Natl. Acad. Sci. Lett. (May–June 2021) 44(3):233–238
https://doi.org/10.1007/s40009-020-00979-z

SHORT COMMUNICATION

Thyroid Disease Prediction Using Machine Learning Approaches

Gyanendra Chaubey[1] · Dhananjay Bisen[1] · Siddharth Arjaria[1] · Vibhash Yadav[1]

Received: 11 November 2019 / Revised: 23 March 2020 / Accepted: 15 April 2020 / Published online: 20 May 2020

Abstract This paper is being written to provide a source of reference for the research scholars who want to work in the area of prediction of thyroid disease. From the different machine learning techniques, compared widely used three algorithms namely logistic regression, decision trees and *k*-nearest neighbor (*k*NN) algorithms to predict and evaluate their performance in terms of accuracy. This study has represented the intuition of how to predict the thyroid disease and highlighted how to apply the logistic regression, decision trees and *k*NN as a tool for the classification. For this, thyroid data set of machine learning repository has used from UC Irvin knowledge discovery in databases archive.

Keywords Machine learning · Decision tree · Thyroid disease · *k*-nearest neighbor · Logistic regression

Introduction

At least a person out of ten is suffered from thyroid disease in India. The disorder of thyroid disease primarily happens in the women having the age of 17–54. The extreme stage of thyroid results in cardiovascular complications, increase in blood pressure, maximizes the cholesterol level, depression and decreased fertility [1].

The hormones, total serum thyroxin (T4) and total serum triiodothyronine (T3) are the two active thyroid hormones produced by the thyroid gland to control the metabolism of body. For the functioning of each cell and each tissue and organ in a right way, in overall energy yield and regulation and to generate proteins in the ordnance of body temperature, these hormones are necessary [2, 3].

The idea for thyroid disease diagnosis and therapy is represented by the functional behavior of the thyroid disease and is the key in most thyroid diseases. The basis of classification of thyroid disease is euthyroidism, hyperthyroidism and hypothyroidism which are denoting normal, excessive or defective levels of thyroid hormones. The state euthyroidism depicts the normal production of thyroid hormones and normal levels at the cellular level by the thyroid gland. The state hyperthyroidism is clinical symptom due to excessive circulation and intracellular thyroid hormones. The state hypothyroidism is most of due to the lack of thyroid hormone generation and poor alternate therapy [4].

Cure of disease is a regular concern for the health care practitioners, and the errorless diagnostic at the right time for a patient is very important. Recently, by some advanced diagnosis methods, the common medical report can be generated with an additional report based on symptoms. The different questions like "what are the causes for affecting the thyroid?", "Which age group of people are affected due to thyroid?", "what is the relevant treatment for a disease?", etc. may find answers on implementing machine learning methods. Health care data can be processed and after implementing with certain methodologies; it can provide information that can be used in diagnosis and

✉ Gyanendra Chaubey
gyanendrachaubey68@gmail.com

Dhananjay Bisen
bisen.it2007@gmail.com

Siddharth Arjaria
arjarias@gmail.com

Vibhash Yadav
vibhashds10@yahoo.com

[1] Department of Information Technology, Rajkiya Engineering College, Atarra, Banda 210201, India

treatment of diseases more efficiently and accurately with better decision making and minimizing the death risk [5].

The large amount of data can be handled using the machine learning techniques. Classification models are well suited for the classification and distinction of the data classes. The handling of both numerical and categorical values can be done by the classification processes. Classification is a two-step classification model in the step one, based on some training data, a model is constructed, and in step two, an unknown tuple is given to the model to classify into a class label [6].

In human life, the classification has a great influence. The comparison of different classification techniques is a non-trivial and has a great dependency on the data set properties. In the statistics community, logistic regression, decision tree and *k*-nearest neighbor have got an esteemed position for classification problems [7].

