Application of Feature Selection and Random Forest Classifier on a Medical Dataset

Aayush Kamath
UG Student
Department of Information Technology
Sardar Patel Institute of Technology, Mumbai-58, India

Gokul Nambiar
UG Student
Department of Information Technology
Sardar Patel Institute of Technology, Mumbai-58, India

Mohammad Izhan
UG Student
Department of Information Technology
Sardar Patel Institute of Technology, Mumbai-58, India

Radha Shankarmani
Professor
Department of Information Technology
Sardar Patel Institute of Technology, Mumbai-58, India

Abstract

For a given medical data set, there is a huge possibility that the data includes hundreds of features, each representing a symptom or a parameter based on which diagnosis can be carried out. While a lot of these features contribute towards the results, it is often the case that quite a few of these features turn out to be either irrelevant or have very little bearing in terms of their overall impact on the results and only end up crowding the data set. Feature selection provides a solution to this problem as the features that provide the highest contribution while predicting an output are retained and the irrelevant features are identified and subsequently eliminated. This helps in the model being trained faster and leads to a better interpretation of the model further allowing better diagnosis of the disease. Apart from feature selection, random forest classifier is being used as a means to predict the outcomes. Since random forest is made up of decision trees, it helps in better classification for a given problem.
Keywords: Feature Selection, Random Forest, Overfitting, Medical Dataset, Classification

I. INTRODUCTION

[1] Machine learning can be applied to a wide variety of fields and there is always scope to come up with great insights and solutions for a given field with the help of machine learning. The approach to a solution naturally varies, based on the field of application and the type of problem at hand. The medical field is one such area that has greatly benefited from machine learning. Since medicine is a field of gargantuan proportions, the applications in the field also exist in a wide range. This wide range of applications include enhancement in the ways in which health related data is managed, predicting illnesses at earlier stages by identifying various disease markers and patterns that pose a health risk, improving the accuracy of the diagnosis for a variety of diseases, etc. [4]While the application may vary, the common goal for using machine learning in the field of medicine, most of the times, is to provide better health care at a lower cost or/and better performance and accuracy. [4] This paper deals with the aspect of large medical datasets that contain a multitude of attributes and need a reduction in the number of attributes as some of them do not contribute much towards the resulting outcome. Since the output variable of the dataset in question concerns itself with prognosis, the problem comes under the umbrella of a classification problem. This is because the diseases covered by the prognosis attribute are to be classified into different categories. With way too many columns dictating the prediction of a disease, the processing speed and efficiency of the model is bound to be less than optimum. [1] Feature selection helps in getting rid of attributes of less importance and makes the dataset more compact and precise. The seeming fixation with lesser number of attributes can be justified by the fact that the complexity of the model is reduced as a result of feature selection which makes it easier for the model to be interpreted and explained. Moreover, feature selection provides a solution for the problem of overfitting. Overfitting is an issue that arises when the model learns about more data than is necessary, i.e., the noise or the unrelated data is identified by the model as necessary data that is to be incorporated in the learning along with the useful data. This leads to inaccuracy in predictions as the factors that do not have an impact on the outcome are believed to be of importance. [3] There are various methods for applying feature selection but these methods can be classified into three categories: filter, wrapper and embedded. Filter feature selection methods such as chi-squared test and information gain assign a score to each feature through statistical means and based on a certain value or threshold the features are either selected or discarded. Wrapper methods look for combinations of features and constantly evaluate and compare them with other feature combinations. Embedded methods carry out feature selection when the model is being trained. We have implemented feature selection by calculating feature importance of all attributes using forest of trees. Finally, random forest classifier is being used for the prediction of diseases. A random forest is a collection of lots of individual trees. Each one of them varying slightly in comparison to the other as each one of them works on a slightly different set of observations. The output of all the individual trees is combined to form a final prediction. This finalprediction is a result of the feature selection and the random forest algorithm that have helped create a model that is in quite a simplified form and helps in better diagnosis of diseases.

