Research Article

Feature Selection: An Assessment of Some Evolving Methodologies

A. Abdul Rasheed[1]*

[1]Professor, Department of Computer Applications, CMR Institute of Technology, Bangalore – 560 037.
[1]*profaar@gmail.com

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Feature selection has predominant importance in various kinds of applications. However, it is still considered as a cumbersome process to identify the vital features among the available set for the problem taken for study. The researchers proposed wide variety of techniques over the period of time which concentrate on its own. Some of the existing familiar methods include Particle Swarm Optimisation (PSO), Genetic Algorithm (GA) and Simulated Annealing (SA). While some of the methods are existing, the emerging methods provide promising results compared with them. This article analyses such methods like LASSO, Boruta, Recursive Feature Elimination (RFE), Regularised Random Forest (RRF) and DALEX. The dataset of variant sizes is considered to assess the importance of feature selection out of the available features. The results are also discussed from the obtained features and the selected features with respect to the method chosen for study.

Keywords: Feature selection, LASSO, Boruta, Recursive Feature Elimination, Regularised Random Forest

1. Introduction

In a predictive model, feature selection is considered as a process of selecting or choosing or reducing the number of attributes. The attributes are synonymously called as features, or input variables. The requirement of reducing the number of features is to reduce the computational cost.

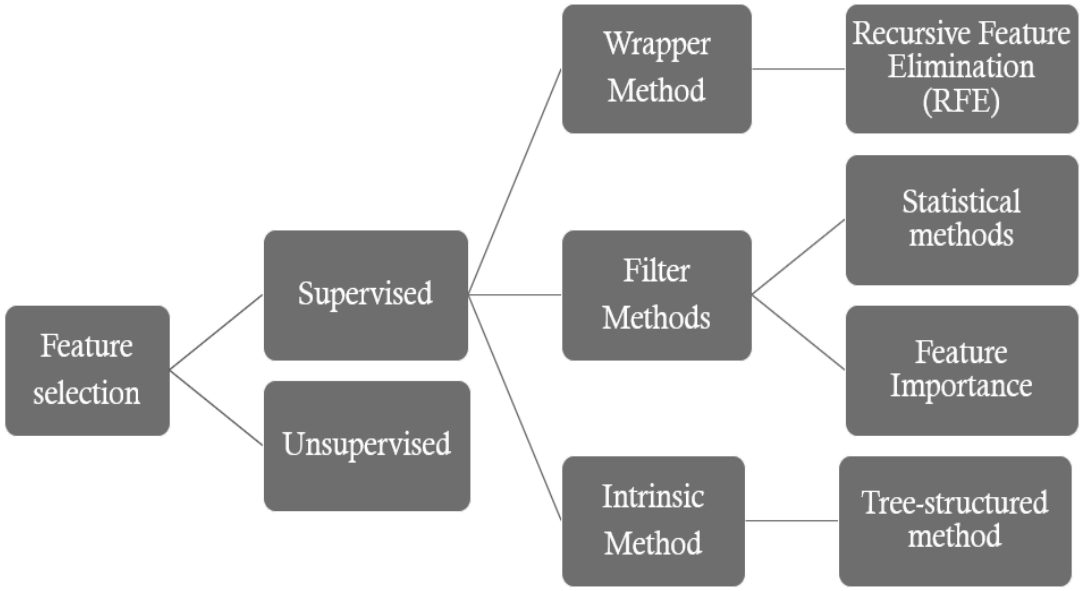

Figure 1. Taxonomy of feature selection methods

The broad spectrum of feature selection methods is the classification as in figure-1. The unsupervised method does not use the target variable hence it is able to remove the redundant variables. In contrast with unsupervised method, the supervised method uses the target variable thereby removing the irrelevant variables. The supervised method has three further classification. (i) The wrapper method uses the subset of features, then decides either to add or to remove the features. Example of such methods include forward selection, backward elimination. (ii) The filter methods use statistical tests to find the correlation between the variables. Some of the statistical tests like Chi-Square, Pearson and ANOVA falls into this category. (iii) Intrinsic methods or embedded methods have in-built feature selection method. The Least Absolute Shrinkage and Selection Operator (LASSO) is a familiar method under this category.

