Feature Selection Using Ensemble Techniques

Yash Kaushik(✉), Muskaan Dixit, Nikhil Sharma, and Monika Garg

Manav Rachna International Institute of Research and Studies, Faridabad, India
monikagarg.fet@mriu.edu.in

Abstract. Data used in Machine Learning tasks need to be pre-processed and prepared to improve its quality. Features or variables in data play a major role in the results obtained after applying Machine Learning models. The features which are irrelevant to the domain should be discarded with the objective of improving the accuracy and validity of results. For this purpose, Feature Selection is used. It is a way of reducing the size, the purpose of settling down to the right element from the original elements by removing the negative, redundant or noisy features. Feature selection can often result in better learning performance, such as, lower computational cost, and better model translation. It is very important to shed some light on the nature of feature selection for student performance measurement, because constructive educational approaches can be found in the appropriate set of features. Feature selection plays a major role in refining the quality of the data models. Increased data quality can produce better results and therefore based options for such quality data can increase the quality of education by predicting performance. In the light of the aforementioned fact, it is necessary to carefully stabilize the selection of the algorithm. Feature selection key can directly affect classification accuracy and simplify operation. Several data experiments were performed to demonstrate the effectiveness of the proposed method. Selective bias may also be the term used in data analytics. In this paper, two ensemble techniques namely, Random Forests and Gradient Boosting Machines have been applied on a dataset, for the purpose of Feature Selection. Experimental results show that Gradient Boosting Machines are better at Feature Selection.

Keywords: Feature selection · Random forests · Gradient boosting machines

1 Introduction

Over the past decade, the enthusiasm for using Feature Selection (FS) techniques has shifted from being a model that is proving to be a real catalyst for model building. In order to use machine learning methods effectively, pre-processing the data is essential. Feature selection is one of the most common and important techniques for early data processing, and it has become a very important part of the machine learning process. It is the process of finding the right features and removing inaccurate, unwanted, or noisy data. This process speeds up data mining algorithms, improves the precision of the guesswork, and increases the accuracy. Our interest is concentrated on high quality data. The sheer amount of high-quality data has posed the greatest challenge to existing

P. K. Singh et al. (Eds.): FTNCT 2020, CCIS 1395, pp. 288–298, 2021.
https://doi.org/10.1007/978-981-16-1480-4_25

machine operating methods. Depending on the supervised learning, Feature Selection provides a set of features for designers using one of the following methods. One is, specified dimensions for support features that maximize test scope, Another is Overall, the bottom set is the best commitment between size and test score. The method of choice (fs) method has changed dramatically over the last few years and, many papers have found domains with (hundreds to tens of thousands of variables). New approaches are being proposed to deal with these stimulating tasks that include many negative and abnormal variables and often the same few examples of training. 1. We are looking for "variable" variable input variables and, "features" built-in input variables. We use null "variables" and "element" where there is no view in selection algorithms, e.g. Text partitioning problem, represented documents, which is an indication of the size of the terminology consisting of word counting, view and understanding of data, decrease in size and storage requirements, reduce training times and use, reduce the size curse to improve predictive performance. The feature selection problem based on supervised instruction is: Given a set of candidate features that selects a standard set defined by one of three methods. A set-size setup that enables an experimenter with a small size satisfies a specific limit on a test scale and is possible to better understand the results obtained by the inducer, reducing its storage capacity. The indirect feature does not apply to import, but not all relevant features are useful. The selection criteria included are that the time spent in training the model is greatly reduced due to the limited number of parameters used.

Authors of a paper [1] give information about robustness of feature selection techniques which needs importance while analyzing the selected feature subsets. This shows that these techniques are of useful for higher dimensional domains and small sample sizes. In addition, they also investigated the effects of integrating feature selection strategies on categorization, providing a new strategy for model selection.

Another paper [2] lets us consider about the contingency of feature selection, which might provide us with a basic classification in hierarchical system of feature selection techniques, and communicate its use, variety and introduction of a few important applications and future bioinformatics applications. This paper also provides taxonomy of these techniques there were areas of application and their diversification in common and the new coming bioinformatics applications.

This paper is organized as follows: Sect. 1 introduces the concept of Feature Selection. Need of Feature Selection is highlighted in Sect. 2. The various categories of Feature selection techniques are explained in Sect. 3. Section 4 surveys the literature. The methodology adopted in this article is explained in Sect. 5. Experiments performed and results obtained are discussed in Sect. 6. Paper is concluded in Sect. 7.