II. LITERATURE REVIEW

The paper [1] lays down the very basics of feature selection and goes on to explain the concept in detail by covering various areas related to it. The papers primary focus lies in creating subsets of features that help in creating a good predictor. Variable Ranking is discussed as a principle selection mechanism and mathematical explanations are provided for various techniques through which it can be carried out such as Correlation Criteria, Single Variable Classifiers, Information theoretic ranking criteria, etc. The paper served as an apt introduction to variable and feature selection and helped in acquiring a fundamental understanding about it. The paper [2] talks about the implementation of mRMR (Minimum Redundancy Maximum Relevance) feature selection method in marketing, focusing on the mathematical aspect of mRMR, while also covering real data examples. Implementation in Production is discussed in terms of Architecture of the Platform, Challenges and Optimization and Online Experiment Evaluation. The implementation of a specific feature selection method on a particular domain helped in understanding the overall approach that is to be taken while implementing a feature selection method to the project. It highlights the parameters that are to be kept in mind during ideation and in its subsequent implementation. The paper [3] discusses the use of machine learning techniques in analyzing data. It gives a brief idea about various feature selection methods namely principal component analysis, factor analysis and attribute ranker. It sheds light on multiple feature selection methods providing us a basic understanding of each one of them and how they can be useful for analyzing data in thefield of medicine. The paper [4] talks about data mining in medical dataset and applying feature selection for classification. Similar to [3], the paper highlights various feature selection methods for analyzing medical data. The difference being, that unlike [3], this paper also discusses the methodology apart from the concepts of those methods. Hence, it helps in understanding the approach that is to be taken, in order to implement feature selection in the domain of medicine. The paper [5] deals with Multilayer perceptronbased feature selection algorithm. The backpropagation algorithm trains the multilayer perceptron to determine the attributes to be removed from the data set. The concept of prominence is used for the objective function for the feature selection algorithm. It basically indicates the real relevance of an attribute for a given task. While this paper takes a different approach by using an MLP based feature selection method but the basic objective of eliminating redundancy and irrelevance matches our objective. The paper [6] discusses feature ranking and feature selection for a linear model. It provides a detailed procedure for ranking a feature, which is an essential step during the implementation of a feature selection method. The entire mathematical procedure is explained in detail, which helps in understanding not just the theoretical aspect but also its actual implementation in various applications.

III. METHODOLOGY

A. Medical Dataset Selection

The primary goal while looking for a dataset was to select one that provided a great deal of information. At the same time, it was important to keep in mind that the information gained from the data needed to be comprehensible. Initially, we leaned towards datasets that have information to predict a particular disease. Most of the datasets that we found using this approach led us to some pretty common datasets dealing with heart related diseases and breast cancer. These datasets had already seen way too many implementations and besides that, the number of columns in it didn't seem to be sufficient enough. We ended up choosing a dataset that focused on multiple diseases with enough data available on them. The dataset that we worked on has 133 columns in total with 132 of those columns signifying various symptoms and the output variable prognosis mentioned the disease that was caused due to these symptoms. Every row signified a patient record and had values of either 1 or 0 which means symptom exists and symptom doesn't exist respectively for 132 columns signifying various symptoms and the last entry for every row stated the disease as mentioned earlier. The number of unique values, i.e., the number of different diseases in the prognosis attribute are 41. With 4920 rows of information for training the model, the dataset is quite useful for the implementation of feature selection. The only bit of data pre-processing done was to classify the string values of the diseases to numbers, for the model to conveniently work on.

B. Feature Selection

With 132 different symptoms existing in the dataset, the objective was to find and eventually eliminate the features that did not contribute much to the final outcome. The implementation of feature selection on the dataset used was done using random forests [7]. This is because random forests use tree-based strategies to naturally rank elements based on how well they improve the purity of a particular node. At the start of the trees, nodes with the greatest decrease in impurity are found. Whereas, at the end of the trees, nodes with the least decrease in impurity are found. The impurity mentioned here is the gini impurity. Used for classification trees, gini impurity is a measure of the frequency of a randomly chosen element from the set being incorrectly labelled if it is being randomly labelled. Hence, the subset of important features is created by pruning trees below a particular node.

Table - 1
Feature importance of certain symptoms

Variable Name	*Feature Importance*
itching	*0.02717919016637643*
skin_rash	*0.017452142464731618*

nodal_skin_eruptions	*0.005726358713313166*
continuous_sneezing	*0.007039175702572009*
shivering	*0.018138450276278598*
chills	*0.02364134810494456*
joint_pain	*0.02747955161277762*
stomach_pain	*0.0019701707809017155*
acidity	*0.003836609029378934*
ulcers_on_tongue	*0.003969434577822112*

Table 1 shows the importance values of 10 features (symptoms) as an example. The selection criteria for a feature was set as the importance value for the given feature being more than or equal to the average importance of a feature. Hence, all features whose importance value was less than the average value was discarded.