2. Related Works

This section describes the works carried out by the researchers over a period of time.

Category	Citation	Observation
Survey / Review related works	[1]	The big data era has voluminous of data. It is required to find the necessary data for the research problem. This paper is an exhaustive survey of methods used for feature selection. It also discusses about the future direction in this research.
	[2]	This paper summarises the earlier work carried out by the researchers to select and extract the features which are applicable for text classification problem.
	[3]	Mutual information (MI) measures the presence or absence of the amount of term used to make classification (c) correctly over the number of terms (t) available. This paper is a comprehension about how MI can be used in different applications.
Methodology related works	[4]	It combines the filter-based feature selection method with wrapper-based method. In support of evaluating the combination of methods, the paper used two benchmarked datasets. The performance is also discussed in the paper.
	[5]	Unsupervised machine learning is a classification model. This paper used the strategy of masking the unwanted or irrelevant feature and thereby choosing the required features which are unmasked.
	[6]	Feature ranking score strategy for the global features is a strategy used in this paper. As claimed by the authors, the proposed method can be used for both supervised and unsupervised models, as it doesn't have any parameter.
	[7]	Image processing research heavily relies on feature selection. This paper discusses how to apply deep convolutional neural network can be used for feature selection as an image processing application by constructing tree classifier. The proposed method was tested over different datasets.
	[8]	Feature selection is considered as an optimisation problem, as it needs to minimise the number of features by the corresponding method. Particle swarm optimisation (PSO) is one such method which is traditionally used for feature selection. This paper used the mutual information along with PSO in order to minimise the number of features.
	[9]	Causality-based feature selection is a type of method for finding the essential features amongst the available set. This paper provides an exhaustive collection that held over the past. A new package was also developed as an outcome of this research.
	[10]	This paper compares three correlation-based feature selection methods and the results are also discussed about the performance.
	[11]	This paper summarises the spectrum of AI techniques that can be used to find the variable which will have impact over the dataset.
	[12]	In order to reduce or minimise the number of variables, this paper discusses the streaming concept of selection process. It will select the importance of variables without taken into consideration of the entire available attributes in the dataset.
	[13]	Graphs are useful to represent various real-time problems. This paper discusses how to select the features by representing the problem space as graphs. The proposed method was evaluated in terms of the execution time and performance.
	[14]	It is a three-step method to find the important variables. The graph is used to model the problem space, find the mutual information and then select the variables which are important. This is the strategy followed in this paper. It is also evaluated by having some of the datasets over the proposed method.

	[15]	The strategy adopted in this proposed method was in converse with the existing methods. The difference between the variables which have marginal value will be considered as one set and the non-selected variables set are measured. This provides the subset of variables which will provide the final set of variables as the resultant set.
	[16]	In this proposed method, the four essential conditions that needs to meet for the feature selection algorithm is satisfied, as it claims. It is a tree-based variable selection algorithm, used to find the important variables. The scalability factor is also satisfied with the proposed method.
	[17]	In order to enhance the efficiency of the variable optimising method, the researchers have proposed randomised algorithms. An automatic breadth searching and attention searching adjustment approaches to further speedup randomized wrapper-based feature selection was proposed by this research. The results are compared with real-time and synthetic datasets.
	[18]	This paper used matrix factorisation way of splitting the variables and then applied to find the important variables over the reduced problem space. It is also evaluated against real-time datasets.
	[19]	Clustering is a machine learning model which is familiarly used to categorise the data labels. This strategy is used to find the important variables in this paper. The unsupervised graph-based representation technique is used to represent the data along with connected components. This will split the variables as number of sets – one for important variables and another one is not important.
Compound Method	[20]	Dimensionality reduction contains two approaches – one is the feature selection and another one is the feature extraction. Most of the research work is carried out in the area of feature selection, whereas this paper concentrated on to combine these two. The authors used a specific method to find two sets of variables – original and transformed.
Application related works	[21]	Dimensionality reduction plays an important role in finding the subset of variables or attributes or features which can be used to improve the performance of a machine learning model. This is applied to applications of variant domains. This paper applied the feature selection method for music emotion recognition.
	[22]	Sentiment analysis expresses the evaluation of the statements given by the users in a common platform such as social networking sites. The number of attributes in a dataset plays a critical role to decide the polarity of the sentiment. A combination of methods is used in this paper to find the required important attributes from the dataset concerned.
Comparative related works	[23]	A comparative analysis of four different methods to predict the required features of household energy consumption is done in this paper.
	[24]	The researchers consider the robustness of variable prediction. This paper discusses about the stability measures, merits and demerits of feature selection methods. It is also provided the comparison by considering some experimental datasets.

