Prediction of Thyroid Disease Based on Classification Using Hierarchical Structure

Charan R, Akash Yadav M, Aprameya N Katti, Mohith P
Global Academy of Technology, Bengaluru Karnataka 560098, India

Prof. Haseeba Yaseen
Global Academy of Technology, Bengaluru Karnataka 560098, India

Abstract: *Thyroid disease is a major cause of concern in medical diagnosis and the prediction of which is a difficult proposition in medical research. The machine learning plays a vital role in the process of disease prediction and this paper handles the analysis of the classification and prediction of the thyroid disease based on the information gathered from the UCI machine learning repository.*

Keywords: *Machine learning, unsupervised and supervised learning, classification, prediction, and threshold.*

1. Introduction

Machine learning is an application of Artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Disease diagnosis is a very complex and tedious task, as it requires lots of experience and knowledge. One of the traditional ways for diagnosis is doctor's examination or a number of blood tests. The main task is to provide disease diagnosis at early stages with higher accuracy.

According to analysis one in ten adults in India's people is suffering from hypothyroidism. This estimation is found on the premise of an analysis conduct by Indian thyroid society. The study also alert for thyroid is 9th ranked in comparison to other type common disease like asthma, cholesterol, depression, diabetes etc. medical practitioner say that thyroid are same as other disorders however, the investigation population are alert to thyroid disorders, know that there are diagnostic tests for finding of this disease.

Thyroid gland creates abnormal type of thyroid hormones like as hypothyroid and hyperthyroid. Hypothyroidism (underactive thyroid or low thyroid) that is called the thyroid hormones are not generating as much as necessary of certain important hormones. Hypothyroidism can justification various health problems such as heaviness, joint pain, unfruitfulness and heart disease. Hyperthyroidism (overactive thyroid) belongs to a position is the thyroid gland delivers a lack of the hormone thyroxin. For this situation, the body's digestion system is quickening essentially, bringing about sudden weight reduction, a fast or irregular heartbeat.

The machine learning plays a vital role in the process of disease prediction. Artificial Neural Network, support vector machine, Naive Bayes and K-Nearest Neighbor are the important modes applied to the prediction of thyroid disease. The experimental study has been conducted using Rapid miner tool and the results shows that the accuracy of K-nearest neighbor is better than all other technique to detect thyroid disease.

2. System Study

i. Machine Learning Techniques for Thyroid Disease Diagnosis:

This is the model for predicting the thyroid disease using DBSCAN algorithm. DBSCAN algorithm classify the data set using hierarchical multiple classifier and also reduce the problems faced by other multiple classifier algorithms with efficient and accurate information. Predicted hypothyroid disease using data mining algorithm called Linear Discriminant Analysis (LDA) to enhance the accuracy. LDA data mining classification techniques are used to classify the hypothyroid disease [1].

Advantages: Proposed model based on the Decision tree using entropy and information gain. This model is compared with KNN and **J**48 and Naïve Bayes algorithm. The proposed model has better accuracy than other models. The LDA Algorithm gives accuracy of 99.62% with cross validation [1].

Disadvantages: For large number of characteristics there is a need to build up a variable determination technique. This research has taken place for two groups A and B and it was observed that radiation exposure found to be 51% depending on the fluid intake.

ii. Applying Classification Algorithms to Predict Thyroid Disease:

In the field of artificial intelligence, Neuro-fuzzy refers to combinations of artificial neural networks and fuzzy logic. Neuro-fuzzy hybridization results in a hybrid intelligent system that synergizes these two techniques by combining the human- like reasoning style of fuzzy systems with the learning and connectionist structure of neural networks [2].

Advantages: The trial of 21 parameters is used. In KNN the prediction accuracy is 80%, ANN the accuracy is 85%. However fuzzy ANN the prediction accuracy is 90%. Therefore comparatively, Fuzzy ANN performs better than other two classification algorithms [2].

Disadvantages: Hard to develop a model from a fuzzy system. Require more fine tuning and simulation before operational. Problems of finding suitable membership values for fuzzy systems. A fuzzy system can be used to solve a problem if knowledge about the solution is available in the form of linguistic if-then rules.

iii. Classification Model Using Random Forest and SVM to Predict Thyroid Disease:

Data Mining Techniques like classification algorithms used are Random forest algorithm and support vector machine algorithm to get the better accuracy.

Advantages: result provides improved accuracy, precision, recall and F-measure by comparing the random forest with LDA algorithm [3].

