EasyChair Preprint

№ 4949

A literature review of feature selection methods

Dyari M. Ameen M. Shareef and Ghader Ali Yosefi

February 3, 2021

A review of feature selection methods

DYARI M.AMEEN M.SHAREEF M.SHAREEF[1], GHADER ALI YOSEFI[2]

1, corresponding author (dma140h@cs.soran.edu.iq),

2, (ghadir.yosefi@visitors.soran.edu.iq@soran.edu.iq),

1,2 Computer Science Department - Faculty of Science - Soran University, Soran, Kurdistan, Iraq

Abstract:

The process of accommodating data is limited by the evolution of hardware and technologies, and the current analytical tools are not sufficient enough to retrieve the current overwhelming flood of data. The agenda of feature selection is to choose a subset of features from the input space while reducing effects, from noise or irrelevant features, and still efficiently describe the input data that ends up in good prediction results. In this work, we present basic knowledge about the fields wherein feature selection methods are important, and show you the results of our review. We have observed that the vast majority of papers we reviewed emphasizes on handling high-dimensional data with the help of human being interference. With having said that authors, most of the time, came up with methods that are less computational than the methods that are currently available in the market. Finally, we suggest some future works which are worth to be worked on and investigate.

Keywords: Big data; Feature Selection Methods; Machine Learning; Data Mining; Data preprocessing

1. Introduction

According to (Petrov, 2020), the amount of generated data was estimated for the year 2020 is roughly 40 zettabytes, internet users generate almost 2.5 quintillion bytes of data daily, Twitter users publish 0.5 million tweets every minute and each person in 2020 was estimated to have generated 1.7 megabytes every second. As such, as the technology evolves and with the fast level of advancement in the world of big data, the sources of generating data are on the increase.

However, the current analytical tools are not sufficient enough to retrieve the information (Hsinchun Chen, 2012). Since datasets may increase in volume, the analytical tools like the computations may even get worse as extra computational overheads will be added upon (Shukla, et al., 2019). With having said that one of the overwhelming challenges is the problem of turning growing data into accessible and actionable knowledge that can later be used.

The attempts to counter these challenges have resulted in a new area called Data mining. Data mining is used as the core task in the process known as Knowledge Discovery which consists of applying computational techniques to extract useful information, including patterns, or eliminate useless information (Alfred, 2005). High-dimensional data like images, gene expressions microarrays and financial time series has become the top obstacle to cope with since it requires high specification hardware.

To handle the problem of the high number of input features, using feature selection methods have become an inevitable part. Feature selection is investigated and used by machine learning and data mining community as it is the process of keeping in the relevant features and discarding the irrelevant features (Bolón-Canedo, et al., 2012).

In this paper, we provide a review of feature selection methods that have been applied mostly in the last decade which took roughly two months. This review was a compulsory requirement asked by the regulations defined by the ministry of higher education for master degree programs. The rest of the paper is organized as follows: in the next section, basic definitions for the mentioned concepts are presented. In the section after that, the state of art studies is presented. In the final section, the conclusion is presented.

2. Background

In this section, the main concepts, that are mentioned in this paper or/and required to be understood to better benefit from this paper, are presented.

2.1 Feature Selection

A feature is an individual measurable attribute of the process feature selection that is being observed. A unique feature means having useful information and provides a score of the feature's usefulness in discriminating the various classes. The agenda of feature selection is to choose a subset of features from the input space while reducing effects from noise or irrelevant features. Feature selection helps in understanding data much better, reducing computations, avoiding the curse of dimensionality[1], and improving the predictor accuracy. To remove an irrelevant feature, a feature selection criterion is a must since the measure of the relevance of every feature should be computed. If irrelevant features are used in a model, the information still will be used by the model and this will lead to poor generalization. The feature selection methods are categorized into filter, wrapper and embedded methods (Chandrashekar & Sahin, 2013) (Bolón-Canedo, et al., 2012).