Based on the research works and literature review, very little work has been done in the classification methods of patients pruned by the thyroid disease. The methods of classification used are the well-known methods. To focus on the above-discussed issues, this paper explains the use of three classification machine learning algorithms: logistic regression classification, decision tree classification and nearest neighbors classification to classify the people pruned by thyroid disease using the thyroid disease database. The paper explain in detail about the preparation, training and testing of the data, step-by-step description of each of the techniques used, and a comparison of the accuracy of the methods used in the prediction.

Research Methods

The data set has been taken from the Graven Institute in Sydney, Australia, uploaded to the UC-Irvin, knowledge discovery in databases [8]. The database has many data sets in this work; the "new-thyroid" data set is taken which contains 5 attributes and 215 instances. Only two most relevant attributes; total serum thyroxin (T4) and total serum triiodothyronine (T3). The outcome of the analysis is prediction of people having thyroid disease or not.

Logistic regression is a very good method to depict and test hypotheses for the two categorical values [9]. Logistic regression is used for classification using a linear decision boundary. Logistic regression works by first looking for linear decision boundaries between the samples of different classes. Then, the logistic function is used to get the probability of belongingness to each class defined with respect to the decision boundaries.

The general formula for the logistic regression classification is:

$$h_B(p) = \frac{1}{1 + e^{-B^t p}} = k(B^t p)$$

$$k(z) = \frac{1}{1 + e^{-z}}$$

The above equation is called the logistic function or sigmoid function. The logistic regression uses the data preparation, splitting the data into training, validation and test set, fitting the line model for classification and finally evaluation of the result.

The decision tree uses the machine learning technique to solve the problem of classification and prediction. Nodes and leaves are the two elements of which the decision trees are formed. Nodes help in the testing of a particular attribute and leaves represents a class [10].

The decision tree implementation is top-down approach. The tree is build with the goal to achieve the maximum homogeneity in leaves as possible. The continuous division of leaves from non-homogenous to homogeneous is the major concern of this algorithm. The steps of training, classification and testing are easy and fast in decision trees. It gives easiness to the users to gain the information by the tree representation of the knowledge [11].

The core algorithm used here is the ID3. It is a greedy search technique with no backtracking of the entire possible branch. The algorithm uses the entropy and information gain to find the possibilities. The formulae for the calculation of entropy and information gain are given below:

1. Entropy:

Entropy using a single attributes:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Entropy using the two attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

2. Information gain:

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

Following steps are used to make a decision tree:

- Data preparation
- Data partition into training, validation and testing set
- Selection of attribute: a method to select the "best" possible attribute for the splitting by the decision tree model
- Evaluation of the model

In the *k*NN classification, the learning is based on analogy that the test tuple is mapped by comparing with the training tuples that are similar to it. When given an

unknown data point, a *k*-nearest neighbor classifier finds the pattern space for the k training tuples that are closest to the unknown data point. The unknown tuple is classified by a majority of its neighbors, and gets assigned to the class most common among its *k*-nearest neighbors. On giving a training tuple *k*-nearest neighbor simply stores it and waits until it is given a test tuple. Thus, it is a "lazy learner" as it stores the training tuples or the instances, they are also known as "instance based learners" [12].

The *k*-nearest neighbor algorithm is based on the distance of the nearest neighbors and uses the following distance formulae to find the nearest neighbors:

1. Euclidean distance:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

2. Manhattan distance:

$$\sum_{i=1}^{k}|x_i - y_i|$$

3. Minkowski distance:

$$\sum_{i=1}^{k}(|x_i - y_i|)$$

The above all the distance are useful in case of the continuous variables. In case of the categorical variables:

4. Use the hamming distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \rightarrow D = 0$$

$$x \neq y \rightarrow D = 1$$

In this work, Euclidean distance is used.

Following four steps are used to do the *k*NN classification:

- Estimate the distance metric between the test data point and all the labeled data points.
- Order the labeled data points in the ascending order of distance metric
- Select the top k-labeled data points and look at the class labels
- Find the class label that majority of these k-labeled data points have and assign it to the test data points

Results and Analysis

The visualization of the training data set will be same for all the three classification methods. The visualization of the new thyroid data set is shown in the Fig. 1a.

