A Literature Review of Feature Selection Methods

Dyari M. Ameen M. Shareef and Ghader Ali Yosefi

February 8, 2021

Feature Selection Methods - Review

Dyari M.Ameen M.Shareef[1], Ghader Ali Yosefi[1]
[1] 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 information from this 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. 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 the authors, very often, came up with methods that are less computational than the methods that are currently available in the market. In this work, we present basic knowledge about feature selection methods, review a number of papers and discuss their disadvantages, and draw conclusions based on our review. Finally, we suggest some future works which are worth to be worked on and investigate.

Keywords: Big data - Feature Selection Methods - Data Mining - Data preprocessing

1. Introduction

According to [1], the amount of generated data that had been estimated for the year 2020 was roughly 40 zettabytes, internet users generated almost 2.5 quintillion bytes of data daily, Twitter users published 0.5 million tweets every minute, and each person in 2020 had been estimated to have generated 1.7 megabytes every second. As the technology evolves and with the fast level of advancement in the world of data, the sources of generating data are on the increase.

However, the current analytical tools are not sufficient enough to retrieve the information [2]. Since datasets may increase in volume, the analytical tools like the computations may even get worse as extra computational overheads will be added upon [3]. With having said that one of the overwhelming challenges is the problem of turning growing data into accessible and actionable knowledge that later can 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 Knowledge Discovery which consists of applying computational techniques to extract useful information, including patterns, or eliminate useless information [4]. 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.

In this paper, we provide a review of feature selection methods that have been applied mostly in the last decade which took us roughly two months. 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, discussion and the state of the art is presented. In the final section, the conclusion is presented.

2. Technological Innovation for applied AI systems

AI is being ranked as one of the most interesting and fastest-growing fields by surveys. It makes trillion dollars a year in revenue. “Its impact will be more than anything in the history of mankind”, the AI expert Kai-Fu Lee predicts. Historically, Scientists and researchers have discovered many different types of AI. Some have referred to intelligent in the sense of fidelity to human performance, while others called it rationality. Eventually and simply put, AI is a fast-growing branch of Computer Science that has to do with building smart machines capable of performing the kinds of tasks that need human intelligence [5]. One of the subfields in the field of AI is Machine Learning which automates analytical works. When an agent learns from experience and observations about the world, s/he improves his/her performance. When the agent is a computer, it is called Machine Learning. A computer can observe some data and builds a model based on the data, then uses the model to predict. Before proceeding, why a machine should be learnt, even though we can program it? Well, there are two main reasons; the first, the builders cannot predicate all possible scenarios. Second, the builders happen to not know how to program a solution themselves. This paper present basic knowledge about feature selection methods, review a number of papers and discuss some future works which are worth to be worked on and investigate.

3. 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.

Feature selection aims to choose a subset of features from the input space while reducing effects from noise or irrelevant features. 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 targets. Feature selection helps in understanding data much better, reducing computations, avoiding the curse of dimensionality, and improving the predictor accuracy. 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 [6] [7].

3.1.1 Filter methods

Filter methods use feature ranking techniques for feature selection, and they are applied before classification to remove the less relevant features [6].

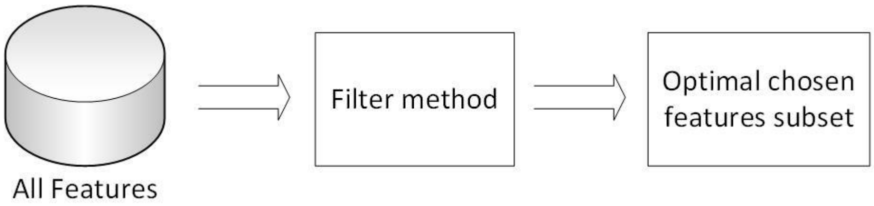

Fig. 1. Schema of the filter feature selection methods

3.1.2 Wrapper methods

Wrapper methods take the predictor as a black box and the predictor performance as the objective function to assess the right feature set [8] [6].

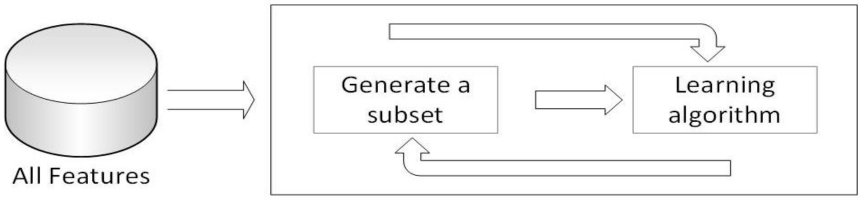

Fig. 2. Schema of the Wrapper feature selection methods

4. Discussion and state of the art

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

[9] proposed an algorithm for unsupervised feature selection, based on feature similarity measures, which requires no search processes. The authors partitioned the original feature subset into clusters based on a similarity function and chose 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. The algorithm is 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.

[10] demonstrated a simple and efficient feature selection method in which the authors observed that 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 is higher than the accuracy of the methods Correlation-based Feature Selection, Information Gain and Gain Ratio. However, they only used the J48 decision tree, as a classifier, and 10-fold cross-validation, as the validation measure for testing, to estimate the accuracy.