2.1.1 Filter methods

Filter methods, as the principal criteria, use feature ranking techniques for feature selection by ordering. A ranking method is a filter method since they are applied before classification to remove

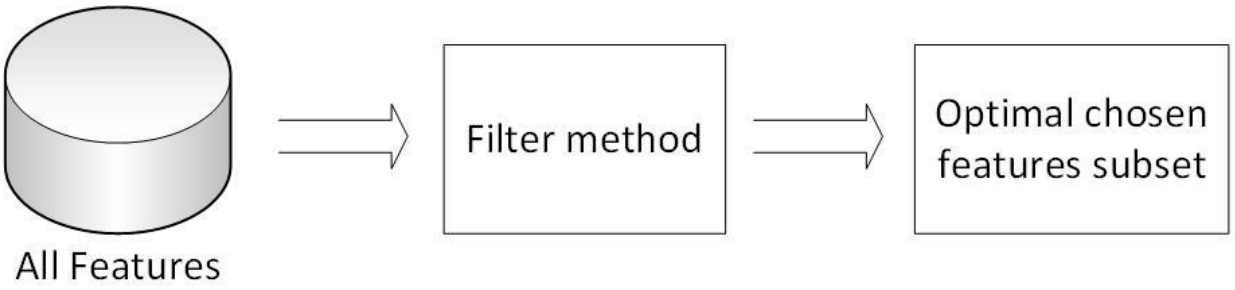

Fig. 1 Schema of the filter feature selection methods

the less relevant features. The filter methods don't depend on any learning algorithm since they choose the features whose ranks are the highest among them. The optimal chosen features subset is chosen based on computed measures in various tests for their correlation with the outcome feature (Chandrashekar & Sahin, 2013).

[1] Occurs when the learning algorithm chosen perform poorly on high-dimensional data.

2.1.2 Wrapper methods

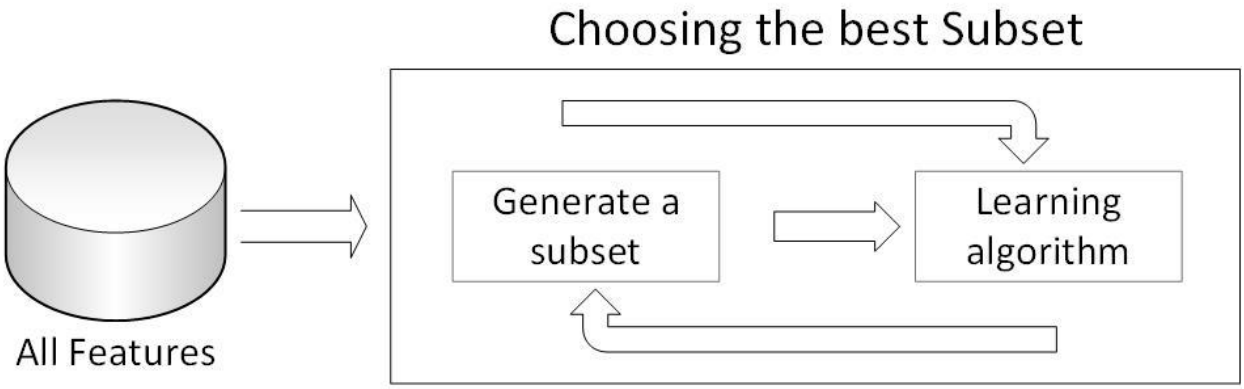

Fig. 2 Schema of the Wrapper feature selection methods

Wrapper methods take the predictor as a black box and the predictor performance as the objective function to assess the right feature set. Any number of search algorithms can be used to figure out the subset of feature which maximizes the objective function, which, in here, is the classification performance. Moreover, based on the conclusions we draw from the previous model, we decide to keep or let go of features, from the so-far selected subset (Remeseiro & Bolon-Canedo, 2019) (Chandrashekar & Sahin, 2013).

2.1.3 Embedded methods

Embedded methods try to reduce the computation time that is being used to reclassify different subsets as the feature selection works as part of the training process. (Remeseiro & Bolon-Canedo, 2019)

