A survey on Feature Selection Techniques

Article · September 2020

CITATION
1

READS
612

1 author:

Nirali Honest
Charotar University of Science and Technology
16 PUBLICATIONS **37** CITATIONS

SEE PROFILE

A survey on Feature Selection Techniques

Nirali Honest

Faculty of Computer Science and Applications, CHARUSAT

Abstract

The feature selection techniques help to remove extra data and allows the model to focus only on the important features of the data, thus is reduces over fitting. By removing the less important features it increases the prediction accuracy and reduces the computation time to process the information. Reducing the model to only few features makes it easy to understand and interpret. This paper focus on feature selection techniques that could be helpful in identifying irrelevant features, remove them and improve the model performance.

Keywords: *Feature selection, Wrapper methods, Filter methods, Embedded methods, Performance parameters,*

1. Introduction

A feature is a variable in the dataset; it is most often present as a column in a dataset. In case of few features, the analysis and discovering of data is easier, but in case of more features it becomes difficult to process the data, so it is in that case feature selection methods come in picture. They allow to choose few but important features and reduce the number of features, without compromising with the resulting accuracy [1][2]. The applications of machine learning or big data analysis may include lots of features, which could become difficult to handle. Several techniques are developed to reduce the irrelevant and redundant features. The concept of feature selection helps in understanding data, reduce the variables to process which leads in improving accuracy [3]. The purpose of feature selection is to find the small portion of variables from the big dataset, that could help reduce computation time, and still increase the prediction accuracy. [4].

The relation among the features plays a major role in deciding what features to keep and what to remove from the dataset [17]. In case of dependent or correlated features only one feature is sufficient to describe the data, as the correlated features provide same information in deciding some class, so they act as noise data. It helps to reduce the dependent features and keep only unique features which contain the maximum information about the classes. When there is no relation among the features, some criterion has to be decided that can measure the relevance of each feature. If a system uses irrelevant variables, it will use this information for new data leading to poor generalization. Removing irrelevant variables must not be compared with other dimension reduction methods such as Principal Component Analysis (PCA) [5] since good features can be independent of the rest of the data [6]. Feature elimination does not create new features since it uses the input features itself to reduce their number. After choosing a feature selection criterion, a process of finding the important and useful features forming the subset of dataset must be carefully developed. Deciding the subset of features for a given data, directly becomes an NP-hard problem as the number of features grow high in number, thus a process must be used to remove the unnecessary data and form the subset of features.

There are different types of methods developed for feature selection. There are three types of feature selection techniques, Wrapper methods, Filter methods, and Embedded methods [7].

2. Wrapper methods

The main criterion of the feature selection in the wrapper methods is the measurement of the prediction accuracy. The core part in this selection technique is the search algorithm whose job is to find the subset of the dataset that gives the highest prediction performance. Initially the prediction is performed with a few specific subset of features and the performance is measured, to calculate the significance of each feature. This process is iterated with different subset of features until the optimal subset is achieved. There are two weaknesses this method, it takes large computation time for dataset having many features and it tends to overfit the model there is small dataset. The most common wrapper methods are forward selection, backward selection, and stepwise selection.

2.1 Forward selection

It starts with zero features, then, for each individual feature, runs a model and determines the p-value associated with the t-test or F-test performed. It then selects the feature with the lowest p-value and adds that to the working model [15][16]. Next, it takes the first feature selected and runs models with a second feature added and selects the second feature with the lowest p-value. Then it takes the two features previously selected and runs models with a third feature and so on, until all features that have significant p-values are added to the model. Any features that never had a significant p-value when tried in the iterations will be excluded from the final model.

2.2 Backward selection

It starts with all features contained in the dataset. It then runs a model and calculates a p-value associated with the t-test or F-test of the model for each feature. The feature with the largest insignificant p-value will then be removed from the model, and the process starts again. This continues until all features with insignificant p-values are removed from the model.

2.3 Stepwise selection

It is a hybrid of forward and backward selection. It starts with zero features and adds the first feature with the lowest significant p-value, then it goes through and finds the second feature with the lowest significant p-value. On the third iteration, it will look for the next feature with the lowest significant p-value, and it will also remove any features that were previously added that now have an insignificant p-value. This allows for the final model to have all of the features included be significant.

