Classification of Thyroid Disease using Machine Learning

Krishna More#1

#Department of Information Technology Engineering, Mumbai University, India.

Abstract*—Thyroid diseases are increasingly common worldwide, affecting health conditions in multiple countries .This disorder is characterized by affecting multiple functionalities of the human body, directly damaging the living condition of people who suffer it. Globally, the use of technology in the field of medicine has made it possible to track the behavior of different disorders or diseases, and in some cases, has even made it possible to predict when a person might or might not suffer from any of them. With the rise of artificial intelligence, different academics and professionals have focused their efforts on making it a tool to help improve the physical condition of the world's population in order to improve the health conditions of people, and thus allow them a better quality of life. This paper aims to compare different classification algorithms used in machine learning as well as an Artificial Neural Network (ANN). The different classification algorithms used are Naive Bayes, Support Vector Machine, k-Nearest Neighbors, Random Forest Classifier, Logistic Regression and also an ANN.*

Keywords— Thyroid Classification, Random Forest, Logistic Regression, Naïve Bayes, KNN, Support Vector Machine, ANN.

1. INTRODUCTION

Thyroid diseases are increasingly common worldwide, affecting health conditions in multiple countries .This disorder is characterized by affecting multiple functionalities of the human body, directly damaging the living condition of people who suffer it. In the case of India it is estimated that there are approximately more than 42 million people who suffer a similar situation [1]. For the national case, it is calculated that approximately 4% of the population suffers from some disorder hormonal [2], and it is estimated that approximately worldwide about 200 million people have a thyroid disorder. Machine learning is not new to Thyroid research. Artificial Neural Networks (ANN) and decision trees have been used in Thyroid detection and diagnosis for nearly 20 years

Machine learning is a set of tools utilized for the creation and evaluation of algorithms that facilitate prediction, pattern recognition, and classification. ML is based on four steps: Collecting data, picking the model, training the model, testing the model [3]. The relation between BC and ML is not recent, it had been used for decades to classify tumors and other malignancies, predict sequences of genes responsible of cancer and determine the prognostic. In this study I use four machine learning classifiers which are Naïve Bayesian Classifier, k-Nearest Neighbors, Support Vector Machine and Artificial Neural Network.

2. BACKGROUND

A. Thyroid Disease Classification

Thyroid Disease Classification uses dataset extracted from the UC Irvine Machine Learning Repository. The Hypothyroid dataset are used for the research and development department for experimental purposes. The dataset contains 3090 instances. In this 149 data comes under hypothyroid and 2941 data is negative cases. In total it has 25 features, which are distributed as follow: 7 continuous variables and 18 categorical variables which are used for classification of Thyroid. The thyroid gland is the primary and biggest gland in the endocrine system. The data mining technique is applied on the hypothyroid dataset to determine the positive and the negative cases from the entire dataset. The classification of dataset is used to give better treatment, decision making, diagnose disease.

B. Machine learning approaches

Machine learning is a field of study that gives computer the ability to learn without being explicitly programmed. It is a branch of artificial intelligence. It can use methods such as statistics, probabilities, absolute conditionality, Boolean logic, and unconventional optimization strategies to classify patterns or to build prediction models. Machine learning can be divided into two categories: supervised learning (classification) and unsupervised learning depending on the used data and their availability. Some of the algorithms that have already been used are in the followingsection.

1) Artificial Neural Networks

Artificial Neural Network is a type of neural network that has been developed by taking inspiration from the brain. It is known that ANNs need to be modified to fit a particular application to achieve good performance. ANNs have been implemented multiple times for breast cancer classification usually using back propagation networks having three

layers a) Input Layer b) Hidden Layer c) Output Layer. Several researchers utilized three layers ANN with sigmoid function

2) Logistic Regression

Logistic regression is one of the most important analytic tools in the social and natural sciences. In natural language processing, logistic regression is the baseline supervised machine learning algorithm for classification, and also has a very close relationship with neural networks

3) Support Vector Machines

This algorithm is from the class of supervised learning algorithms and it reduces the overflowing of trained data. Its goal is to find the optimized decision boundaries to help predict breast cancer at the earlier stage that amongst three diverse data mining techniques.