The analysis and explanation of each algorithm is reported below.

Logistic Regression Classification

The logistic classification classifies the data based on the sigmoid function. The classification of the thyroid data set by logistic regression classification is shown in Fig. 1b. The data are divided into three parts:

- Training set (70%)
- Validation set (15%)
- Test set (15%)

On evaluating the logistic regression classifier on this thyroid data set, it shows a validation misclassification percentage of 18.75% and test misclassification percentage of 15.625%. The confusion matrix drawn on the random selection of test data on the random selection of training data is shown in Fig. 1c. The confusion matrix explains about the how much the model is accurate. The formula for the calculation of accuracy from the confusion matrix is given as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FN}) + (\text{FP} + \text{TN})} \quad (1)$$

where *TP* true positive, *FP* false positive, *FN* false negative, *TN* true negative.

Putting the values in the formula,

$$\text{Accuracy} = \frac{2 + 24}{(2 + 0) + (6 + 24)}$$

$$\text{Accuracy} = \frac{26}{32} = 0.8125$$

Hence, the accuracy is 81.25%.

Decision Tree

Total serum thyroxin and total serum triiodothyronine are selected as the feature names for making the decisions. The class that the output produce will be class 0 (having thyroid) and class 1 (normal). To prepare the model, data set is

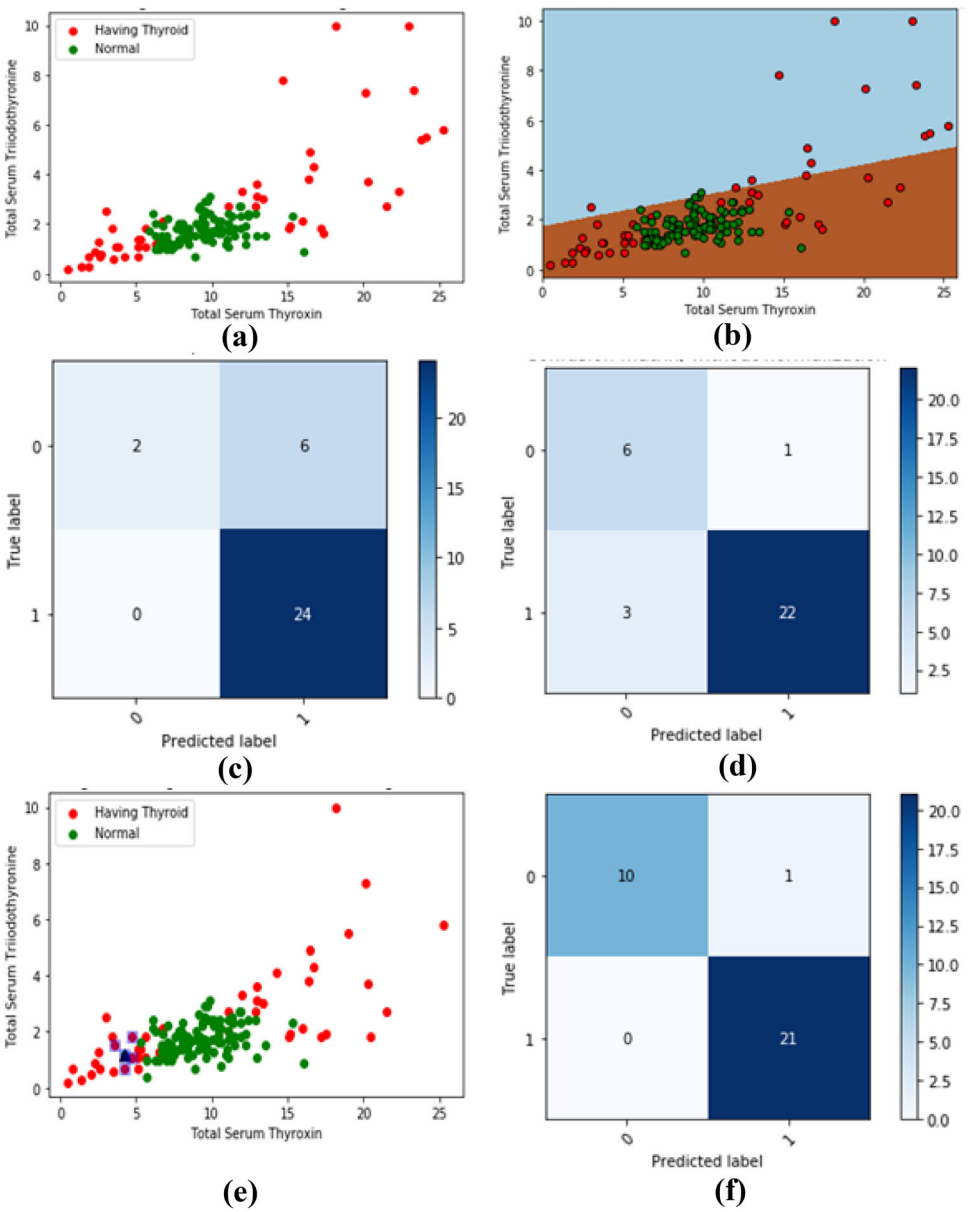

Fig. 1 **a** Visualization of data set. **b** Visualization decision boundary of logistic regression model. **c** Confusion matrix of logistic model. **d** Confusion matrix of decision tree. **e** Working of the *k*NN algorithm. **f** Confusion matrix of *k*NN model

divided into training set (70%), validation set (15%) and test set (15%).

On evaluating the performance of the algorithm, it shows validation misclassification percentage of 12.5% and test misclassification percentage of 3.125%.