2 Need of Feature Selection

Feature selection as the name describes is to pick out important features from a specific data set or in simple terms to cut back number of input variables while developing a predictive model and is additionally referred to as variable selection. Nowadays, datasets have abundant information with data which is collected using countless of IoT devices and sensors. Now the times datasets have huge number of attributes in which not all are

useful a large number of features make a model bulky, time-taking, high dimensional and harder to implement in production.

Feature selection is vital as there is noisy data which needs to be removed. Lots of low frequent features, multi-type features, too many features compared to samples, complex models, samples in real scenario is non-homogenous with training and test samples. All these points make it very important for data to be cleaned properly. Data cleaning is the most important and tedious work in data mining. If the data is not cleaned and arranged properly it will take much more time to get the desired results. It also reduces the computation time involved to get the model and helps us to focus on what is more important. It reduces the risk of Over-fitting that is also known as curse of dimensionality the dimensionality of the features space increases, number of configurations can grow exponentially which becomes very difficult if we have huge amount of data. It improves Accuracy which leads to less misleading data means and thus improving accuracy, data training time is reduced so that less data means that algorithms train faster. Feature selection also reduce the computational cost of modelling [3].

Seeing all the points we realize how much important it is to select the correct features accurately. One feature may not seem as important but it can be related to some other attribute too. Feature Selection methods helps with these problems by reducing the dimensions, reducing redundancy, reducing noisy data without much loss of the total information. By applying feature selection methods we get to know the significance of each feature in improving the model.

Feature selection can be done manually by data analysts but when the amount of attributes are in enormous amount it is not feasible for a person to do and it might take longer duration of time to complete a given project. Here comes the need for various feature selection methods.

3 Feature Selection Techniques

Feature Selection techniques can be categorized into following categories. Techniques which need to be applied depends upon the type on the type of data whether it is numerical or categorical.

3.1 Filter Methods

Filter methods are not dependent on any learning algorithm, but depend on the complete aspects of the training data. Filter-based feature selection methods [4] use statistical measures to check the correlation or dependency between the given input variables that how much are they dependent on each other that can be filtered to choose the most relevant features. Statistical measures for feature selection must be chosen carefully based on the data type of the input variable and the output or response variable. In filter methods, features are selected by selecting the most relevant attributes without using any machine learning algorithm. Various statistical tests are done like chi square test, ANOVA test, LDA, information gain, fisher score, correlation coefficient, variance threshold etc. and features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Here correlation is considered as a subjective term.

3.2 Wrapper Methods

Wrapper methods use learning algorithms to evaluate the features. Wrapper methods [5] are of various kinds too and we try to use subset of features and train a data from it. These processes have large number of steps and we also need to consider which method is to applied when depending upon our input data and the type of result e get whether it is numerical or categorical data.

3.3 Hybrid Methods

Hybrid methods combine properties of both Filter and Wrapper methods.

3.4 Embedded Methods

There is also another technique referred to as embedded technique of feature selection and is most advanced one. In our paper we have applied two different embedded techniques only. The difference between the filter and wrapper method is that Wrapper methods measure the "usefulness" of features supported on the classifier performance. On the opposite hand, the filter methods focus and acquire the intrinsic properties of the features measured via univariate statistics instead of cross-validation performance. There are different algorithms in java, python, R. Each one has its importance and some disadvantages like few feature selection techniques can be applied only on data having not large number of attributes or in some cases the accuracy is increased if there are larger number of variables.

In this paper, we have applied two Feature selection algorithms, namely Gradient Boosting Machines (GBM) and Random Forests (RF) [6]. These two techniques are described as follows:

3.4.1 Gradient Boosting Machines

In Gradient Boosting [7, 8], additive regression models are built by iteratively fitting a simple base learner to currently updated pseudo-residuals by applying least squares at every further iteration. The objective of this method is to evolve a function F*(x) which maps x to y, so that when the joint distribution of all values (y, x) is taken, the expected value of $\Psi(y, F(x))$ which is some specified loss function is minimized. This relation is depicted in Eq. (1).

Where y is the random output or dependent variable and $x = \{x_1, x2, \ldots x_n\}$ is a set of random input variables.

$$F^*(x) = \arg\min E_{y,x} \Psi(y, F(x)) \tag{1}$$

3.4.2 Random Forests

Random Forests [9–11] were developed as an extension to the popular ensemble technique called Bagging. It is a tree-based ensemble technique where each tree depends on

a set of random variables. Random Forests are used for prediction in various domains like Traffic estimation and congestion on roads due to traffic [12].

