A REVIEW: FEATURE SELECTION METHODS FOR TEXT CLASSIFICATION

Nijeesha Joseph
M Tech Research Scholar
School of Computer Sciences
Mahatma Gandhi University, Kottayam, India.

Abstract : Feature selection is the process of eliminating the irrelevant and redundant data from a large dataset. Feature selection mainly three types. They are filter method, Wrapper method and embedded method. Filter method and the wrapper method are the two primary feature selection methods. Feature selection is the preprocessing step in machine learning. In feature selection picks subset from large dataset and makes an efficient model for describing the selected subset. Other than selecting the subset, it also have some other advantages, such as dimensionality reduction, adjust the amount of data which are required for learning process and progress in predictive accuracy. The main aim of this work is to get more idea about the concept of feature selection and various feature selection methods.

IndexTerms **- Filter method; Wrapper method; Embedded method.**

I. INTRODUCTION

Machine learning concept is working according to a simple rule: if you place the trash, you will only get the trash out. Trash means noise. The common problem in machine learning is selecting a group of relevant features for constructing a good classification model. The most important function of feature selection is avoiding inappropriate and redundant data that facilitates improving the performance of learning algorithms. Feature selection is a dynamic and productive research field such as machine learning area, data mining and pattern recognition. Feature selection is the process of selecting relevant data and avoid unwanted data that are not improves the performance of the final model. Feature selection is also called variable selection or attributes selection. In feature selection method it only selects the important and useful data from the large dataset. One of the important methods for improving the performance of the classification algorithm is done by feature selection. The proper Feature selection method helps to increase the accuracy of the classification model. Using fewer data, by avoiding redundant and irrelevant data reduces the complexity of the final classification model that is very simple and very easy to understand.

Main objectives of variable selection or feature selection are,

- Provide a better understanding of the underlying process that generated the data.
- Provide faster and more profitable predictors.
- To improve the prediction performance of the predictors.

Feature selection is considered in here is because of the following specified reasons, which includes,

- Fast training process
- To obtain high accuracy
- To reduce the complexity

II. FEATURE SELECTION METHODS

The feature selection method collects group of subset of features and analysis the performance of the features by using machine learning algorithms. Main feature selection methods are,

- Filter Method
- Wrapper Method
- Embedded Method

FILTER METHOD

Filter methods are used in most cases as a preprocessing step. The selection of features is not dependent on any machine learning algorithm. The features are selected based on their values in several statistical tests. The feature selection is based on the concepts of correlation between each feature. The features are classified according to the scores that can be getting from different statistical tests. Then the selected features are to be saved or deleted from the dataset according to its score.

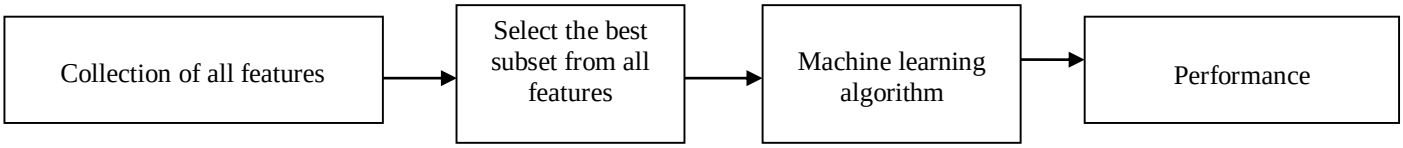

Figure1. Filter Method

From the given collection of large dataset, select all important features. Then choose best subset of features for applying machine learning algorithm. Then compute the performance of the classification model. Filter method cannot get any performance feedback.

WRAPPER METHOD

The wrapper method is also known as the feedback method. It always uses a feedback from the previous performance output. In wrapper methods consider several subsets from total collection of attributes. And evaluate the subset by machine learning algorithm. Then the result is given to the input of next subset. Then iteratively checks all the subsets. Then find the best subset from all the attributes. By using this subset build a model. Then assign a score based on the accuracy of that model. The wrapper method is very expensive. The common examples of wrapper methods are,

Forward Selection: Forward selection is an iterative method. At the starting time it does not have any features. During each iteration it adds features that help to improve the performance of the model. And avoids the features that decrease the performance of the model. This process is continuous till find a feature that decreases the performance of the model.

Backward Elimination: In backward elimination, start with all collection of the features and eliminate the least significant feature, that improves the performance of the final model. Repeat this method till there's no improvement discovered on removal of features.

Recursive Feature Elimination: It is a kind of greedy improvement rule, its main aim is to seek out the simplest subset from a bunch of subsets. That subset provides more performance. Then it repeatedly creates models and neglects the simplest or the worst performance features in every iteration. Build subsequent model with the left features. This method is dispensed till all features are exhausted. Then classify the features in keeping with the order of their removal.