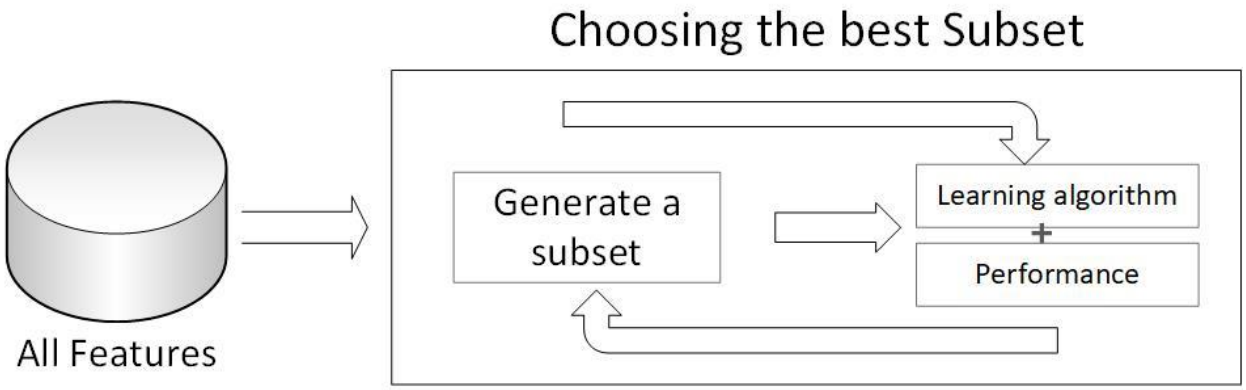

Fig. 3 Schema of the Embedded feature selection methods

3. State of art studies

In this section, several works regarding feature selection methods are presented and discussed.

(Mitra, et al., 2002) proposed an algorithm for unsupervised feature selection based on feature similarity measures which requires no search processes. The authors partition the original feature subset into clusters based a similarity function and choose their corresponding representatives from each cluster to form the sample. Maximal Information Compression Index (MICI) is used as the feature similarity measure, and partitioning of features is based on the K-NN principle with k acting as a scale parameter. Consequently, the algorithm was fast and the authors associate the speedup achieved by the algorithm with the new use of the MICI measure, and the algorithm itself. The algorithm has complexity O(D2) (D refer to the original feature set). However, it doesn't take any account of any clustering structure while selecting the features.

(H, et al., 2013) demonstrated a simple and efficient feature selection method in which Authors observed there is a negative correlation between Attribute Ratio and accuracy. The method takes advantage of the attributed average of total and each class data, and then the decision tree classifier evaluated with the method will detect four different data types. The method uses Attribute Ratio which is calculated by mean and frequency of features. The authors used a dataset of 41 features which had three types of features: Numeric, Nominal and Binary. The accuracy achieved was higher than the accuracy of the methods Correlation-based Feature Selection, Information Gain and Gain Ratio. However, they only used J48 decision tree, as a classifier, and 10-fold cross validation, as the validation measure for testing, to estimate the accuracy.

(Shih-Wei, et al., 2012) showed an intelligence algorithm to find the most relevant features in dataset by combining three learning techniques, namely Support Vector Machine, Decision Tree and simulated annealing. In the very first place, the authors pre-processed the dataset as training and testing data, set the initial values of parameters and generated initial solution randomly. The proposed algorithm starts with four parameters where Simulated Annealing and Support Vector Machine are simultaneously implemented to optimize some parameters and select the best subset of features to maximize accuracy of classification. After that, Simulated Annealing and Decision Tree are implemented to optimize some other parameters to increase the testing accuracy for selected features and build the decision rules to increase the testing accuracy. Finally, if the termination creation is met, which is defined in advance, it will report the best testing accuracy,

selected features, and decision rules. Otherwise, the two steps are repeated to reach the best testing accuracy for the selected features. The results were positive since the performance of the proposed combining algorithm outperforms the performance of other mentioned learning algorithms when they are performed alone. However, the algorithm computationally is demanding and has a lot of concerns while performing.

(Hui, et al., 2014) proposed a SVM algorithm based on a novel stable wrapper feature selection using linear SVM, polynomial SVM, gaussian SVM and sigmoid SVM. Predictions of different SVM models on each candidate feature are converted into ranking-order information of each feature. Moreover, the algorithm combines the two computed statistical values of mean and standard deviation based on the ranking-order information to calculate a feature selection index of SVM. Finally, the index is used to select the most relevant features. To evaluate the performance of proposed approach, the authors witnessed that Gaussian SVM and linear SVM with Statistics-based Wrapper Feature Selection improves the predictive performance as compared to other SVM models. However, this algorithm computationally is demanding since it goes through a lot of steps and has an exponential size on the input.

(Opeyemi, et al., 2016) presented an ensemble-based multi-filter feature selection method which reduces the feature set while improving and maintaining the classification accuracy using a decision tree classifier. The method starts with all features, then, the most relevant features using the filter methods Information Gain, Gain Ratio, Chi-Squared and Relief are achieved, and the output of one-third split of the filter methods are combined. Finally, the method uses a predefined threshold and a simple majority vote in order to reach the final feature set. The results of the proposed method were positive since it achieved better results with the final feature set. However, the method has a high computational complexity since it takes advantage of three methods.