The confusion matrix is drawn here for calculating the accuracy of the model is shown in Fig. 1d. The accuracy of this matrix can be calculated using the Eq. (1). Here, putting the values in the above equation

$$\text{Accuracy} = \frac{6+22}{(6+3)+(1+22)}$$

$$\text{Accuracy} = \frac{28}{32} = 0.875$$

So, the accuracy calculated here is 87.5%.

***k*NN**

While applying the algorithm at random chosen a point [4.2 1.2] as query point. The true class of the query point is 0. On applying the algorithm, the nearest neighbors of the query point are: ([4.2 1.2] [4.2 0.7] [4.7 1.1] [3.6 1.5] [4.7 1.8]), classes of the nearest neighbors are: ([1] [0] [0] [0] [0]) and predicted class for query point is also 0. The visualization of working of *k*NN is shown in Fig. 1e.

On evaluating the performance of the *k*-NN classifier, the test misclassification percentage = 3.125%.

The confusion matrix of the test data is shown in Fig. 1f. For calculating the accuracy of the matrix, the Eq. (1) is used. Here, putting the values from the matrix,

Table 1 Result analysis

	Logistic regression classification (%)	Decision tree classification (%)	*k*-NN classifier (%)
Test misclassification percentage	18.75	12.5	3.125
Validation misclassification percentage	15.625 .	3.125	6.25
Accuracy	81.25	87.5	96.875

Table 2 Compare with previous work

Research/algorithms	Decision tree accuracy	*k*NN accuracy
Ankita Tyagi and Ritika mehra [13]	75.76% (Much lower accuracy)	98.62% (little better accuracy)
Proposed method	87.5%	96.875%

Table 3 Compare with previous work

Research/algorithms	Decision tree accuracy	*k*NN accuracy
Rafi khan et al [14]	98.89% (Better accuracy)	91.62% (Much lower accuracy)
Proposed method	87.5%	96.87%

$$\text{Accuracy} = \frac{10+21}{(10+0)+(1+21)}$$

$$\text{Accuracy} = \frac{31}{32} = 0.96875$$

So, the accuracy calculated here is 96.875%.