Disadvantages: main drawbacks include the poor performance on imbalanced data (rare outcomes or rare predictors) and lack of an interpretable model.

3. Proposed System

If we observe the diagram, first classification is done using the threshold value given by the medical expert or internet source for the particular attribute. Based on the threshold frequency the tuples in the dataset will be classified as normal or abnormal. During Classification the attributes will be prioritized and will be allotted priorities this is called **Training.**

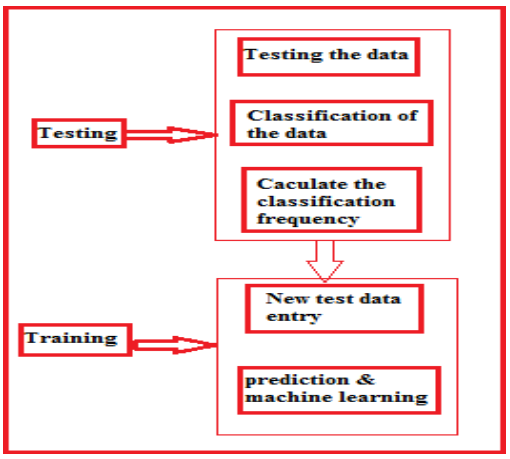

Figure 1. Training and Testing Data

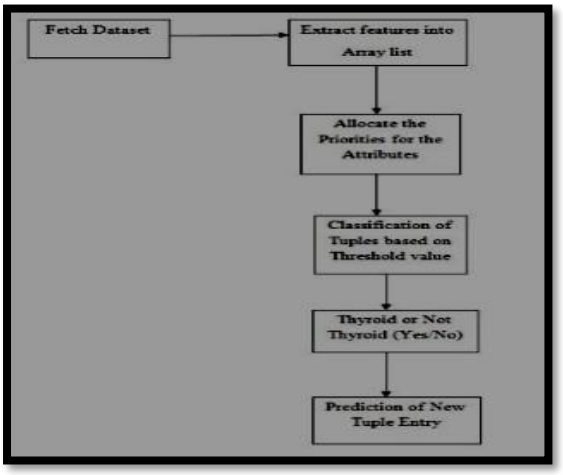

Figure 2. Proposed System

The figure demonstrates about the proposed system for Thyroid Prediction technique. It mainly consists of two phases namely

1) Classification (training).
2) Prediction (testing).

If we observe the diagram first the data will be loaded to the memory by using POIFS (Poor Obfuscation Implementation File System). The attributes will be prioritized and will be allotted priorities. The threshold value for Prioritized attributes will be extracted which is obtained by the help of internet source/medical expert.
We consider only prioritized attributes because those are very important attributes which are essential for prediction and classification of the rows and those attributes have numerical value.

Some prioritized attributes:-

1. Thyroxine level (0.7-0.9ul/ml).
2. Query on Anti thyroid.
3. TSH (0.5-0.6ul/ml) measured.
4. T4 (Serum Thyroxine) (4.6-12ug/dl) measured.
5. FTI (Free Thyroxine Index) (4-11) measured.
6. TBG (Thyroxine Binding Globin) (12-20ug/dl) measured.

Based on the threshold all the rows gets the violated and non violated frequency per tuple and will be classified as normal or abnormal.
First **classification** is done using the Threshold of prioritized attributes the tuples in the dataset will be classified as normal or abnormal this is called **Training.**
Prediction of new tuple entry is predicted using KNN algorithm called as **Testing.** KNN algorithms use a data and classify new data points based on a similarity measures (e.g. distance function).Classification is done by a **majority vote** to its neighbors. Based on the new entry given either by Doctor or patient KNN will be applied and will be traced. By the application of the Euclidean distance all the tuple result will be prioritized and need to remove the items which are having less value than the K entry value. Based on the Remaining value prediction will happen this prediction can be learned by machine.