C. Random Forest Classifier

The classifier was used initially so as to apply feature selection[8] using forest of trees[9]. Once a subset of features was obtained from the original set of features through feature selection, the classifier was trained for a second time in order to run the model with the new dataset consisting of the retained features. We used a forest of 120 trees in our classifier. Individual trees do not provide accuracy for predictions. A random forest consisting of largely uncorrelated trees come up with accurate predictions. The existence of uncorrelated trees is ensured through the concept of bagging which allows individual trees to randomly take samples for training from the dataset rather than allocating the same data to every tree. For example, let's assume that a training data has 5 rows. Instead of allowing a tree to train using those 5 rows, the tree is randomly allotted two instances of row 2, and 1 instance each of row 1, row 3 and row 5. Similarly, other trees are randomly allotted slightly different data resulting in each tree being slightly different than the others. Since the overall prediction is based on the prediction of majority of trees, the errors of individual trees get overshadowed by the accurate prediction of other trees. These characteristics of random forest[10] classifier informed our decision to use it for our model.

IV. RESULT

Once the importance of all features was calculated using feature selection [11], the features having an importance less than the average importance were to be discarded. This resulted in close to 70 features being removed out of 133. A lot of those features hardly featured as symptoms for most of the diseases in the dataset. Once random forest [12] classifier was applied on the existing features, we fed certain symptoms to test our model. For example, on entering symptoms of diarrhoea and vomiting, the model predicted Gastroenteritis as the condition with an accuracy of 85%. On providing itching and blister as the symptoms, Drug Reaction was the prognosis made by the model with 95% accuracy. While testing for Hepatitis C an accuracy of 90% was obtained.

V. FUTURE SCOPE

One key improvement that can be made is that multiple diseases are simultaneously detected when the symptoms are entered. There are many cases in the real world where patients suffer from multiple diseases. And these diseases are most often linked to each other due to various health reasons. The model we have implemented at present detects one disease at a time for a given set of symptoms. There is a possibility that the patient suffers from another disease but it isn't detected because the set of symptoms entered match to a high extent with the predicted disease. Apart from that, various feature selection methods could be implemented on the medical dataset and compared amongst each other in order to come up with a method that provides the highest accuracy for disease prediction.

VI. CONCLUSION

With data getting bigger by the day, it has become a necessity that people keep up with it and try to get useful insights from it as much as is possible. [4] Looking at the field of medicine, there is so much scope for the usage of machine learning, right from providing better care at low cost to making higher accuracy predictions for diseases. Feature Selection is just one of the many tools available, that helps achieve the aforementioned objectives. Yet it's relevance can be credited to the fact that it allows for faster processing of data, better interpretation of a model and better accuracy for predictions.

REFERENCES

[1] Isabelle Guyon, Andre Elisseeff, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Researc h 3 (2003) 1157-1182.
[2] Zhenyu Zhao, Radhika Anand, Mallory Wang, “Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform”, The IEEE International Conference on Data Science and Advanced Analytics, 2019.
[3] Rahul Samant, Srikantha Rao, “A study on Feature Selection Methods in Medical Decision Support Systems”, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 11, November - 2013 IJERT ISSN: 2278-0181

[4] Prof.K.Rajeswari, Dr.V.Vaithiyanathan, Shailaja V.Pede, "Feature Selection for Classification in Medical Data Mining", International Journal of Emerging Trends and Technology in Computer Science.
[5] E. Gasca, J.S. Sánchez, R. Alonso, "Eliminating redundancy and irrelevance using a new MLP-based feature selection method", The Journal of the Pattern Recognition Society.
[6] Herve Stoppiglia, Gerard Dreyfus, Remi Dubois, Yacine Oussar, "Ranking a Random Feature for Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1399-1414.
[7] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.
[8] Ms. Shweta Srivastava , Ms. Nikita Joshi , Ms. Madhvi Gaur (2013). A Review Paper on Feature Selection Methodologies and Their Applications. International Journal of Engineering Research and Development (IJERD) 2278-067X.
[9] Cuong Nguyen, Yong Wang, Ha Nam Nguyen,"Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.", 2013, Journal of Biomedical Science and Engineering.
[10] Guanglu Sun, Shaobo Li, Yanzhen Cao, and Fei Lang , "Cervical Cancer Diagnosis based on Random Forest", 2017, International Journal of Performability Engineering, 446-457.
[11] Jundong Li,Kewei Cheng,Suhang Wang,Fred Morstatter,Robert P. Trevino,Jiliang Tang,Huan Liu, "Feature Selection : A Data Perspective" (2017),ACM Computing Surveys, 94.
[12] Gerard Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research (2012), 1063-1095.