These selection methods can be used as a starting point, for selecting the important features from the dataset. It effectively selects the significant features from the large dataset, but all the features combination is not tested so it may not end up in the best possible model. Also it has inflated beta coefficients due to related features which is not good for predicting accuracy.

3. Filter methods

Filter methods are used to rank the features [10] while cleaning and processing the data set and the highly ranked features are selected [11] and applied to a predict the model outcome [8][9]. The errors are ignored during the cleaning and preprocessing phase; a subset of the features is selected through ranking them by a useful descriptive measure [12]. These techniques take less computation and will not over fit the data. The filters methods apply the essential properties of data to evaluate feature subsets [18][19][20]. But the major drawback with this technique is that it will not consider the correlations among the features. The most common filter methods include Pearson correlation, ANOVA and variance thresholding.

3.1 Pearson correlation

It is a measure of the similarity of two features that ranges between -1 and 1. A value close to 1 or -1 indicates that the two features have a high correlation and may be related. It can be used to create a model with reduced features, it can pick the features that have the highest correlation with the response predictor variable. The cutoff value of high correlation vs low correlation depends on the range of correlation coefficients within each dataset. A general measure of high correlation is $0.7 < |\text{correlation}| < 1.0$. This will allow the model that uses the features selected to encompass a majority of the valuable information contained in the dataset.

3.2 ANOVA

The Analysis of variance test looks at the variation within the treatments of a feature and also between the treatments. These variances are important metrics for this specific filtering method because we can determine whether a feature does a good job of accounting for variation in the dependent variable. If the variance within each specific treatment is larger than the variation between the treatments, then the feature hasn't done a good job of accounting for the variation in the dependent variable. To carry out an ANOVA test, an F statistic is computed for each individual feature with the variation between treatments in the numerator and the variation within treatments in the denominator. This test statistic is then tested against the null hypothesis (H_0 : Mean value is equal across all treatments) and the alternative (H_α : At least two treatments differ).

4 Embedded methods

Embedded methods [13][14] include variable selection as part of the training process without splitting the data into training and testing sets.

Embedded methods perform feature selection as a part of the model creation process. This generally leads to a happy medium between the two methods of feature selection previously explained, as the selection is done in conjunction with the model tuning process. Lasso and Ridge regression are the two most common feature selection methods of this type, and Decision tree also creates a model using different types of feature selection.

4.1 Ridge regression

Occasionally you may want to keep all the features in your final model, but you don't want the model to focus too much on any one coefficient, ridge regression can do this by penalizing the beta coefficients of a model for being too large. Basically, it scales

back the strength of correlation with variables that may not be as important as others. This takes care of any multicollinearity (relationships among features that will inflate their betas) that may be present in your data. Ride Regression is done by adding a penalty term (also called ridge estimator or shrinkage estimator) to the cost function of the regression. The penalty term takes all of the betas and scales them by a term lambda (λ) that must be tuned (usually with cross validation: compares the same model but with different values of lambda). Lambda is a value between 0 and infinity, although it is good to start with values between 0 and 1. The higher the value of lambda, the more the coefficients are shrunk. When lambda is equal to 0, the result will be a regular ordinary least squares model with no penalty.

4.2 Lasso Regression

It is another way to penalize the beta coefficients in a model, and is very similar to Ridge regression. It also adds a penalty term to the cost function of a model, with a lambda value that must be tuned. The most important distinction from Ridge regression is that Lasso Regression can force the Beta coefficient to zero, which will remove that feature from the model. This is why Lasso is preferred at times, especially when you are looking to reduce model complexity. The smaller number of features a model has, the lower the complexity. In order to force the coefficients to zero, the penalty term added to the cost function takes the absolute value of the beta terms instead of squaring it, which when trying to minimize the cost, can negate the rest of the function, leading to a beta equal to zero.

An important note for Ridge and Lasso regression is that all of your features must be standardized. Many functions in Python and R do this automatically, because the lambda must be applied equally to each feature. Having one feature with values in the thousands and another with decimal values will not allow this to happen, hence the standardization requirement.

4.3 Decision Tree

Another common way to model data with feature selection is called Decision Tree, which can either be a regression tree or classification tree depending on whether the response variable is continuous or discrete, respectively. This method creates splits in the tree based on certain features to create an algorithm to find the correct response variable. The way the tree is built uses a wrapper method inside an embedded method. What we mean by that is, when making the tree model, the function has several feature selection methods built into it. At each split, the function used to create the tree tries all possible splits for all the features and chooses the one that splits the data into the most homogenous groups. In plain terms, it chooses the feature that can best predict what the response variable will be at each point in the tree. This is a wrapper method since it tries all possible combinations of features and then picks the best one.