4. System Architecture

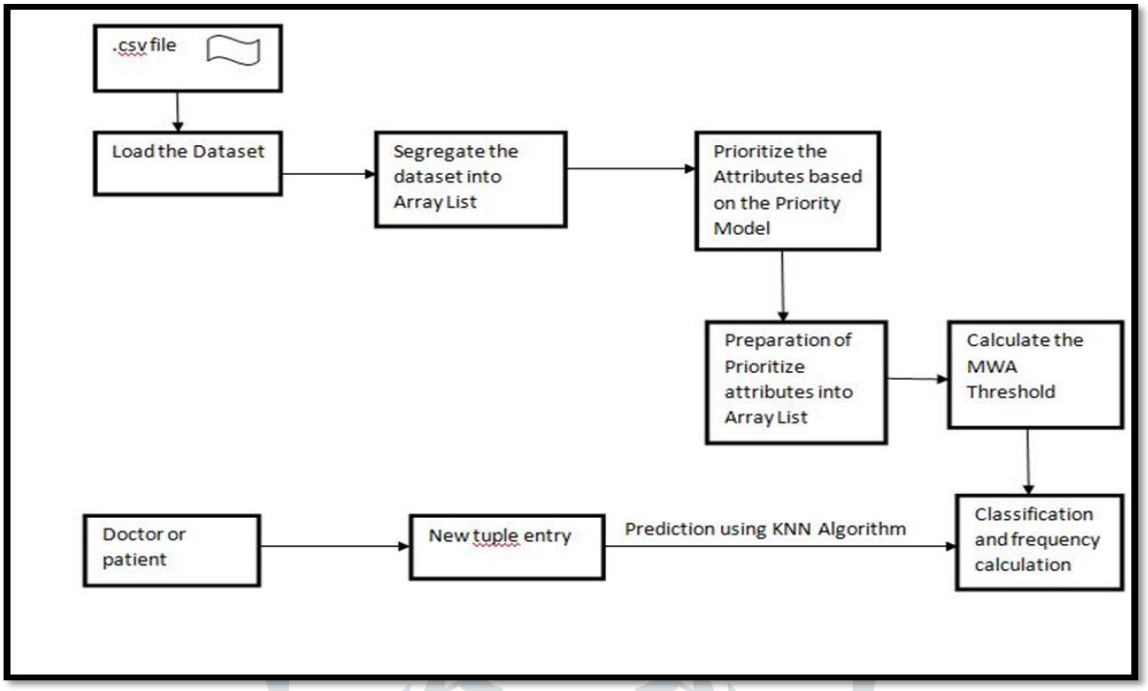

Figure 3: System Architecture

Load dataset takes the thyroid dataset available on ASCII-file. In dataset available we have,

- Number of attributes: 21 (15 attributes are binary, 6 attributes are continuous)
- Number of classes: 3
  a) Normal (no thyroid)
  b) Hyper function (Hyperthyroid)
  c) Subnormal functioning (Hypothyroid).
- Number of learning examples: 3772
- Number of testing examples: 3428

Attribute Name	Possible Values
age:	continuous.
sex:	M, F.
on thyroxine:	f, t.
query on thyroxine:	f, t.
on antithyroid medication:	f, t.
sick:	f, t.
pregnant:	f, t.
thyroid surgery:	f, t.
I131 treatment:	f, t.
query hypothyroid:	f, t.
query hyperthyroid:	f, t.
lithium:	f, t.
goitre:	f, t.
tumor:	f, t.
hypopituitary:	f, t.
psych:	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.
referral source:	WEST, STMW, SVHC, SVI, SVHD, other.

Figure 4: Dataset with attribute and possible values

Segregate Dataset takes the attributes and store them in the data structure that is store the attributes in the array list for further process. prioritized attributes are considered which are age, Thyroxine level, Query on Anti thyroid, TSH measured, TT4 measured ,Pregnant, FTI measured, TBG measured prioritized attributes are any number we can choose.

Classification is done by comparing the Threshold value (which is given by expert) of the prioritized attribute with the value of the prioritized attribute in the dataset. Basically each and every attribute will have a threshold from which classification can be done. The tuples are classified into two categories:

1. Normal (not affected/No)

2. Abnormal (affected/Yes).

```
29,F,f,f,f,f,f,f,t,f,f,f,f,f,t,0.3,f,?,f,?,f,?,f,?,f,?,other,-[840801013]
29,F,f,f,f,f,f,f,f,f,f,f,f,f,t,1.6,t,1.9,t,128,f,?,f,?,f,?,other,-[840801014]
41,F,f,f,f,f,f,f,f,f,t,f,f,f,f,f,f,?,f,?,f,?,f,?,f,?,t,11,other,-[840801042]
36,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,?,f,?,f,?,f,?,f,?,t,26,other,-[840803046]
32,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,?,f,?,f,?,f,?,f,?,t,36,other,S[840803047]
60,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,?,f,?,f,?,f,?,f,?,t,26,other,-[840803048]
77,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,?,f,?,f,?,f,?,f,?,t,21,other,-[840803068]
28,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.7,t,2.6,t,116,f,?,f,?,f,?,SVI,-[840807019]
28,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1.2,t,1.8,t,76,f,?,f,?,f,?,other,-[840808060]
28,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1.9,t,1.7,t,83,f,?,f,?,f,?,other,-[840808073]
54,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1.9,t,2.3,t,133,f,?,f,?,f,?,other,-[840810016]
42,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1,t,1.8,t,105,f,?,f,?,f,?,other,-[840813022]
51,M,t,f,f,f,f,f,f,f,f,f,f,f,f,t,0.5,f,?,f,?,f,?,f,?,f,?,other,-[840813060]
51,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.7,t,2.4,t,116,f,?,f,?,f,?,other,-[840813068]
37,F,f,f,f,f,f,f,f,f,f,f,t,f,f,f,f,?,t,2.9,f,?,f,?,f,?,f,?,other,-[840814014]
16,M,f,f,f,f,f,f,f,f,f,f,f,f,f,t,2.6,f,?,f,?,f,?,f,?,f,?,other,-[840814057]
54,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1,t,2,t,122,f,?,f,?,f,?,SVHC,-[840815016]
43,M,f,f,f,f,f,f,f,f,f,f,f,f,f,f,?,t,2.1,t,116,f,?,f,?,f,?,other,-[840815020]
63,F,t,f,f,t,f,f,f,f,f,f,f,f,f,t,68,f,?,t,48,t,1.02,t,47,f,?,other,F[840815067]
36,F,f,f,f,f,f,f,f,t,t,f,f,f,f,f,t,1.5,t,2.4,t,90,t,1.06,t,85,f,?,other,-[840815068]
40,F,f,t,f,f,f,f,f,f,f,f,f,f,f,f,?,f,?,t,79,t,0.94,t,84,f,?,other,-[840815069]
40,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1.2,t,2.3,t,104,t,1.08,t,96,f,?,other,-[840816001]
40,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,5.9,t,2.1,t,88,t,0.84,t,105,f,?,other,-[840816002]
77,F,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.05,t,2.4,t,107,t,1.13,t,95,f,?,other,-[840816003]
77,?,f,f,f,f,f,f,f,f,f,f,f,f,f,t,4,t,2,t,126,f,?,f,?,f,?,SVHC,-[840816004]
```

Figure 5: Dataset with attribute and corresponding values

Prediction: new tuple patient entry is given as input now the task is to predict whether the patient is affected with thyroid or not? Which is prediction (testing). Based on the classification results the new entry given is classified by using KNN algorithm. Predicted result may be either patient is normal or abnormal.

If new row entry is classified as Normal then we use Decision tree to predict what is possibility of this patient getting the thyroid? This can be predicted by the help of threshold value comparison and constructing the decision tree on prioritized attributes.

If new tuple entry is classified as Abnormal then we predict either person is suffering from hypothyroid or hyperthyroid. This is done using comparing the Threshold value with the patient entered value. This also can be done by constructing the decision tree.

5. KNN Algorithm

KNN makes classifications using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is **Euclidean distance**.

Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

Euclidean *Distance(x, xi) = sqrt (sum((xj – xij)^2))*

Euclidean is a good distance measure to use if the input variables are similar in type (e.g. all measured widths and heights). It is a good idea to try many different values for K and see what works best for your problem. The computational complexity of KNN increases with the size of

the training dataset. For very large training sets, KNN can be made stochastic by taking a sample from the training dataset from which to calculate the K-most similar instances.KNN has been around for a long time and has been very well studied. As such, different disciplines have different names for it, for example:

1) **Instance-Based Learning**: The raw training instances are used to make predictions. As such KNN is often referred to as instance-based learning or a case-based learning (where each training instance is a case from the problem domain).
2) **Lazy Learning**: No learning of the model is required and all of the work happens at the time a prediction is requested. As such, KNN is often referred to as a lazy learning algorithm.
3) **Non-Parametric**: KNN makes no assumptions about the functional form of the problem being solved. As such KNN is referred to as non-parametric machine learning algorithm.

6. Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by **learning decision rules** inferred from prior data (training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each **internal node** of the tree corresponds to an attribute, and each **leaf node** corresponds to a class label.

1) Decision Tree Algorithm Pseudo code:

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

In decision trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. We continue comparing our record's attribute values with other **internal nodes** of the tree until we reach a **leaf node** with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value.

