javed K, Babri HA, Saeed M. Feature selection based on class-dependent densities for high-dimensional binary data. IEEE Transactions on Knowledge and Data Engineering. 2012; 24(3):465–77. https://doi.org/10.1109/TKDE.2010.263.

11. Ghamisi P, Benediktsson JA. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. IEEE Geoscience and Remote Sensing Letters. 2015; 12(2):309–13. https://doi.org/10.1109/LGRS.2014.2337320.

12. Erguzel TT, Tas C, Cebi M. A wrapper-based approach for feature selection and classification of major depressive disorder–bipolar disorders. Computers in biology and medicine. 2015; 1(64):127–37. https://doi.org/10.1016/j.compbiomed.2015.06.021. PMid:26164033 .

13. Javidi MM, Emami N. A hybrid search method of wrapper feature selection by chaos particle swarm optimization and local search. Turkish Journal of Electrical Engineering and Computer Sciences. 2016; 24(5):3852–61. https://doi.org/10.3906/elk-1404-220.

14. Elitism based multi-objective differential evolution for feature selection: A filter approach with an efficient redundancy measure [Internet]. [cited 2017 Aug 15]. Available from: https://www.sciencedirect.com/science/article/pii/S1319157817301003.

15. Das S, Singh PK, Bhowmik S, Sarkar R, Nasipuri M. A harmony search based wrapper feature selection method for holistic bangla word recognition. Procedia Computer Science. 2016; 89:395–403. https://doi.org/10.1016/j.procs.2016.06.087.

16. Hu Z, Bao Y, Xiong T, Chiong R. Hybrid filter–wrapper feature selection for short-term load forecasting. Engineering Applications of Artificial Intelligence. 2015; 40:17–27. https://doi.org/10.1016/j.engappai.2014.12.014.

17. Ali JB, Saidi L, Mouelhi A, Chebel-Morello B, Fnaiech F. Linear feature selection and classification using PNN and SFAM neural networks for a nearly online diagnosis of bearing naturally progressing degradations. Engineering Applications of Artificial Intelligence. 2015; 42:67–81. https://doi.org/10.1016/j.engappai.2015.03.013.

18. Santana LE, de Paula Canuto AM. Filter-based optimization techniques for selection of feature subsets in ensemble systems. Expert Systems with Applications. 2014; 41(4):1622–31. https://doi.org/10.1016/j.eswa.2013.08.059.

19. Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data classification: Algorithms and applications; 2014. p. 1–33.

20. A survey of feature selection and feature extraction techniques in machine learning [Internet]. [cited 2014 Oct 09]. Available from: https://ieeexplore.ieee.org/document/6918213.

21. Feature selection: A data perspective [Internet]. [cited 2018 Aug 26]. Available from: https://arxiv.org/abs/1601.07996.