4) Random Forest Classifier

Random forest is an ensembles of unpruned classification or regression like bootstrapping algorithm with number of decision trees. It is the blend of tree predictors where each tree relies on the values of the vector selected randomly and independently. When new input data is given, the algorithm makes trees of those input data and places them in forest. Random forest commonly provides a massive improvement than the single tree classifier

3. PROPOSED ALGORITHMS

This paper delves into the performance of ANNs as it is a very popular choice from classification problems. However in there been a rise in popularity of other algorithms such as Naïve Bayesian Classifier , k-Nearest Neighbors , Random Forest ,Naïve Bayes and Logistic regression in the field of medical research. Therefore I have decided to compare a few popular algorithms (ANN and Support Vector Machine) along with some other algorithms that are gaining popularity.

A. Dataset

Thyroid Disease Classification uses dataset extracted from the UC Irvine Machine Learning Repository. The Hypothyroid dataset are used for the research and development department for experimental purposes. The dataset contains 3090 instances. In this 149 data comes under hypothyroid and 2941 data is negative cases. In total it has 25 features, which are distributed as follow: 7 continuous variables and 18 categorical variables which are used for classification of Thyroid.

Table - 1
Hypothyroid Dataset

Attribute Name	*Value type*
age	*continuous,?.*
sex	*M,F,?.*
on_thyroxine	*f,t.*
query_on_thyroxine	*f,t.*
on_antithyroid_medication	*f,t.*
thyroid_surgery	*f,t.*
query_hypothyroid	*f,t.*
query_hyperthyroid	*f,t.*
pregnant	*f,t.*
sick	*f,t.*
tumor	*f,t.*
lithium	*f,t.*
goitre	*f,t.*
TSH_measured	*f,t.*
TSH	*continuous,?.*
T3_measured	*f,t.*
T3	*continuous,?.*
TT4_measured	*f,t.*
TT4	*continuous,?.*
T4U_measured	*f,t.*
T4U	*continuous,?.*
FTI_measured	*f,t.*
FTI	*continuous,?.*
TBG_measured	*f,t.*
TBG	*continuous,?.*

B. Support Vector Machine

1) Algorithm:

1. Prepare and format dataset
2. Normalize the dataset
3. Select activating function(usually sigmoid)
4. Optimize parameters c and g using search algorithm after cross validation
5. Train SVM network
6. Test SVM network

C. Naïve Bayesian Classifier

1) Algorithm:

1. Separate the data into a block of 2 classes and 2 sets of features
2. Find the standard deviation and mean for each feature and class
3. Find probability of each feature using density of normal distribution
4. Calculate probability of each class as a multiplication of probabilities of all features
5. Predict class of an instance using the probabilities

D. K-Nearest neighbors

1) Algorithm:

1. Split the dataset into training and testing set
2. Choose an instance and find its distance from the training set
3. Arrange the distances in ascending order
4. Class of the instance is the most common class of of the first k^{n} training instances

E. Description

For these three algorithms similar data preprocessing and data scaling steps were used to optimize the performance and get a better accuracy.

Steps

1. Dropping the first column from the data set as it is irrelevant(ID)
2. Use of standard scaling to normalize the data set as some algorithms have drastically improved performance after data scaling
3. Using 10-fold cross validation.
4. Using Confusion matrix to evaluate the performance of the algorithms for classification.

F. Artificial Neural Network

An ANN was implemented using Keras to classify the dataset. There is one input layer, two hidden layers and one output layer. Data was split into training set and testing set and then fit to the classifier. The results of all the algorithms have been documented below.

G. Logistic Regression

Logistic Regression was implemented using scikit-learn to classify the dataset. First the dataset was split into training and testing set. The scikit-learn library was imported from which the Logistic Regression was used to train by fitting and transforming the split dataset. The result of this algorithm is given below with comparison of the rest algorithms.