From our research work, it is shown that how can thyroid disease be predicted and give an intution how to apply the logistic regression, decision tree classification and *k*NN algorithms. According to the data set, the following results results are obtained.

The result (Table 1) shows that the *k*NN classifier is a better algorithm for this data set in thyroid disease prediction.

The efficiency of an algorithm depends upon the data set and its features selected for the prediction. Some papers written during 2018–2020 have less accuracy than proposed algorithms, and some algorithms have a better accuracy which is due to the data set they have chosen. The paper given in below in Ref. [13] has shown less accuracy in case of decision tree, while in case of *k*NN they have better accuracy shown in Table 2: compare with previous work.

The UCI thyroid repository itself contains many data sets for thyroid disease. For proposed work, "new-thyroid" data set has been taken [8]. The paper authors [13] might have taken different data set of the same UCI thyroid repository. This is the reason of variation of result. Another work [14] has shown much less accuracy in case of *k*NN (91.82%) while decision tree has a better accuracy of 98.89% represented in Table 3: compare with previous work.

Conclusion and Future Work

Rafikhan et al. [14] has used a clinical data of Kashmir of 807 patients and UCI thyroid repository of "new thyroid" has only 215 instances. Proposed method has not taken this data set for thyroid prediction; it will consider in future work and measure accuracy using decision tree and *k*NN. Hence, according to the data set which is used in this work, the accuracy obtained is satisfactory.

The current scenario is of the developing of the models that help in the various sectors of life using the machine learning. The availability of data and its generation day by day increased a chance for the computer scientists to make prediction and analysis on such data sets that make the human life better and comfort. This study is concern with this motivation. The prediction and classification of any data depends on the data set itself and the various algorithms that are used. If anyone organizes a better data set of real time and applies various other machine leaning and deep learning algorithms such as SVM, Naïve Bayes, auto encoders, ANNs and CNNs then further better results may be achieved.

References

1. Chen Ling, Li Xue, Sheng Quan Z, Peng W-C (2016) Mining health examination records—a graph-based approach. IEEE Trans Knowl Discov Eng 28:2423–2437
2. Temurtas F (2009) A comparative study on thyroid disease diagnosis using neural networks. Expert Syst Appl 36:944–949

3. Ulutagay G (2012) Modeling of thyroid disease: a fuzzy inference system approach. Wulfenia J 19(1):346–357
4. Monaco Fabrizio (2003) Classification of thyroid diseases: suggestions for a revision. J Clin Endocrinol Metab 88:1428–1432
5. Ionita I, Ionita L (2016) Prediction of thyroid disease using data mining techniques. Broad Res Artif Intell Neurosci 7(3):115–124
6. Gorade SM, Deo A, Purohit P (2017) A study of some data mining classification technique. Int Res J Eng Technol 4(4):3112–3115
7. Bichler M, Kiss C (2004) A comparison of logistic regression, *k*-nearest neighbor, and decision tree induction for campaign management. In: Proceedings of the tenth Americas conference on information systems, New York
8. http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/
9. Peng CYJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96(1):3–14
10. Mesarić J, Sebalj D (2016) Decision trees for predicting the academic success of students. Croat Oper Res Rev 7:367–388
11. Patel BN, Prajapati SG, Lakhtaria K (2012) Efficient classification of data using decision tree. Bonfring Int J Data Min 2(1):6–12
12. Introduction to machine learning edition 2, by Ethem Alpaydin. https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf
13. Tyagi A, Mehra R (2018) Interactive thyroid disease prediction system using machine learning technique. In: 5th IEEE international conference on parallel, distributed and grid computing (PDGC-2018), 20–22 Dec, Solan, India
14. Sidiq U, Aaqib SM, Khan RA (2019) Diagnosis of various thyroid ailments using data mining classification techniques. Int J Sci Res Comput Sci Eng Inf Technol 5(1):2456–3307