The most important features in predicting the response variable are used to make splits near the root (start) of the tree, and the more irrelevant features aren't used to make splits until near the nodes of the tree (ends). In this way, decision tree penalizes features that are not helpful in predicting the response variable (embedded method). After a tree has been made, there is an option to go back and 'prune' some of the nodes that do not provide any additional information to the model. This prevents overfitting, and is usually done through cross validation with a holdout test set.

Conclusion

To conclude the process of feature selection , one must first understand the feature, it is most often a column in a dataset

then optimizing a model by selecting a subset of the features to use. This can be done by applying Wrapper methods in which models with different subsets of features are picked with the best combination. For example in forward selection adding features one by one to reach the optimal model, in backward selection features are removed one by one to reach the optimal model and in stepwise selection combination of forward and backward selection is performed so the features can be added and/or removed one by one to reach the optimal model. In applying filtering, a subset of features is selected by a measure other than error, for example pearson correlation is used to a measure of the linear correlation between two variables, variance thresholding is used to select the features above a variance cutoff to preserve most of the information from the data. Analysis of variance a group of statistical estimation procedures and models that is used to observe differences in sample means; can be used to tell when a feature is statistically significant to a model. Interacting term quantifies the relationship between two of the features when they depend on the value of the other. Multicollinearity occurs when two or more independent variables are highly correlated with each other. Embedded method are used for selecting and tuning the subset of features during the model creation process for example ridge regression is a modified least squares regression that penalizes features for having inflated beta coefficients by applying a lambda term to the cost function, lasso regression is similar to ridge regression, but different in that the lambda term added to the cost function can force a beta coefficient to zero, and decision tree is a non-parametric model that uses features as nodes to split samples to correctly classify an observation. After understanding the feature selection techniques one can apply it based on the application domain and data set.

Acknowledgments

I would like to thank Charotar University of Science and Technology for providing all the resources to carry out the work.

References

[1] H. Motoda and H. Liu, “Feature selection, extraction and construction” In: Towards the Foundation of Data Mining Workshop, Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002), Taipei, Taiwan, pp. 67–72, 2002.

[2] N. Chumerin and V. Hulle, M. M, “Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information” In: Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, pp. 343–348, 2006.

[3] L. Ladla and T. Deepa, “ Feature Selection Methods And Algorithms”, International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.

[4] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[5] Alpaydin E. Introduction to machine learning. The MIT Press; 2004.

[6] Law MH, Figueiredo M rio AT, Jain AK. Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 2004;26.

[7] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

[8] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997;97:245–70.

[9] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Proc 11th int conf mach learn; 1994. p. 121–9.

[10] Caruana R, de S V. Benefitting from the variables that variable selection discards. J Mach Learn Res 2003;3:1245–64.

[11] [20] Koller D, Sahami M. Towards optimal feature selection. In: ICML, vol. 96; 1996. p. 284–92.

[12] [21] Davidson JL, Jalan J. Feature selection for steganalysis using the mahalonobis distance. In: Proc SPIE 7541, Media Forensics and Security II 7541; 2010.

[13] Langley P. Selection of relevant features in machine learning. In: AAAI fall symp relevance; 1994.

[14] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997;97:245–70.

[15] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection. Pattern Recog Lett 1994;15:1119–25.

[16] [34] Reunanen J. Overfitting in making comparisons between variable selection methods. J Mach Learn Res 2003;3:1371–82.

[17] A. G. K. Janecek and G. F. Gansterer et al, “On the Relationship between Feature Selection and Classification Accuracy”, In: Proceeding of New Challenges for Feature Selection, pp. 40-105, 2008.

[18] Yu, L. and Liu, H. Feature Selection for High-Dimensional Data: A Fast CorrelationBased Filter Solution. Proc. 20th Int’l Conf. Machine Learning, pp. 856-863, 2003

[19] Hall, M. A. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. Proc. 17th Int’l Conf. Machine Learning, pp. 359-366, 2000

[20] Dash, M., Choi, K., Scheuermann, P. and Liu, H. Feature Selection for Clustering - A Filter Solution. in Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02), pp. 115-122, 2002