4. OBSERVATIONS AND RESULTS

Accuracy Performance of Support Vector Machine, Random Forest Classifier, Logistic Regression, Naïve Bayesian Classifier and k-Nearest Neighbors' was very poor before any kind of data scaling and cross validation was used but after feature scaling I was able to attain better Accuracy as shown in Table 1 . The observations are as follows

Table 1: Comparison between KNN, Random Forest Classifier, Logistic Regression, Naïve Bayes and SVM

Algorithm	**Accuracy**
Logistic Regression	96.1929%
KNN	95.5584%
Random Forest Classifier	95.5584%
SVM	95.4315%
Naïve Bayes	36.2944%

As we can observe that Naïve Bayes has the lowest accuracy amongst the tested algorithms. Logistic Regression was a very high accuracy of 96.5736% along with a reduced training time. However for very large data sets it is possible that Naïve Bayes will be a better algorithm due to the computational costs of SVM.

The performance of exceeds the performance of the aforementioned algorithms. The accuracy of ANN was found to be 96.7573%.

Fig 2: Confusion Matrix of ANN

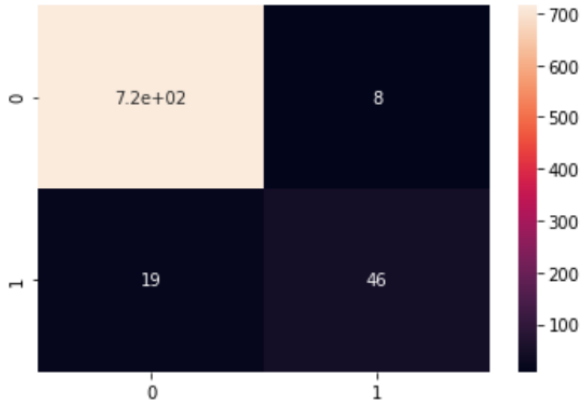
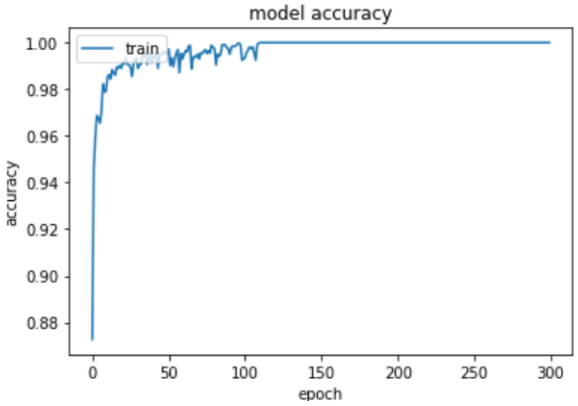


Fig 3: ANN Model Accuracy



5. CONCLUSION

Six algorithms were implemented on the Thyroid diseases dataset (Naïve Bayes, k-Nearest neighbors', Support Vector Machine, Random Forest Classifier, Logistic Regression and Artificial Neural Network). The observation was the ANN has the highest accuracy of 96.7573% however Logistic Regression has a good accuracy of 96.1929% and if the dataset is larger, the computational costs of the ANN will increase.

6. FUTURE SCOPE

In the future I will test these algorithms on different datasets with a larger number of instances so that we can confirm the conclusions we have made in this paper. It would be better to use real life datasets from different fields of science to exhaustively test these algorithms and compare their performances.

REFERENCES

[1] L. Adi Tarca, V.J.C., X. Chen, R. Romero, S. Drăghici, "Machine Learning and Its Applications to Biology", PLoS Comput Biol,, Vol. 3, pp. 116122, 2007.

[2] RuiXu, Anagnostopoulos, G.C. And Wunsch, D.C.I.I.,"Multiclass Cancer Classification Using Semi supervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol.4, No.1, Pp. 65-77, 2007.

[3] Thyroid Data Prediction using Data Classification Algorithm (ijirst.org)

[4] 5.pdf (stanford.edu)

[5] http://www.eluniversal.com.co/salud/el-4-de-los-colombianos-sufren-problemas-de-tiroides-278828-HBEU394560