(Ienco & Meo, 2008) came up with a hybrid approach consisting of hierarchical clustering and the representatives of the clusters. First, the appropriate distance measure is selected, followed by grouping features based on clustering methods. Finally, the most representative feature of each cluster is selected to reach the reduced subset. The approach was tested over 40 datasets and showed a better classification accuracy. However, even though the approach doesn’t require any special nor any complex parameter tuning process but the design of packaging method does not only increase the time cost but also will have the learning method biased.

(Yu & Liu, 2004) proposed that feature relevance is not sufficient when it doesn't come with redundancy analysis. The authors developed a correlation-based method that uses C- and F-correlations for relevance and redundancy analysis. Additionally, a new algorithm called FCBF is implemented. Finally, the results are verified by two learning algorithms as the proposed method verifies its efficiency and effectiveness in supervised learning. The features often, in this paper, can be divided into four groups; Irrelevant features, Weakly irrelevant and redundant features, Weakly relevant but non-redundant features and/or Strongly relevant features. Supervised feature selection results should include the groups Weakly relevant but non-redundant features and Strongly relevant features. However, the proposed method's values need to be discrete numbers because it is symmetrical uncertainty measure. So, this method cannot deal with regression problems since the classes have continues values.

(Vandenbroucke, et al., 2000) reached a new approach for feature selection which uses a competitive learning scheme to differentiate the samples and assess the scale of clusters. The scheme then will have the original set grouped into several reduced subsets, following by a judgement function, designed for the average dispersion within classes, which is calculated for each feature subset. The feature subset, that maximize the judgment value, is selected to choose the candidate feature. In the end, if the correlation coefficient calculated between the candidate feature and the selected feature happened to be greater than 0.75, the candidate feature will be dropped.

(Ahmad, et al., 2017) proposed a machine learning-assisted algorithm which is a mixture of consistency subset evaluation and DDoS characteristic features. Authors first carried out feature selection on a total of 42 features. Then, they applied DDoS characteristic based features and Consistency-based Subset Evaluation in a parallel manner. Finally, they used a simple majority vote technique on the output of these two feature selection methods to select the most relevant feature based on a predefined threshold. The proposed algorithm was better than traditional ones such as Information Gain, Gain Ratio, Correlated features selection. However, the algorithm does nothing to tackle down feature redundancy.

(Lifang, et al., 2010) presented a novel clustering method combined with feature ranking. The method provides the linear correlation coefficient for feature ranking and Modified Global K-means algorithm (MGKM) where, as the number of top-ranked features falls, a point where the

cluster function value falls heavily is selected, and the final selected features are identified afterwards. The method works like this: at initial state, all variables are selected, then, the method is used to identify the cluster structure. Next, the linear correlation coefficient is used for feature ranking. Finally, the feature ranking result is used to inform and recalculate the clusters. This method can adaptively select the working feature vector according to various patterns of data with low complexity. However, this method could not capture correlations between features that are not linear in nature.

(Hongbin & Guangyu, 2002) used the tabu search for subset generation and compared it with classical algorithms. The authors evaluated the generated subset using classification error criteria to find the better feature subset. During their experiments, the results were very positive and encouraging since the experiments showed that the tabu search didn't only provide them the optimal or almost optimal subset, but also requires less computational time as compared with the branch and bound method and most other currently used suboptimal methods. However, the datasets used for experiments are synthetic.

(Sebastián & Richard, 2009) demonstrated a novel wrapper method for feature selection by combing SVM with a specific Kernel function. The method commences with all features and determine each feature's contribution to the corresponding classifier. Moreover, the one having the least impact on the classification accuracy in an independent validation subset is removed in each iteration until a termination criterion indicates that a better solution has been found. This method performs a sequential backward elimination of features to generate the reduced feature subset. The authors used the number of errors in a validation subset as the measure to decide which feature to remove in each iteration. However, this algorithm can be very expensive if the number of input feature goes high.

(Julia, et al., 2004) showed a number of novel embedded approaches for simultaneous feature selection and classification within a general optimization framework. The authors included both linear and nonlinear SVMs and applied difference of convex functions programming to solve their problems and they did their experiments on both real-world and synthetic data.

(B.Azhagusundari & Antony, 2013) came up an algorithm based on discernibility matrix and Information gain to find optimal feature subset. Consequently, the results were better in terms of number of features selected and accuracy as compared to applying methods separately. The

authors came up with the fact that when a feature does not have very much impact on the data classification, it can be discarded without any effect on the detection accuracy of a classifier because it has very small information gain. However, as did the majority of papers, this paper as well has not applied on real-world datasets.

