Annals of Mathematics and Computer Science
ISSN: 2789-7206
Vol 4 (2021) 48-54

FEATURE SELECTION: STATE-OF-THE-ART SURVEY

SAIDJON KAMOLOV[1*]

Abstract. Feature selection is an important preprocessing step in many machine learning algorithms. In this paper, we review the current literature on feature selection algorithms. We discuss the major approaches and identify the advantages and disadvantages of each approach.

1. Introduction

Feature selection is an important preprocessing step in many machine applications. It is particularly relevant in big data. Feature selection is used in a wide range of fields including medicine, text processing, intrusion detection [14], microarray data, and others. Feature selection has a number of benefits. A model with fewer features is easier to interpret. A simpler model is usually scientifically more desirable than a more complicated model. Simple models are preferred to more complex models according to the established scientific principles. Fewer number of features has a positive effect on the computational complexity of the models. Concretely, with fewer features the size of the data is reduced and consequently the training times are also lower. In some machine learning algorithms such as decision trees the training times are directly related to the number of features. In decision tree classifier, the algorithm traverses each feature to find the optimal split point. Thus, with fewer feature there are fewer steps in finding the optimal split position. A model with fewer feature is also more interpretable. It is easier to extract knowledge and create rules from a small model compared to a large model where the relationships between the features and the target variable are more complicated and harder to untangle.

In many applications such as microarray data analysis the number of features can reach thousands. A model based on such a large number of features is not tractable. Even if a machine learning algorithm can be trained with so many features it would be a black box model. In many scientific fields such as medicine black box models do not pass the test for scientific rigor and cannot be used in practice. Therefore, it is imperative to obtain models with fewer features that can be analyzed and used for knowledge extraction.

Date: Received: Oct 10, 2021; Accepted: Nov 5, 2021.

* Corresponding author.

2010 *Mathematics Subject Classification.* Primary 46L55; Secondary 44B20.

Key words and phrases. feature selection; filter method; wrapper method; genetic algorithm; survey.

High dimensional data poses a challenge called the curse of dimensionality. Given a large number of features, it requires an exponentially larger amount of samples to learn meaningful patterns within data. If there are not enough samples then it is hard to distinguish between samples. In particular, samples appear to be sparse and dissimilar. Another issue with high-dimensional data is distance concentration which refers to the problem of all the pairwise distances between different samples/points in the space converging to the same value as the dimensionality of the data increases. A number of machine learning models including clustering or nearest neighbors methods use distance-based metrics to identify similar or proximity of the samples. Because of distance concentration, the concept of proximity or similarity of the samples may not be qualitatively relevant in higher dimensions.

There exist a wide range of feature selection methods [8, 29]. The field of feature selection continues to be actively researched with dozens of new feature selection methods appearing in the literature of a consistent basis. Feature selection methods can be generally grouped into three groups: i) filter methods, ii) wrapper methods, and iii) embedded methods. The three methods differ by the nature of identifying the relevant features.

In filter methods, a similarity metric is used to evaluate feature importance. There exist a number of similarity metrics such mutual information and chi^2 for discrete variables [13] and Pearson's correlation for continuous variables. Given a metric, the similarity between the target variable and each feature is calculated . The features are ranked according to their scores. Subsequently the top features are selected for model training. The number of selected features is either determined directly by the user or a heuristic. In many cases, a similarity metric can also be used to evaluate subsets of features. In this case, instead of choosing the set of optimal features based on the top ranked variables the optimal subset is identified directly based on the similarity scores between the subsets of the target variable. Although there are a wide range of pseudo-metrics used for feature evaluation, the most commonly used metric is mutual information which is given by the equation

$$\mathrm{I}(X,Y) = \sum_x \sum_y p_{(X,Y)}(x,y) \log \frac{p_{(X,Y)}(x,y)}{p_X(x) p_Y(y)}, \tag{1}$$

where $p_{(X,Y)}(x,y)$ is the joint probability distribution, $p_X(x)$ and $p_Y(y)$ are the marginal distributions.

2. Feature Selection Methods

A great volume of research has been dedicated to the topic of feature selection. It is important to distinguish between feature selection and dimensionality reduction in the literature landscape. In the former the outcome is a subset of the original features while in the later the outcome is a set of new features albeit of smaller dimension. There a several dimensionality reduction methods including Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in addition other popular techniques [9]. These are unsupervised methods that

create a new reduced subspace without using the class label information [3]. On the other hand, feature selection techniques can be classified as supervised methods, where the class information is utilized to identify the best subset of features. The current feature selection methods can be broadly divided into four categories: i) filter, ii) wrapper, iii) embedded methods, and iv) genetic algorithms. Filter methods are the most basic and the most efficient category of selection methods. Filters employ an independent metric such as information gain [30] or χ^2-statistic [20] to evaluate the relevance of a feature subset. Wrapper methods train a base model on a given feature subset to measure its importance. Wrapper methods normally perform better than filter methods but require more computational power. Embedded methods are learning models that carry out automatic feature selection. Embedded techniques are often based on model regularization [27]. Feature subsets selected via wrapper and embedded methods perform well on the base models, but may fail to generalize to other learning models. Subsets selected using filter methods perform more regularly over a range of different models.