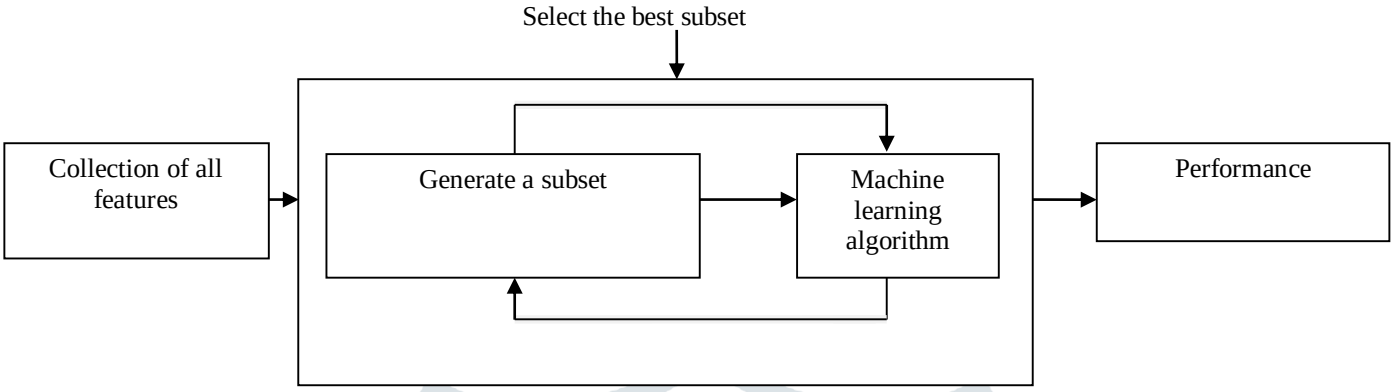

Figure2. Wrapper Method

COMPARISON BETWEEN FILTER AND WRAPPER METHOD

There are several differences between filter method and wrapper method for feature selection.

- Wrapper method is very expensive compared to filter method. Filter method is faster than the Wrapper method. Because filter method does not train any model while the wrapper method train the model. So the wrapping method is very expensive.
- The wrapper method is uses cross validation technique for evaluate the subset of features; while the filter method uses statistical methods for evaluation of subset of features.
- The use of the subset of characteristics of the wrapper methods makes the model more prone to over fitting compared to the use of the subset of characteristics of the filter method.
- Wrapper method is always finding the best subset of features from the all collections of the features. But the filter method never finds any subset from all the features.
- The filter method analysis the importance of the features by their correlation with the dependent variable, but in the wrapper method analysis the usefulness of a subset of features when actually training a model in it.

EMBEDDED METHOD

Embedded method is the combination of filter method and wrapper method. Because it uses the qualities of both filter method and the wrapper method. Embedded method functioning with several algorithms that contains built in functions. Feature selection is carried out in the training phase of the classification. The working of embedded method is little tough compared to filter method and wrapper method. Because from the collection of all features it selects subset of relevant features. Then that subset is analysed by a machine learning algorithm and finds its performance. Then iteratively check the performance of all relevant subsets. Finally it takes a best subset that has high performance. Regularization method is the most common embedded feature selection method. The alternative name of regularization method is the penalty method. This method is help to reduce the complexity of the model.

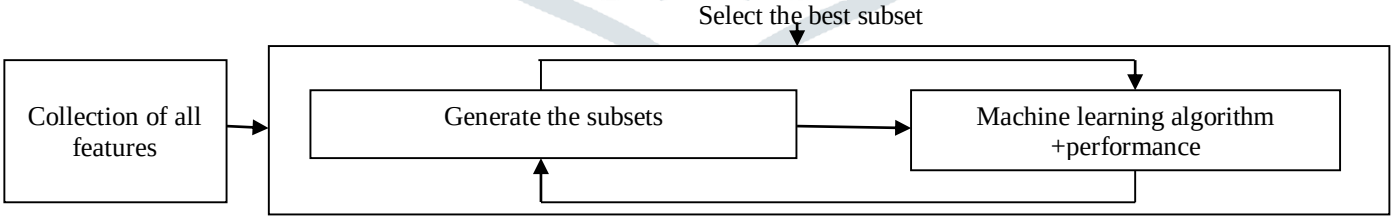

Figure3. Embedded Method

III. CONCLUSION

The main aim of feature selection is to select the relevant subset and avoid irrelevant and redundant information from the large dataset. This paper mainly focused on basic concept of feature selection and different feature selection methods. The feature selection methods are flexible and capable of providing a solution to any kind of problem faced when performing feature selection. Instead of using filter and wrapper methods separately, use embedded method, improves the classification accuracy, increase the speed and also reduce the error rate.

IV. ACKNOWLEDGMENT

My thanks to the Guide, Ms. Sunitha C S Centre Head of Centre for Development and Advanced Computing (C-DAC) Kochi and Chinju K Assistant Professor at School Of Computer Sciences, Mahatma Gandhi University gives me the courage, enthusiasm, comments, suggestion and positive feedback during the presentation of this paper.

REFERENCES

[1] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2014, pp. 1-6.

[2] M. B. Çatalkaya, O. Kalıpsız, M. S. Aktaş and U. O. Turgut, "Data Feature Selection Methods on Distributed Big Data Processing Platforms," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018, pp. 133-138.

[3] Yun Jiang, Xi Liu, Guolei Yan, Jize Xiao "Modified Binary Cuckoo Search for Feature Selection: A Hybrid Filter-Wrapper Approach" 13th International Conference on Computational Intelligence and Security-2017

[4] Ms. Priyanka Patel, Ms. Khushali Mistry "A Review: Text Classification on Social Media Data" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. IV (Jan – Feb. 2015), PP 80-84 www.iosrjournals.org

[5] H. Zhang, Y. Ren and X. Yang, "Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree," 2013 10th Web Information System and Application Conference, Yangzhou, 2013, pp. 446-449.