(Robert, et al., 2011) presented a study of feature selection methods using a number of combination methods. The authors performed experiments on 18 various multi-class text categorizations and a number of ranking merging methods for combining features from multiple methods. As a result, the single methods showed to be generally better than combinational methods. However, in their study, they didn't report time complexity. So, the authors' so-called "no combination showed to be generally superior to the best single methods" statement in paper is right only in some limited capacity.

(Chih-Fong & Yu-Chi, 2019) examined two different combination orders of feature selection and discretization in sense of classification accuracies and computational times. Moreover, the authors focused on reaching the best combination of feature selection and discretization, and used the principal component analysis as the filter method, the genetic algorithm as the wrapper method, and the C4.5 decision tree as the embedded method to represent the three types of feature selection methods. They then compared 12 different combinations over 10 datasets, and found that the best choice for the SVM classifier is the minimum description length principle followed by the C4.5 decision tree, and second-best choice is the C4.5 decision tree followed by the minimum description length principle. Despite insignificant differences in performance the combination of the minimum description length principle followed by the C4.5 decision tree, generally, is suggested since it provides both the highest classification accuracy and the least computational time. However, as, specifically, for the decision tree classifier, the optimal combination is the C4.5 decision tree followed by the minimum description length principle.

4. Conclusion

In feature selection era, there are two main approaches; Individual evaluation and subset evaluation in which the former assesses each feature and assign them scores per their degree of relevance whereas the latter assesses candidate feature subsets based a technique. In our work, after reviewing more than 10 papers in detail, we have observed that both evaluations have their disadvantages; the individual evaluation is unable to remove redundancy because redundant features are most likely to have similar weights whereas subset evaluation can cause the methods to suffer from the problem caused by searching among candidate feature subsets. Consequently, for high-dimensional data, which may contain a huge number of redundant features, the individual evaluation may produce results beyond optimal, because as long as features look relevant, they will be selected even though many of them might highly be correlated to each other. On the other hand, though there exist many search types including the heuristic searches, the majority of them still incur time complexity, which prevents them to be scalable to large datasets, so the methods that can be used in subset evaluations can suffer from this. As a result, we recommend that the current feature selection methods are better to go beyond the concepts of relevance and redundancy.

According to our research-based knowledge, the goodness of a feature selection method is to have high accuracy while having less time and space complexity. Even though there are a large number of reviews on the feature selection methods, they are necessarily emphasizing on specific research fields. So, it would be interesting to investigate and explore measures that can handle all types of values, and come up with methods and approaches that can be combined to handle all (if not the majority) types of value and fields equally. In literature, the proposed algorithms are being implemented on large datasets where they handle millions of samples and features at a time, but the majority of the state-of-art selection methods are not able to cope with these growing vast datasets since they are not being developed under that assumption or they are being tested on famous or synthetic datasets. So, it would be very unique and important to explore more sophisticated methods, such as parallel programming, that can cope with big data and real-life datasets. Not to mention the very majority of papers proposed static methods without even mentioning online feature selection, in which the data changes over time. So, exploring more methods to deal with online feature selection would be of need.

To sum up, the current feature selection techniques are either computationally practicable but not optimal, or they are optimal or very close to optimal but cannot handle computational complexity of feature selection problems of realistic size.

5. Bibliography

Ahmad, R. et al., 2017. *Adaptive feature selection for denial of services (DoS) attack.* Miri, Malaysia, IEEE.

Alfred, R., 2005. In: *Knowledge Discovery: Enhancing Data Mining and Decision Support Integration.* Heslington, York YO10 5DD, United Kingdom: The university of York, pp. 6-8.

Alpaydin, E., n.d. What is Machine Learning. In: *Introduction to Machine Learning.* s.l.:s.n., pp. 3-4.

B.Azhagusundari & Antony, S., 2013. Feature Selection based on Information Gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* 2(2), pp. 2278-3075.

Berthold, M. R., Borgelt, C., Höppner, F. & Klawonn, F., 2010. Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data. In: *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data.* s.l.:Springer, pp. 1-3.

Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A., 2012. A review of feature selection methods on synthetic data. *Knowledge and Information Systems,* p. 1.

Cai, J., Luo, J., Wang, S. & Yang, S., 2018. Feature selection in machine learning: A new perspective. *Neurocomputing,* Volume 300, pp. 70-79.