[11] presented an ensemble-based multi-filter feature selection method which reduces the feature set while improving and maintaining the classification accuracy by 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 to reach the final feature set. The results of the proposed method are promising since it achieves better results with the final selected feature set. However, the method has a high computational complexity since it takes advantage of three methods.

[12] 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 the packaging method does not only increase the time cost but also will have the learning method biased.

[13] 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. However, the method has a lot of computational concerns since it takes advantage of many functions.

[14] 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 the initial state, all features 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.

[15] 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 a 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 with the optimal or almost optimal subset, but also required 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.

[16] demonstrated a novel wrapper method for feature selection by combing SVM with a specific Kernel function. The method commences with all features and determines 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.

[17] came up with an algorithm based on discernibility matrix and Information gain to find optimal feature subset. Consequently, the results were better in terms of the number of features selected and accuracy as compared to applying the 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 to real-world datasets.

[18] 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 statement "no combination showed to be generally superior to the best single methods", in the paper, is right only in some limited capacity.

5. Conclusion and future work

In this section, the conclusion based on our review is given and several future works are discussed.

5.1 Individual evaluation and subset evaluation

In our work, 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 triggering by searching matters. 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.

5.2 Generalizing

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 and other kinds of papers 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.

5.3 Large and real Datasets and online feature methods

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.

5.4 Computation and complexity

Finally, we have also observed that in the current market, the 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.

References

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

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

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

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

[5] S. Russell and P. Norvig, "What is AI?," in *Artificial Intelligence: A modern approach: Fourth Edition*, Pearson, 2020, p. 5.

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* 2013.

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

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

[9] P. Mitra, A. Murthy and K. Pal, "Unsupervised feature selection using feature similarity," *IEEE,* vol. 24, no. 3, 2002.

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

[11] O. Opeyemi, C. Haibin, R. Kim-Kwang, D. Ali, X. Zheng and D. Mqhele, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing," *Springer,* vol. 130, 2016.

[12] D. Ienco and R. Meo, "Exploration and Reduction of the Feature Space by Hierarchical Clustering," Atlanta, Georgia, USA, 2008.

[13] N. Vandenbroucke, L. Macaire and J.-G. Postaire, "UNSUPERVISED COLOR TEXTURE FEATURE EXTRACTION AND SELECTION FOR SOCCER IMAGE SEGMENTATION," Vancouver, BC, Canada, Canada, 2000.

[14] Z. Lifang, Y. John and W. Xin-Wen, "Adaptive Clustering with Feature Ranking for DDoS Attacks Detection," Melbourne, VIC, Australia, 2010.

[15] Z. Hongbin and S. Guangyu, "Feature selection using tabu search method," *Pattern Recognition,* vol. 35, no. 3, pp. 701-711, 2002.

[16] M. Sebastián and W. Richard, "A wrapper method for feature selection using Support Vector Machines," *Information Sciences,* vol. 179, no. 13, pp. 2208-2217, 2009.

[17] B.Azhagusundari and S. Antony, "Feature Selection based on Information Gain," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 2, no. 2, pp. 2278-3075, 2013.

[18] N. Robert, M. Rudolf and N. Kjetil, "Combination of Feature Selection Methods for Text Categorisation," *European Conference on Information Retrieval,* vol. 6611, pp. 763-766, 2011.

[19] H. Polat, H. Danaei and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector," *Journal of Medical Systems,* 2017.

[20] E. Alpaydin, "What is Machine Learning," in *Introduction to Machine Learning*, pp. 3-4.

[21] R. Sheikhpour, M. A. Saram, S. Gharaghani and M. A. Z. Chahooki, "A Survey on semi-supervised feature selection methods," *ELSEVIER,* 2017.

[22] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* p. 1, 1997.

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

[24] D. Laney, "3D Data Management: Controlling Data Volume, Velocity and Variety," 2001.

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

[26] P. C. Zikopoulos, D. Deroos and K. Parasuraman, "Harness the power of big data : the IBM big data platform," 2013.

[27] M. R. Berthold, C. Borgelt, F. Höppner and F. Klawonn, "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*, Springer, 2010, pp. 1-3.

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

[29] J. R. Vergara and P. Este´vez, "A review of feature selection methods based," *Springer-Verlag London 2013,* 2013.

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

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

[32] L. Shih-Wei, Y. b. Kuo-Ching, L. Chou-Yuan and L. Zne-Jung, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Applied Soft Computing,* vol. 12, no. 10, pp. 3285-3290, 2012.

[33] J. Cai, J. Luo, S. Wang and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing,* vol. 300, pp. 70-79, 2018.

[34] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *Journal of Machine Learning Research,* vol. 5, pp. 1205-1224, 2004.

[35] R. Ahmad, I. Nur, S. Ali, H. Hazlina and T. Mohd, "Adaptive feature selection for denial of services (DoS) attack," Miri, Malaysia, 2017.

[36] N. Julia, S. Christoph and S. Gabriele, "SVM-based Feature Selection by Direct Objective Minimisation," *Joint Pattern Recognition Symposium,* vol. 3175, pp. 212-219, 2004.