3. Materials and Methods

There are four different datasets taken from UCI Machine Learning repository, for consideration. The iris dataset [25] is widely used for fundamental research in different domains including image processing and machine learning. It has four attributes with 150 instances. The Wisconsin Breast Cancer dataset [26] is a benchmarking dataset. It has ten attributes and 699 instances. The CT scan slice location is a medical diagnosis dataset [27]. It has 386 attributes with 53,500 instances. The QSAR dataset [28] is from the physical sciences domain. It has 1024 attributes with 8992 instances.

The ID column has been removed from the dataset, if it is in existence, for further processing. It is the pre-processing step applied over the datasets taken for study. The reason towards this is since it is not going to play any role.

The following are the methods considered in this article and they are implemented in R environment along with the required package specific to the method.

(i) Boruta: it is a feature selection algorithm which has basic working principle from random forest. It finds the rank for the feature's existence in the dataset. It decides the importance of the variable which are statistically significant. The algorithm runs for some minimum number of times, by default it is 100. It will permute the features and runs the algorithm by finding the rank with the specified number of times.

(ii) moDel Agnostic Language for Exploration and eXplanation (DALEX): It is basically a machine learning model that can also be used for feature selection. It can be achieved by finding the link between predicate variable and the predictor variable. The working principle of this method is on regression model.

(iii) Least Absolute Shrinkage and Selection Operator (LASSO): It is a regression based variable selection technique. It considers the cost of coefficients for the attributes. Over the evolution of the method, it reduces the value of the coefficient thereby reducing the feature from the existing features. It is also an NP-hard feature selection method.

(iv) Recursive Feature Elimination (RFE): It works with the principle of k-fold cross validation process. It repeats the same for some specified number of times with the required parameters such as size and control.

(v) Regularised Random Forest (RRF): It constructs the trees and combines the tree ensembles. At every construction of trees, it determines the information gain and this process proceeds until the required gain is more for the corresponding feature (attribute).

4. Results

The summary of the datasets is shown in the Table-1.

Table1. Summary of the datasets

Name of the dataset	Number of features	Number of instances
Iris	4	150
Breast Cancer Wisconsin (Original)	10	699
CT Scan slice location	386	53,500
QSAR	1,024	8992

The proposed methods like Boruta, DALEX, LASSO, RFE, and RRF are applied to select the most prominent features from the datasets. The results are shown in table-2 and subsequently in figure-X.

Table 2. Number of selected features

Name of the Dataset	Features	No. of selected features by the Method				
		Boruta	DALEX	Lasso	RFE	RRF
Iris	4	4	4	4	4	4
Breast Cancer Wisconsin (Original)	9	9	9	9	9	9
CT Scan slice location	386	136	302	274	283	196
QSAR	1024	293	596	674	729	529