2.1. **Filter Methods.** The most elementary approach in feature selection is feature filtering. In a filter method the degree of the relationship between a feature subset and the target variable is calculated based on a chosen measure. The resulting value represents the importance of the feature subset. In the simplest case, individual feature variables are ordered according to their metric values and the top ranked features are selected. However, such approach ignores feature interactions. There is a possibility that a pair of features that are irrelevant individually can have a high predictive power when combined together as in the case of the XOR problem. In [23, 33], the authors propose a method that attempts to handle feature interactions by evaluating features based on their relevancy with respect to the class variable and their redundancy with respect to other feature variables. The features that are highly correlated with the target variable but uncorrelated with other feature variables are given the preference. The authors in [7], suggest combining mutual information and χ^2 values into a ranked vector. The use of both mutual information and χ^2-statistic decreases the deviation of feature evaluation scores. The proposed method was effectively employed to select the relevant factors in autism classification [28]. Other measures of feature relevance have also been used in literature. In [6], the authors propose to evaluate feature subsets based on an inconsistency measure. In this approach a subset is deemed inconsistent if there are at least two instances with identical feature labels that belong to distinct classes. The authors discuss different search methods for finding the optimal subset based on the inconsistency measure. In a novel approach proposed in [8, 15] the authors use orthogonal variance decomposition to evaluate features. The orthogonality of the decomposition allows to directly calculate the total contribution of each feature to the output variance. As a result, they obtain a feature selection algorithm which takes into account feature interactions.

2.2. **Wrapper Methods.** Wrapper techniques utilize a base learning model on a feature subset and use the resulting prediction error as the subset score. The

difference between wrapper and filter methods is in the scoring process. While the former uses model accuracy, the later uses a statistical metric to score feature subsets. Wrapper methods are computationally intensive but often yield better results for a given learning model. To account for feature interactions, wrapper methods ought to be applied to subsets inplace of individual features which enlarges the search space. Recursive Feature Elimination (RFE) algorithm is at the intersection of filter and wrapper approaches. RFE is a simple iterative algorithm. At each iteration, the base model is trained and the weakest feature in the model is dropped. If that the dataset is normalized, then a model feature can be scored according to the value of its coefficient. For instance, in an ordinary least squares (OLS) regression model the absolute value of the coefficients represent their grandeur in the model. RFE is a widely used method and has implementations with many estimators including SVM [19], Random Forests [18], and MLP [32].

2.3. **Evolutionary Algorithms.** In recent years, evolutionary algorithms have become popular in feature selection [5, 31, 26]. An evolutionary algorithm (EA) is an optimization technique that follows evolutionary processes. An EA model starts with a population consisting of subsets of features called genes. Each subset is evaluated based on a fitness measure, usually classifier accuracy. Then a number of genetic operations such as selection, crossover and mutation are carried out on the population. This process is repeated several times (generations) until the stopping criterion is achieved. A number of other optimization techniques along the same lines have also been considered for feature selection. For instance, in an artificial bee colony, the selection process is based on the behavior of honey bee swarm Hancer. In [22], the authors employ the swarming behavior of grasshoppers to construct a feature selection method. An evolutionary approach based on the behavior of gray wolf packs was employed in [1] to choose the optimal feature subset. In another approach, a nested couple of genetic algorithms was applied to feature selection in cancer microarray data [25]. Despite some positive findings, these methods are computationally demanding. In addition, since EA models are stochastic the results can be unstable and convergence to the global minimum is not guaranteed.

2.4. **Penalized Interaction.** Penalized interaction approaches are used to select important interaction terms in nonlinear regression models. An important aspect of variable selection in interaction models is the penalty term in the cost function that follows the hierarchy of the importance of the features. An often utilized approach is to include an interaction term only after the corresponding main effects are present [35]. One of the popular approaches that uses hierarchical penalty is hierarchical Lasso. It is based on the traditional Lasso regression with the addition of weak hierarchy constraints between the main effects and the interaction effects [4, 21]. There exist a number of other techniques for hierarchical selection of variables include CAP [34] and VANISH [24]. Although penalized interaction algorithms are used to select both the main effects and interactions terms they are primarily designed for the latter.

2.5. **Principal Component Analysis.** Principal component analysis (PCA) is a dimensionality reduction algorithm that generates a new orthogonal basis that best captures the variance in a dataset. It is commonly used to reduce the dimension of the dataset by projecting each data point onto only the first few principal components to obtain lower-dimensional data. At the same time, it maximally preserves the variance of the dataset [2]. There are two key differences between PCA and feature selection. First, PCA is an unsupervised algorithm. It does not employ the class label information to construct the new basis vectors. Second, PCA is not a feature selection algorithm but instead a dimensionality reduction algorithm. PCA generates a set of new features which can be used to reduce the dimension of the dataset. It does not supply direct information about the significance of the original features.