Equation (2) shows the mean square generalization error in Random Forests for a numeric predictor h(x)

$$E_{X,Y}(Y - h(X))^2 \tag{2}$$

The Random Forest predictor is constructed by taking the mean over k of the trees $\{h(x, \theta_k)\}$.

Equation (3) states the case for infinite numbers of trees in the forest

$$E_{X,Y}(Y - av_k h(X, \theta_k))^2 \rightarrow E_{X,Y}(Y - E_\theta h(X, \theta))^2 \tag{3}$$

4 Literature Survey

This section reviews the work done by some researchers.

This paper [13] introduces a new feature selection algorithm based on the wrapper process using neural networks. An important feature of this algorithm is the automatic determination of neural network structures during the process. Their algorithm uses a constructive approach that incorporates linking information in selecting features and determining neural network structures. The test results show the essence of a constructive approach to selecting features with integrated structure.

In this work [14], a propose a method of selecting a multi-filter element based on an method that includes the issuance of four filtering methods to achieve optimal selection. Then they also conducted a detailed evaluation of our proposed method using a benchmark acquisition dataset for intrusion detection, NSL-KDD and decision-making tree planning. The findings show that our proposed approach can effectively reduce the number of features from 41 to 13 and has a higher level of accuracy and classification accuracy compared to other classification strategies.

Recent advances [4] in computer technology in terms of speed, cost, and acquisition of large amounts of computer power and the ability to process large amounts of data in a timely manner have stimulated interest in increasing data mining requests to extract useful information from data. Machine learning has become one of the most widely used methods in these data mining applications. In this study, the authors examined the various options between classes and selected methods of using distance in relation to their effectiveness in advancing inclusion data to attract decision trees. The results of their research showed that intermediate-stage measures lead to better performance compared to expected measures, in general.

In this paper [15], writers have presented a review of the state of the art approach to the selection of scientific-theoretic features. The concepts of the importance of feature importance and redundancy are well defined, as well as Markov's outfit. The problem with the right feature selection is explained. An integrated theoretical framework has been developed, which can re-establish successful success strategies, reflecting the limitations made in each case. There are many open-ended issues in the field that are also being presented.

The authors in this paper [16] applied three Machine Learning techniques namely-Support Vector Machines, Multilayer Perceptron and Random Forests to predict the residency of teachers in Indian universities related to Information and Communication Technology awareness. Results showed that Random Forests was far better at prediction than other two algorithms with prediction accuracy of 72.8% and prediction time of 0.12 s.

Feature Selection is an important phase in Data Analysis and Machine Learning. In a paper [17], the authors have applied Machine Learning algorithms to analyze few categories of Cyber attacks.

In another paper [18], Feature selection is performed to select genes of Microarray datasets, with the objective of applying it for cancer diagnosis. Then Support vector machines technique is applied for classification purpose.

A paper [19] introduces the concept of variable and Feature selection and its objectives. The article emphasizes on various aspects like Feature construction, Feature ranking, multivariate Feature selection and some assessment methods to validate features.

5 Methodology

The research outline adopted in this paper is defined in Fig. 1. It consists of several phases which are explained as follows.

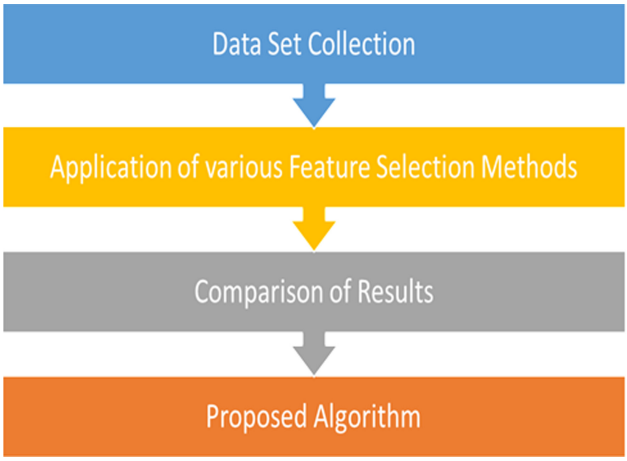

Fig. 1. Methodology

5.1 Dataset Collection

Dataset of store sales was collected and its parameters are described below in Table 1:

Table 1. Dataset Parameter Description

Parameters	Description
Shipping.Cost	This parameter determines the shipping cost used for shipping items to consumers
Order.Quantity	It is the number of units to be added in order to reduce inventory costs
Product.Category	It is all the products offering the same general functionality
Product.Container	This parameter is the art of incorporating or protecting a product for distribution, sale and use
Order.ID	is a unique **number** which you'll need to identify and track your **orders**
Product.Sub.Category	It is with respect to a given category, a sub-category while the category is a group
Ship.Date	Shipping date is the date the order is sent from the merchant or the last store to the customer
Order.Date	Order Date means the date on which the decision and final order of the Decision is issued by the Commission
Ship.Mode	Shipping Mode is a shipment name that defines travel mode
Province	Province is the main division of land or sovereignty
Order.Priority	The term used in fulfilling the practices that place certain submissions before others
Customer.Segment	We divide customers into groups based on general characteristics so that companies can sell to each group effectively and efficiently

5.2 Algorithm Application

As aforementioned, Random Forests and Gradient Boosting Machines were applied on the collected dataset for Feature Selection.

5.3 Results Comparison

The results of feature selection obtained from both the algorithms, RF and GBM were compared.

5.4 Proposed Algorithm

Since the results of GBM are better as compared to Random Forests, we propose GBM for feature selection.

6 Experiments and Results

In this paper, Random Forests and Gradient Boosting Machines were applied on the dataset with the objective of Feature Selection. The experiments have been performed using R Studio. Table 2 shows the results obtained from both algorithms.

Table 2. Feature Importance given by RF and GBM

Random forests	Gradient boosting machines
Order.ID	Shipping.Cost
Order.Date	Order.Quantity
Order.Priority	Product.Category
Order.Quantity	Product.Container
Ship.Mode	Order.ID
Shipping.Cost	Product.Sub.Category
Province	Ship.Date
Customer.Segment	Order.Date
Product.Category	Ship.Mode
Product.Sub.Category	Province
Product.Container	Order.Priority
Ship.Date	Customer.Segment

It is evident from the results shown in Table 2 that GBM is better at feature selection and importance as compared to RF. Therefore, we propose GBM for the purpose of Feature Selection.

Table 2 shows the priority order of the parameters which are responsible for Store sales by both - RF method and GBM method. GBM method is more accurate as compared to RF method. While using GBM method we were able to arrange the data in the order which is based on the priority that the parameter with the maximum value is at the top i.e. Shipping.Cost and the parameter with the least value is placed in the bottom i.e. Customer.Segment. GBM method gives the result with the values and by using these values we are also able to produce the Feature Importance graph. This graph is more visually understandable and we are also able to decide that which parameter is more important and is more responsible for the sales in store.

Table 3 shows the percentage assigned by GBM to each parameter in sequence.

Table 3. Feature importance percentage by GBM

Gradient boosting machines	Percentage
Shipping.Cost	38.81237
Order.Quantity	15.8415
Product.Category	8.672646
Product.Container	7.289848

(*continued*)

Table 3. (*continued*)

Gradient boosting machines	Percentage
Order.ID	6.758149
Product.Sub.Category	5.424211
Ship.Date	4.711707
Order.Date	4.274776
Ship.Mode	3.511199
Province	2.477245
Order.Priority	1.326632
Customer.Segment	0.899724

Figure 2 shows the various features, ordered as per importance using GBM technique. It clearly states that the parameter Shipping.Cost has the highest value from all other parameters.