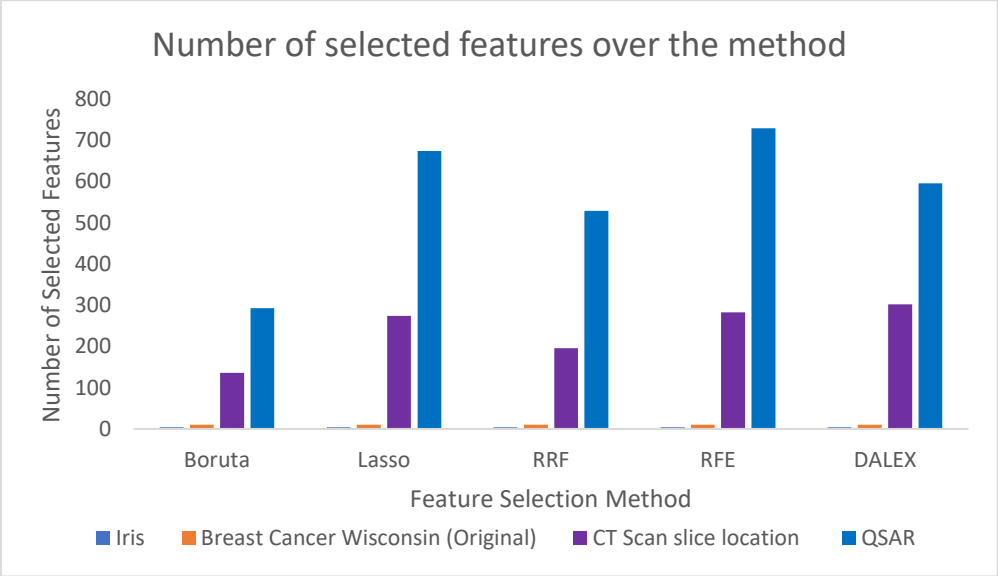

Figure 2. Number of selected features over the methods

Please note: the selected features from the iris dataset are invisible in the above figure, as the number of features is very less (that is only 4 features).

5. Discussion

This article assesses about some of the emerging methods related to the feature selection. There are five different methods such as Boruta, DALEX, Lasso, RFE, and RRF considered for study. The datasets of variant in sizes in terms of the number of attributes and the number of instances is considered for study. The number of features in the iris dataset and the Breast cancer Wisconsin dataset has no impact over all the methods. The other two datasets have variant results in the individual methods of feature selection. The number of features is more and the number of selected features is considerably less over the methods. There is a significant reduction in the percentage of the features selected with the methods. There is a 65% reduction in Boruta method and around 50% of reduction in RRF method for the CT Scan slice location dataset. There is a 71% reduction in Boruta and around 50% reduction in RRF method for QSAR dataset. There is a reduction at the rate of 42% in DALEX method over QSAR dataset, at the same time it has only 22% reduction in CT scan slice location dataset. The reduction is not so significant for both the datasets, in RFE whereas It has slight variation in Lasso method. There is a marginal reduction in the number of features for these two datasets over the Lasso method. Still, the methods adopted in this article would definitely help the researchers to find the appropriate features over the dataset they have taken for study and the method which would provide significantly better result.

6. Conclusion

Feature selection is used to find the useful features available in the dataset. It is still considered as a cumbersome process. The researchers proposed wide spectrum of methods which can be used for the same purpose. At the same time, the individually proposed method concentrates on its own merits. This article tried to assess some of the emerging feature selection methods. In order to achieve this, five different methods were used over four different datasets of variant in sizes in terms of the number of features (attributes) and the number of instances. The results shown that there is no reduction in the number of features over the two datasets whereas it shows significant reduction in the number of features that can be selected over the other two datasets. The reduction is evident in terms of its percentage for two methods. It has marginal impact for two of the datasets for a method. However, the researchers have choices to be considered in order to fit it with the research problem, the size of the dataset and the method that needs to be considered while choosing the feature selection.

References

1. J. Li *et al.*, “Feature Selection: A Data Perspective,” *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, Jan. 2018, doi: 10.1145/3136625.
2. F. P. Shah and V. Patel, “A review on feature selection and feature extraction for text classification,” *2016 International Conference on Wireless Communications, Signal Processing*

and Networking (WiSPNET), Mar. 2016, doi: 10.1109/wispnet.2016.7566545.