Chandrashekar, G. & Sahin, F., 2013. A survey on feature selection methods. *Computers & Electrical Engineering.*

Chen, H., Chiang , R. H. L. & Storey, V. C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly.*

Chih-Fong, T. & Yu-Chi, C., 2019. The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences,* Volume 505, pp. 282-193.

Fayyad, M., Piatetsky-Shapiro, G. & Smyth, P., 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Advances in Knowledge Discovery and Data Mining.*

H, C., B, J., SH, C. & T, P., 2013. Feature Selection for Intrusion Detection using NSL-KDD. *Recent advances in computer science.*

Hongbin, Z. & Guangyu, S., 2002. Feature selection using tabu search method. *Pattern Recognition,* 35(3), pp. 701-711.

Hsinchun Chen, R. H. L. C. a. V. C. S., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly,* p. 24.

Hui, L., Chang-Jiang, L., Xian-Jun, W. & Jie, S., 2014. Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine. *Applied Soft Computing,* Volume 19, pp. 57-67.

Ienco, D. & Meo, R., 2008. *Exploration and Reduction of the Feature Space by Hierarchical Clustering.* Atlanta, Georgia, USA, Proceedings of the SIAM International Conference on Data Mining.

Julia, N., Christoph, S. & Gabriele, S., 2004. SVM-based Feature Selection by Direct Objective Minimisation. *Joint Pattern Recognition Symposium,* Volume 3175, pp. 212-219.

Kohavi, R. & John, G. H., 1997. Wrappers for feature subset selection. *Artificial Intelligence,* p. 1.

Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety.

Lifang, Z., John, Y. & Xin-Wen, W., 2010. *Adaptive Clustering with Feature Ranking for DDoS Attacks Detection.* Melbourne, VIC, Australia, International Conference on Network and System Security, NSS.

Mitra, P., Murthy, A. & Pal, K., 2002. Unsupervised feature selection using feature similarity. *IEEE,* 24(3).

Opeyemi, O. et al., 2016. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *Springer,* Volume 130.

Petrov, C., 2020. *25+ Impressive Big Data Statistics for 2020.* [Online]
Available at: https://techjury.net/blog/big-data-statistics/#gref

Polat, H., Danaei, H. & Cetin, A., 2017. Diagnosis of Chronic Kidney Disease Based on Support Vector. *Journal of Medical Systems.*

Remeseiro, B. & Bolon-Canedo, V., 2019. A review of feature selection methods in medical applications. *Computers in Biology and Medicine,* p. 1.

Robert, N., Rudolf, M. & Kjetil, N., 2011. Combination of Feature Selection Methods for Text Categorisation. *European Conference on Information Retrieval,* Volume 6611, pp. 763-766.

Russell, S. & Norvig, P., 2020. What is AI?. In: *Artificial Intelligence: A modern approach: Fourth Edition.* s.l.:Pearson, p. 5.

Sebastián, M. & Richard, W., 2009. A wrapper method for feature selection using Support Vector Machines. *Information Sciences,* 179(13), pp. 2208-2217.

Sheikhpour, R., Saram, M. A., Gharaghani, S. & Chahooki, M. A. Z., 2017. A Survey on semi-supervised feature selection methods. *ELSEVIER.*

Shih-Wei, L., Kuo-Ching, Y. b., Chou-Yuan, L. & Zne-Jung, L., 2012. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. *Applied Soft Computing,* 12(10), pp. 3285-3290.

Shukla, A. K., Yadav, M., Kumar, S. & Muhuri, K. P., 2019. Veracity handling and instance reduction in big data using interval type-2. *ELSEVIER.*

Vandenbroucke, N., Macaire, L. & Postaire, J.-G., 2000. *UNSUPERVISED COLOR TEXTURE FEATURE EXTRACTION AND SELECTION FOR SOCCER IMAGE SEGMENTATION.* Vancouver, BC, Canada, Canada, IEEE International Conference on Image Processing.

Vergara, J. R. & Este´vez, P., 2013. A review of feature selection methods based. *Springer-Verlag London 2013.*

Vergara, J. R. & Estévez, P. A., 2013. A review of feature selection methods based on mutual information. *Neural Computing and Applications,* p. 1.

Yu, L. & Liu, H., 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research,* Volume 5, pp. 1205-1224.

Zikopoulos, P. C., Deroos, D. & Parasuraman, K., 2013. Harness the power of big data : the IBM big data platform.