Diagnosis of Four Stages of Thyroid using United States' Data

Coral Jain Yashovardhan Sharma

EECS
Washington State University
Pullman, WA-99164

Abstract

Thyroid is one of the deadliest diseases in United States. More than half the population of United States is assumed to be suffering from this disorder. It is of utmost concern for all of us that we design a classifier that makes it easy for the medical practitioners to classify the disease on the basis of data available to them. In this study, we have used Support Vector Machine algorithms along with other machine learning concepts to classify the data into four classes. The approach that we have used here is one-vs-all where we divide the problem into four binary classification problems and then classify the data. We have used thyroid dataset from the official website of American Thyroid Association which is divided into multiple classes. The dataset is further exploited Exploratory Data Analysis (EDA) and Feature Engineering to understand the dataset in a better manner and make improvements before training the model. The trained model is then used for classifying the data into four categories namely, Euthyroid, Euthyroid Sick, Hypothyroidism and Hyperthyroidism.

1 Introduction

Thyroid is one of the deadliest diseases that exist today. Thyroid is a gland which secrets a specific type of hormone called thyroid hormone. It has various functions in the body. Some of which regulates the basic functioning of the body. It includes controlling heart rate and providing a balance in the functioning of the heart. Secondly, it also manages the weight of the body. Thirdly, it strengthens the breathing and muscle power in human beings. Thyroid disease is very common in countries like USA, India, and many European countries such as France, Sweden, etc. It is categorized into four stages, namely, Euthyroid, Euthyroid Sick, Hypothyroidism and Hyperthyroidism. Although the treatment of this disease exists today, it is not very predictable and does not generate very rewarding results. While our main area of concern is United States, it is to be noted that according to the official website of American Thyroid Association, around 27 million people in United States are predicted to have been suffering from this deadly disease. Amongst these, majority of the patients are women. As per the above source, around 25% of all the men in America will die of inflamed thyroid in coming years. Moreover, the website also reports that more than 50% of the population in American is misdiagnosed, which means the diagnosis of this disease is a major challenge in the medical industry of the world. Therefore, the imperative nature of the situation must be understood, and it has to be realized that we need better classifiers for the diagnosis of this disease. One of the very recent innovation is the classification of thyroid diseases using Convolutional Neural Network. However, there is a lot of work that needs to be done before this classifier is implied.

2 Problem Definition

As per the statistics mentioned above, it is a big concern for the medical industry that we provide a means by which we can carry out effective methods to diagnose four stages of disease. Our problem is not just finding the means to detect the disease but to find a means through which we can classify with less computation power. Another concern here is that when we are dealing with multi-dimensional arrays, it is difficult to manipulate using algorithms which require high computation power. Thus, we need to find a way to manipulate high-dimensional arrays using an algorithm and still use less computation power. Thus, main objective of this project is to provide an efficient, productive and accurate classifier that solves our problem at hand. Our classifier will provide a better solution that uses for diagnosis of four stages of disease stated above. Our instrument will produce an enormous decline in misdiagnoses as it is capable of differentiating between problems of the thyroid gland and other different disorders in the body. In addition to this, we are also offering the facility to diagnose thyroid disease before it leads to a more damaging problem.

3 Solution Approach

Let us consider a diagnosis case of thyroid disease, the patient has undergone some tests to determine whether he is suffering from the disease because he/she showed symptoms of the corresponding disease. After the test, the medical practitioner checks the report and realized that the functioning of the thyroid appears healthy. In such a case, what result does the practitioner give to his/her patient? Obviously, the patient is reported to be not suffering from the disease. Well, who is to be blamed in such a scenario? You cannot blame the practitioner as the tool and the reports clearly show that the disease looks inexistent in the patient's body. However, this leads to intricacies in the body and thus, the tool needs to be blamed for this as it only provides the practitioner information about only one kind of thyroid disease, which in general, is of four types. In such a scenario we need to define a classifier which takes the data and utilizes it classify into four categories.

One of the main challenges in diagnosis of this disease is that there are a number of classifiers that are in use, but none of them are efficient enough to classify between the multiple stages of thyroid disease. All the four stages are difficult to be diagnosed and although, there are classifiers exist, the main challenge for them is less computation power. Moreover, in our case, since there are four classes that need to be handled, it is a case of multi-class classification and with other machine learning algorithms it is difficult to obtain high accuracy.