3. X. Su and F. Liu, "A Survey For Study of Feature Selection Based on Mutual Information," *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Sep. 2018, doi: 10.1109/whispers.2018.8746913.
4. A. K. Uysal, "On Two-Stage Feature Selection Methods for Text Classification," *IEEE Access*, vol. 6, pp. 43233–43251, 2018, doi: 10.1109/access.2018.2863547.
5. C. Fahy and S. Yang, "Dynamic Feature Selection for Clustering High Dimensional Data Streams," *IEEE Access*, vol. 7, pp. 127128–127140, 2019, doi: 10.1109/access.2019.2932308.
6. D. Wang, F. Nie, and H. Huang, "Feature Selection via Global Redundancy Minimization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015, doi: 10.1109/tkde.2015.2426703.
7. G. He, J. Ji, H. Zhang, Y. Xu, and J. Fan, "Feature Selection-Based Hierarchical Deep Network for Image Classification," *IEEE Access*, vol. 8, pp. 15436–15447, 2020, doi: 10.1109/access.2020.2966651.
8. H. S. Baruah, J. Thakur, S. Sarmah, and N. Hoque, "A Feature Selection Method using PSO-MI," *2020 International Conference on Computational Performance Evaluation (ComPE)*, Jul. 2020, doi: 10.1109/compe49325.2020.9200034.
9. K. Yu *et al.*, "Causality-based Feature Selection," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–36, Oct. 2020, doi: 10.1145/3409382.
10. N. Gopika and A. M. kowshalaya, "Correlation Based Feature Selection Algorithm for Machine Learning," *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, Oct. 2018, doi: 10.1109/cesys.2018.8723980.
11. S. Cateni, M. Vannucci, M. Vannocci, and V. Coll, "Variable Selection and Feature Extraction Through Artificial Intelligence Techniques," *Multivariate Analysis in Management, Engineering and the Sciences*, Jan. 2013, doi: 10.5772/53862.
12. X.-T. Wang and X.-Z. Luan, "Bayesian Penalized Method for Streaming Feature Selection," *IEEE Access*, vol. 7, pp. 103815–103822, 2019, doi: 10.1109/access.2019.2930346.
13. Y. Akhiat, M. Chahhou, and A. Zinedine, "Feature Selection Based on Graph Representation," *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, Oct. 2018, doi: 10.1109/cist.2018.8596467.
14. Z. Zhang and E. R. Hancock, "A Graph-Based Approach to Feature Selection," *Graph-Based Representations in Pattern Recognition*, pp. 205–214, 2011, doi: 10.1007/978-3-642-20844-7_21.
15. Y. Liu, F. Tang, and Z. Zeng, "Feature Selection Based on Dependency Margin," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1209–1221, Jun. 2015, doi: 10.1109/tcyb.2014.2347372.
16. Z. Xu, G. Huang, K. Q. Weinberger, and A. X. Zheng, "Gradient boosted feature selection," *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2014, doi: 10.1145/2623330.2623635.
17. Z. Wang, X. Xiao, and S. Rajasekaran, "Novel and efficient randomized algorithms for feature selection," *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 208–224, Sep. 2020, doi: 10.26599/bdma.2020.9020005.
18. J. K. Valadi, P. T. Ovhal, and K. J. Rathore, "A Simple Method of Solution For Multi-label Feature Selection," *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Feb. 2019, doi: 10.1109/icecct.2019.8869493.
19. S. Yang, F. Nie, and X. Li, "Unsupervised Feature Selection with Local Structure Learning," *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct. 2018, doi: 10.1109/icip.2018.8451101.
20. [Sreevani and C. A. Murthy, "Bridging feature selection and extraction: Compound feature generation (extended abstract)," *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, Apr. 2017, doi: 10.1109/icde.2017.30.
21. E. Widiyanti and S. N. Endah, "Feature Selection for Music Emotion Recognition," *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Oct. 2018, doi: 10.1109/icicos.2018.8621783.
22. F. R. Saputra Rangkuti, M. A. Fauzi, Y. A. Sari, and E. D. L. Sari, "Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection," *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, Nov. 2018, doi: 10.1109/siet.2018.8693211.
23. A. M. Pirbazari, A. Chakravorty, and C. Rong, "Evaluating Feature Selection Methods for Short-Term Load Forecasting," *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Feb. 2019, doi: 10.1109/bigcomp.2019.8679188.

24. P. Mohana Chelvan and K. Perumal, "A comparative analysis of feature selection stability measures," *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, May 2017, doi: 10.1109/icoei.2017.8300901.

Dataset References

25. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
26. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
27. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
28. D. Ballabio, F. Grisoni, V. Consonni, R. Todeschini (2019), Integrated QSAR models to predict acute oral systemic toxicity, Molecular Informatics, 38, 180012; doi: 10.1002/minf.201800124