2) Assumptions while creating Decision Tree

The below are the some of the assumptions we make while using Decision tree:

- At the beginning, the whole training set is considered as the **root.**
- Feature values are preferred to be categorical. If the values are continuous then they are converted to discrete values prior to building the model.
- Records are **distributed recursively** on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

Decision Trees follow **Sum of Product (SOP)** representation. It's a sum of product representation. The Sum of product (SOP) is also known as **Disjunctive Normal Form**. For a class, every branch

from the root of the tree to a leaf node having the same class is a conjunction (product) of values, different branches ending in that class form a disjunction (sum).

The primary challenge in the decision tree implementation is to identify which attributes we need to consider as the root node and each level. Handling this is the attributes selection. We have different attributes selection measure to identify the attribute which can be considered as the root note at each level.

The popular attribute selection measures:

- Information gain
- Gini index

If dataset consists of **"n"** attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy. For solving this attribute selection problem, researchers worked and devised some solutions.

They suggested using some criterion like **information gain, gini index,** etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e., the attribute with a high value (in case of information gain) is placed at the root. While using information Gain as a criterion, we assume attributes to be categorical, and for gini index, attributes are assumed to be continuous.

3) Information Gain

By using information gain as a criterion, we try to estimate the information contained by each attribute. We are going to use some points deducted from information theory.
To measure the randomness or uncertainty of a random variable X is defined by **Entropy**.
For a binary classification problem with only two classes, positive and negative class.

- If all examples are positive or all are negative then entropy will be zero i.e., low.
- If half of the records are of positive class and half are of negative class then entropy is one i.e., high.

By calculating **entropy measure** of each attribute we can calculate their **information gain**.

7. Conclusion

Disease diagnosis plays a major role and it is indispensable for any busy clinician. Thyroid disease is one such disease and prediction of which is a difficult aspect without a computer technology. In this paper, we have given an elaborate work that has been done by using Machine learning models. Based on the usage of these models, we have tried to show the path for prediction and classification. In this paper we have applied classification (KNN) and predicting (decision tree) model on thyroid data set to predict the new patient entry accurately.

KNN algorithm is used for classifying the thyroid disease with the related prioritized symptoms (attributes). User can predict and test their health with the symptoms (attributes). The user can predict the probability of future occurrence of thyroid disease with related symptoms before going to the hospital and check with the doctor using this proposed work.

8. Scope of the Paper

Our work analyzed the Thyroid prediction system using Machine learning framework. Our

methodology gives accurate result for both predictive modeling and information retrieving with more efficiently. Most of the people are willing to spend time and money to know the prediction for thyroid disease. Our system explains about people to know the prediction for thyroid disease and also to know the prediction details and level of disease anywhere in the world.
In this work we utilized KNN algorithm and other classification models to predict patients with Hypo Thyroid. The proposed system utilized demonstrated its execution in foreseeing with the best outcomes in provisions of accuracy and least execution time. We study the effectiveness and performance analysis of our proposed system with an experiment set, consisting of scalability, quality and accuracy.

References

[1] S.Umadevi, Dr. K.s.jeenmarseline research scholar principal department of computer science Sri Krishna arts and Science College, Coimbatore, India "Applying classification algorithms to predict thyroid disease" volume 07, issue 10, 2017.

[2] Sudipto guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, member, ieee, and liadan o'callaghan, "clustering data streams: theory and practice ieee transactions on knowledge and data engineering, vol. 15, no. 3, may/June 2013".

[3] Irina IoniŃă Informatics, Computer Science, Mathematics, and Petroleum-Gas University of Ploieşti, Romania, Liviu IoniŃă. LiviuIoniŃă Informatics, Computer Science, Mathematics, and Petroleum-Gas University of Ploieşti, Romania , Liviu IoniŃă, "Prediction of Thyroid disease using data mining techniques" vol. 1, no. 3 august 2016.

[4] Roshan Banu D, P.G. Student, Department of Computer Science, Stella Maris College, Tamilnadu, India, Sharmeli K C, Assistant Professor, Department of Computer Science, Stella Maris College, Tamilnadu, India, "Classification Model Using Random Forest and SVM to Predict Thyroid Disease" ,Volume:4 Issue: 2, International Conference on Advancements in Computing Technologies - ICACT 2018.

[5] Farhad soleimanian gharehchopogh, computer engineering department, urmia branch, islamic azad university, Iran maryam molany and freshte dabaghchi mokri,computer engineering department, science and research branch, islamic azad university, west azerbaijan, iran," Using artificial neural network in diagnosis of thyroid disease: a case study'',international journal on computational sciences & applications (ijcsa) vol.3, no.4, august 2013.