As a solution for the above problem, we need to solve our objective to build a hyperplane that distinctly classifies the data points. Hence, we are using Support Vector Machine (SVM) algorithm in our case as it obtains better accuracy than other algorithms and there are also no issues of overfitting the data. Also, decision boundary of SVM is very simple and thus it is very easier to implement and classify into multiple classes using this algorithm. Due to the above reasons, we are utilizing machine learning concepts incorporated with SVM algorithm to tackle the challenges. For this, we need a hyperplane to perform multi-classify task at hand. The hyperplanes are decision boundaries that help classify the data points. Technically speaking, the main function of the implementation of the classifier is to find a decision boundary that classifies the data points in

space. There exist many possible hyperplanes, but the one that has the maximum margin, or distance, between data points of the classes is ideal. This maximum marginal distance provides confidence that future data points are assigned to the correct class. If a data point falls on either side of the hyperplane, it is attributed to different classes. In short, we are in search of a hyperplane that will solve our problem. Our instrument will produce an enormous decline in misdiagnoses as it is capable of differentiating between problems of the thyroid gland and other different disorders in the body. In addition to this, we are also offering the facility to diagnose thyroid disease before it leads to a more.

4 Experiments

4.1 Dataset

The data has been retrieved from UCI Machine Learning Repository (*link*: https://archive.ics.uci.edu/ml/datasets/thyroid+disease). The directory consists of 6 databases and each dataset consists of approximately 2800 training data instances and 972 test instances. The dataset is divided some classes, some of the major classes are:

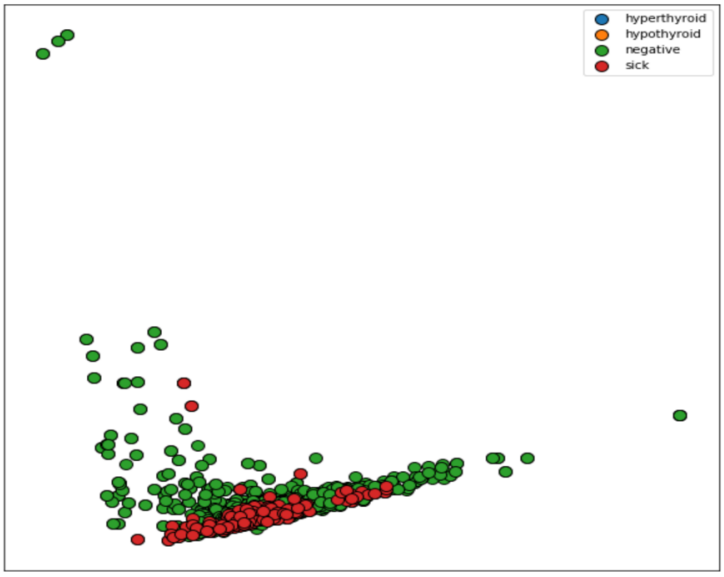
- Age: The age of the person.
- Sex: The gender of the person, whether they are Male(M) or Female(F).
- On Thyroxine: A medicine that is used to treat patients with hypothyroid. Denoted as True(T) or False(F).
- On Anti-thyroid medicine: Medicine used to treat patients with hyperthyroid. Denoted as True(T) or False(F).
- Sick: To denote whether patient is suffering from any type of thyroid disease. Denoted as True(T) or False(F).
- Pregnant: As a metric for women which are pregnant as this affects the conditions for cure. Denoted as True(T) or False(F).
- Thyroid Surgery: Denotes patients who underwent thyroid surgery for any prior thyroid disease. Denoted as True(T) or False(F).
- TSH: Measure of Thyroid stimulating hormone, which is a major hormone responsible for imbalance of thyroid gland imbalance.

Some other metrics we considered were T3, TT4 hormones, also we considered other diseases like tumors which might affect a patient's conditions significantly. The image below represents the data distributed in a hyperplane.

4.2 Exploratory Data Analysis

The main goal or objective of doing this is to make corrections before training model. It is a preprocessing step which carries out imputation of missing values for two kinds of values given in our data. Our dataset consists of values which are either continuous or categorical. The continuous values such as age are substituted in with the mode of the column whereas the categorical values such as sex are erased from the dataset as it is obvious that the gender of a person would not matter in case of the diagnosis. Secondly, the balanced data has been leaned

heavily towards the negative samples and if it is found once, it would be used to train model effortlessly.