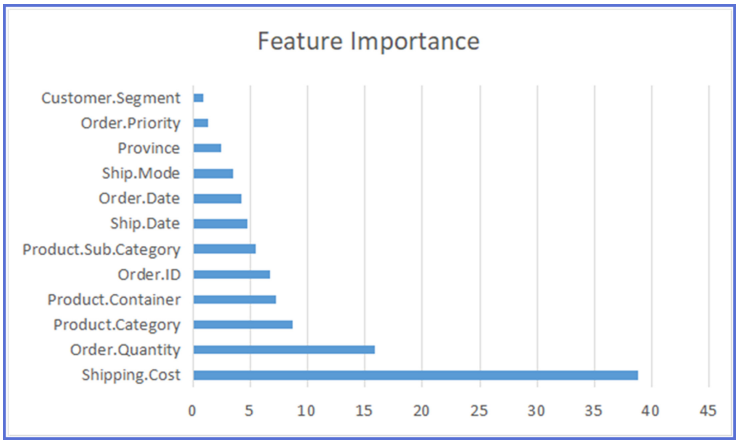

Fig. 2. Feature importance using GBM

7 Conclusion and Future Scope

This paper presents a survey of the feature selection methods expected in the literature. A few popular feature selection methods like Filter method, Wrapper method, Embedded methods were introduced. From embedded feature selection category, Random Forests and Gradient Boosting Machines were applied on a dataset for selecting important features. Experiments were performed in R Studio and the obtained results show that GBM performs better than RF in terms of Feature Selection. In future, other well-known techniques for Feature Selection can be experimented.

References

1. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87481-2_21
2. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
3. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
4. Osanaiye, O., Cai, H., Choo, K.-K., Dehghantanha, A., Xu, Z., Dlodlo, M.: Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. EURASIP J. Wirel. Commun. Netw. **2016**(1), 1 (2016). https://doi.org/10.1186/s13638-016-0623-3
5. Jović, A., Brkić, K., Bogunović, N.: A review of feature selection methods with applications. In: 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200–1205. IEEE (2015)
6. Goyal, M., Pandey, M.: Towards prediction of energy consumption of HVAC plants using machine learning. In: Batra, U., Roy, N.R., Panda, B. (eds.) REDSET 2019. CCIS, vol. 1229, pp. 254–265. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-5827-6_22
7. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
8. Goyal, M., Pandey, M.: Extreme gradient boosting algorithm for energy optimization in buildings pertaining to HVAC plants. EW, EAI (2020). https://doi.org/10.4108/eai.13-7-2018.164562
9. Prasad, A.M., Iverson, L.R., Liaw, A.: Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems **9**(2), 181–199 (2006)
10. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
11. Cutler, A., Cutler, D.R., Stevens, J.R.: Random forests. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning. Springer, Boston, MA (2012). https://doi.org/https://doi.org/10.1007/978-1-4419-9326-7_5
12. Khanna, A., Goyal, R., Verma, M., Joshi, D.: Intelligent traffic management system for smart cities. In: Singh, P.K., Paprzycki, M., Bhargava, B., Chhabra, J.K., Kaushal, N.C., Kumar, Y. (eds.) FTNCT 2018. CCIS, vol. 958, pp. 152–164. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-3804-5_12
13. Kabir, M.M., Islam, M.M., Murase, K.: A new wrapper feature selection approach using neural network. Neurocomputing **73**(16–18), 3273–3283 (2010)
14. Piramuthu, S.: Evaluating feature selection methods for learning in data mining applications. Eur. J. Oper. Res. **156**(2), 483–494 (2004)
15. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**(1), 175–186 (2013). https://doi.org/10.1007/s00521-013-1368-0
16. Verma, C., Illés, Z., Stoffová, V.: Predictive modeling to predict the residency of teachers using machine learning for the real-time. In: Singh, P.K., Sood, S., Kumar, Y., Paprzycki, M., Pljonkin, A., Hong, W.-C. (eds.) FTNCT 2019. CCIS, vol. 1206, pp. 592–601. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-4451-4_47
17. Malhotra, H., Dave, M., Lamba, T.: Security analysis of cyber attacks using machine learning algorithms in eGovernance projects. In: Singh, P.K., Sood, S., Kumar, Y., Paprzycki, M., Pljonkin, A., Hong, W.-C. (eds.) FTNCT 2019. CCIS, vol. 1206, pp. 662–672. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-4451-4_52
18. García-Nieto, J., Alba, E., Jourdan, L., Talbi, E.: Sensitivity and specificity based multiobjective approach for feature selection: application to cancer diagnosis. Inf. Process. Lett. **109**(16), 887–896 (2009)

19. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(Mar), 1157–1182 (2003)
20. Singh, P., Paprzycki, M., Bhargava, B., Chhabra, J., Kaushal, N., Kumar, Y.: Futuristic trends in network and communication technologies. FTNCT 2018. Communications in Computer and Information Science **958**, 141–166 (2018)
21. Singh, P., Sood, S., Kumar, Y., Paprzycki, M., Pljonkin, A., Hong, W.C.: Futuristic trends in networks and computing technologies. FTNCT. Commun. Comput. Inf. Sci. **1206**, 3–707 (2019)