2.6. **Feature Selection in Imbalanced Data.** Imbalanced data refers to skewed distribution of class labels [16]. Feature selection in the context of imbalanced data requires a careful approach. Since the number of samples in the majority class heavily out weight the number of samples in the minority class, typical feature selection methods may not be appropriate. For instance, mutual information based filter methods would unfairly be biased towards the majority class labels. At the same time, correct classification of the minority samples is often of greater importance. Therefore, the features that help correctly identify the minority samples are of added importance. To carry out feature selection in imbalanced data, macro metrics that consider unweighted aggregate outcomes are desirable.

3. Conclusion

Machine learning is a growing field of research with a wide range of applications including computer vision, speech recognition, medicine, astronomy, finance and many others [10, 11, 12]. Feature selection is an important preprocessing step in big data machine learning applications. In this paper, we reviewed the latest advancement is feature selection research. We highlighted the classic as well as the modern approaches. We discussed the benefits and drawbacks of each approach. In the future, we recommend taking into account feature interactions.

References

1. Al-Tashi, Q., Kadir, S. J. A., Rais, H. M., Mirjalili, S., & Alhussian, H. (2019). Binary optimization using hybrid grey wolf optimization for feature selection. IEEE Access, 7, 39496-39508.
2. Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.
3. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.
4. Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. Annals of statistics, 41(3), 1111.
5. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.
6. Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. Artificial intelligence, 151(1-2), 155-176.

7. Kamalov, F., & Thabtah, F. (2017). A feature selection method based on ranked vector scores of features for classification. Annals of Data Science, 4(4), 483-502.
8. Kamalov, F. (2018, December). Sensitivity analysis for feature selection. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1466-1470). IEEE.
9. Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 19(01), 2040013.
10. Kamalov, F., Gurrib, I., & Rajab, K. (2021). Financial Forecasting with Machine Learning: Price Vs Return. Journal of Computer Science, 17(3), 251-264.
11. Kamalov, F., Smail, L., & Gurrib, I. (2020, December). Forecasting with Deep Learning: S&P 500 index. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 422-425). IEEE.
12. Kamalov, F., Smail, L., & Gurrib, I. (2020, November). Stock price forecast with deep learning. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 1098-1102). IEEE.
13. Kamalov, F. (2020). Generalized feature similarity measure. Annals of mathematics and artificial intelligence, 88, 987-1002.
14. Kamalov, F., Moussa, S., Zgheib, R., & Mashaal, O. (2020, December). Feature selection for intrusion detection systems. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 265-269). IEEE.
15. Kamalov, F. (2021). Orthogonal variance decomposition based feature selection. Expert Systems with Applications, 182, 115191.
16. Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. Information Sciences, 512, 1192-1201.
17. Hancer, E., Xue, B., Zhang, M., Karaboga, D., & Akay, B. (2018). Pareto front feature selection based on artificial bee colony optimization. Information Sciences, 422, 462-479.
18. Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 83(2), 83-90.
19. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), 389-422.
20. Jin, X., Xu, A., Bie, R., & Guo, P. (2006, April). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In International Workshop on Data Mining for Biomedical Applications (pp. 106-115). Springer, Berlin, Heidelberg.
21. Liu, Y., Wang, J., & Ye, J. (2016). An efficient algorithm for weak hierarchical lasso. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(3), 1-24.
22. Mafarja, M., Aljarah, I., Heidari, A. A., Hammouri, A. I., Faris, H., AlaM, A. Z., & Mirjalili, S. (2018). Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems. Knowledge-Based Systems, 145, 25-45.
23. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.
24. Radchenko, P., & James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. Journal of the American Statistical Association, 105(492), 1541-1553.
25. Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications, 121, 233-243.
26. Suárez, R. R., Valencia-Ramírez, J. M., & Graff, M. (2014, November). Genetic programming as a feature selection algorithm. In 2014 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC) (pp. 1-5). IEEE.

27. Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 273-282.
28. Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. International journal of medical informatics, 117, 112-124.
29. Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least Loss: A simplified filter method for feature selection. Information Sciences, 534, 1-15.
30. Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. Neural computing and applications, 24(1), 175-186.
31. Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(4), 606-626.
32. Yang, J. B., Shen, K. Q., Ong, C. J., & Li, X. P. (2009). Feature selection for MLP neural network: The use of random permutation of probabilistic outputs. IEEE Transactions on Neural Networks, 20(12), 1911-1922.
33. Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5(Oct), 1205-1224.
34. Zhao, P., Rocha, G., & Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics, 37(6A), 3468-3497.
35. Zhao, J., & Leng, C. (2016). An analysis of penalized interaction models. Bernoulli, 22(3), 1937-1961.

[1]Faculty of Engineering, Tajik Technical University, Dushanbe, Tajikistan.
Email address: said.kamolov@ttu.tj