In addition to all this, the baseline method that we have presented here is one-versus-all scheme. SVMs are generally used for binary classification and not multi-class classification. In order to deal with this problem, we are using one-vs-all approach, in which we convert a multi-class problem into multiple binary class problems. Thus, this approach will differentiate between one and the rest multiple number of times and finally give us output multiple times. This will lead to classification at the end.

4.3 Features

For our program, the features that were chosen were age, pregnant, thyroid surgery, etc. The continuous values such as age are substituted in with the mode of the column whereas the categorical values such as sex are erased from the dataset as it is obvious that the gender of a person would not matter in case of the diagnosis.

5 Results

We selected SVM as it is one of the better algorithms for classification and the data, we had was linearly separable. We measured the performance of our model using an index called F1 index which relies on two parameters p and r, which are precision and recall respectively. F1 score is the harmonic mean of p and r. Below image represents the classification report as we can see that calculated our results for four major categories hyperthyroid, hypothyroid, negative and sick where we calculated their F1 score along with their accuracy, macro average and weighted average.

```
Accuracy: 0.6935483870967742
**********************************************

                    Classification Report

              precision    recall  f1-score   support

hyperthyroid       0.21      0.68      0.32        19
 hypothyroid       0.25      0.64      0.36        55
    negative       0.95      0.70      0.80       689
        sick       0.30      0.67      0.42        43

    accuracy                           0.69       806
   macro avg       0.43      0.67      0.48       806
weighted avg       0.85      0.69      0.74       806
```

The other performance metrics that we have used are Classification Accuracy, which is generally defined as the correct number of predictions divided by the total number of samples. In our case, the classification accuracy is found to be 69.3%. in addition to this, confusion matrix is used for evaluating the results we have obtained. The matrix is obtained as shown below:

```
[[482  46 100  61]
 [  5  13   0   1]
 [ 15   0  35   5]
 [  7   3   4  29]]
```

6 Conclusions and Future Work

In this project we achieved accuracy of 69.3% which is higher than we expected but still leaves much to be desired. Through this project we were able to classify patients showing symptoms of thyroid in four major categories Hyperthyroid, Hypothyroid, Sick and showing no signs of Thyroid disease that is negative. While measuring accuracy we used F1 score as a measure which takes into account two parameters precision(p) and recall(r),precision can be described as number of correct positive divided by number of all positive results returned by the classifier and recall is number of correct positive results divided by all relevant samples. Then F1 score is the harmonic mean of precision and recall and takes a value from 0 to1 with 1 being the best and 0 being the worst.

In our project the F1 score is 0.69 with weighted avg of 0.74 which is on the upper side of the spectrum. We believe that in future we would try to improve the program and try to implement them for samples which show early signs of Thyroid disease so as to help patients in early stages of this disease further we would also like that we train the algorithms such that the samples which currently show no signs of Thyroid disease we can also classify them. This would be beneficial for people to precautious about their health.

7 Acknowledgment

This project gave us a deep insight of how to implement a machine learning model to a real-world problem and we would like to extend our gratitude to some of our peers and instructors. Firstly, we are grateful to our Professor Dr. Janardhan Rao Doppa who gave us the opportunity for this project and guided us throughout this project, whenever we had doubts, he was always eager to help us however he could Secondly our Teaching Assistant Mr. Aryan Deshwal who despite his busy schedule was always ready to guide us. Our peers also played significant part as the questions they asked during class and on the online forum Piazza were quite enlightening and acted as guidelines for our approach in this project. Lastly to Ms. Isabelle Guyon who first implemented support vector machine in 1991, as her implementation is the backbone of our project.

8 References

[1] Support vector machine

[2] Linear model

[3] MATHUR, A., AND FOODY, G.M. Multiclass and binary SVM classification: implications for training and classification users, *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 241-245 (2008).

[4] FEI, B., LIU, J. Binary tree of SVM: a new fast multiclass training and classification algorithm IEEE trans, *Neural Netw.*, vol. 17, no. 3, pp. 696-704 (2006).

[5] SVM Multi-class classification. *Apache Ignite. URL:* https://apacheignite.readme.io/docs/svm-multi-class-classification, Accessed on 7 December 2019.

[6] Quinlan,J.R., Compton,P.J., Horn,K.A., & Lazurus,L. (1986). Inductive knowledge acquisition: A case study. *In Proceedings of the Second Australian Conference on Applications of Expert Systems*. Sydney, Australia.