Thyroid Data Prediction using Data Classification Algorithm

Ammulu K.
Research Scholar
Rayalaseema University, Kurnool

Venugopal T.
Associate Professor & Head
JNTUH College of Engineering, Sultanpur, Medak, Telangana

Abstract

Thyroid is the major disorder occurs due to the lack of thyroid hormone among women than man. The test report of thyroid includes number of attributes such as TSH, T3, TT4, T4U and more. Manually determining the disorder for number of peoples test report is not easier. So, using the data mining approach will made this task simpler by predicting the disorder from the large dataset. Traditionally, Linear Discriminant Analysis (LDA) data mining technique is used to predict the thyroid disorder. In our proposed work, the random forest approach is utilized to predict the hypothyroid disorder by collecting the dataset from UCI repository. The performance measure is calculated from the confusion matrix with the accuracy. The experimental result is obtained from the Weka tool.

Keywords: Random Forest, thyroid, classification, LDA

I. Introduction

Ten out of one Indians suffer from thyroid disorder. This disorder primarily happens for women at the age of 17-54. The extreme stage of thyroid leads to increase in blood pressure, maximize the cholesterol level, cardiovascular complications, decreased fertility, and depression. An Electronic Health Record (EHR) contains the digitally stored information about the health information about an individual which includes the observations, laboratory tests, diagnostic reports, medications, procedures, patient, identifying information, and allergies[3]. Thyroid hormone is produced by the thyroid gland which is one of the endocrine glands. The main function of this hormone is to accelerate the human body metabolism, burn calories, protein, and restrict the other hormonal gland while there is excessive secretion[4]. The thyroid gland

Identifying the thyroid disorder from the tested report is complex and tedious job which can be determined only by the experienced and knowledge. Traditionally, there are two approaches, one is examining the blood tests by lab technicians and the other is doctor's diagnosis based on the signs, symptoms, and physical examination of patient[1] to predict the thyroid. Since this is not easier to examine each and everyone report from the large dataset to predict the result. The main task is to predict the thyroid disorder with better accuracy.

Thyroid gland secretes thyroid hormones to control the body's metabolic rate. The malfunction of thyroid hormone will leads to thyroid disorders. The thyroid or the thyroid gland is an endocrine gland. The thyroid gland releases thyroxine (T4) and triiodothyronine (T3) into the blood stream as the principal hormones. The functions of the thyroid hormones are to regulate the rate of metabolism and affect the growth. There are two most common problems of thyroid disorder or thyroid disease. They are Hyperthyroidism - releases too much thyroid hormone into the blood due to over active of thyroid and Hypothyroidism - when the thyroid is not active and releases too low thyroid hormone into the blood[2].

The existing work is carried out using the LDA algorithm which has the main disadvantage of LDA does not work well if the design is not balanced (i.e. the number of objects in various classes are (highly) different). The LDA is sensitive to overfit and validation of LDA models is at least problematic. (However other methods as RDA, ANN, SVM etc. are even worse). LDA is not applicable (inferior) for non-linear problems (separation of orange- banana shape point clouds, class in class situations). The proposed work includes:

- The thyroid dataset collection,
- Classifying using the Random forest approach,
- Implementing it in weka tool.

II. Related works

In paper [1], proposed a precise technique for detecting the thyroid by utilizing the back propagation algorithm. Artificial Neural Network is developed using the back propagation of error to identify the preliminary thyroid prediction. ANN is trained subsequently for testing the experimentally, but not the same training sets. The training can be done in two ways as supervised learning and unsupervised learning. The experimental result is carried out in MATLAB Neural Network Toolbox Software. This provides better performance than the simple gradient descent algorithm.

In paper [2], classification approaches are discussed that are utilized for the prediction of class label. This classification of dataset helpful for the prediction of various diseases from large volume of patient's dataset. Diabetic's dataset is used for the classification based on the decision table from the support and confidence to obtain better accuracy. The naïve bayes and fuzzy KNN are processed together for the medical dataset which provides better accuracy.

In paper[3], Ling Chen et al., proposed a graph-based semi-supervised learning algorithm called SHG-Health (Semi-supervised Heterogeneous Graph on Health) to predict the risky patients from the electronic health record. The Cause Of Death(COD) database are prepared the high risk dataset from the GHE database. The risky patients are classified using the semi-supervised learning with label propagation which includes the patient personal details, metal report and physical report.

In paper[5], Sudesh Kumar and Nancy proposed a clustering and data mining technique. The normalization approach is used to retrieve the efficient information with efficient factor. This approaches are Min-Max, Z-Scaling, decimal scaling normalization. The K-means clustering algorithm is utilized to cluster in less time. The clustering process helps to detect the homogeneous groups of objects based on the values of their attributes. The Z-score normalization technique is combined with the K-means clustering technique.

III. PROPOSED METHODOLOGY

A. *Dataset Description*

The dataset is 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. The attributes are shown in the table below:

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. *Random Forest Approach*

The dataset collected from the source is classified using the random forest algorithm. 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 such as CART and C4.5. The main advantage of the random forest algorithm are as follows:

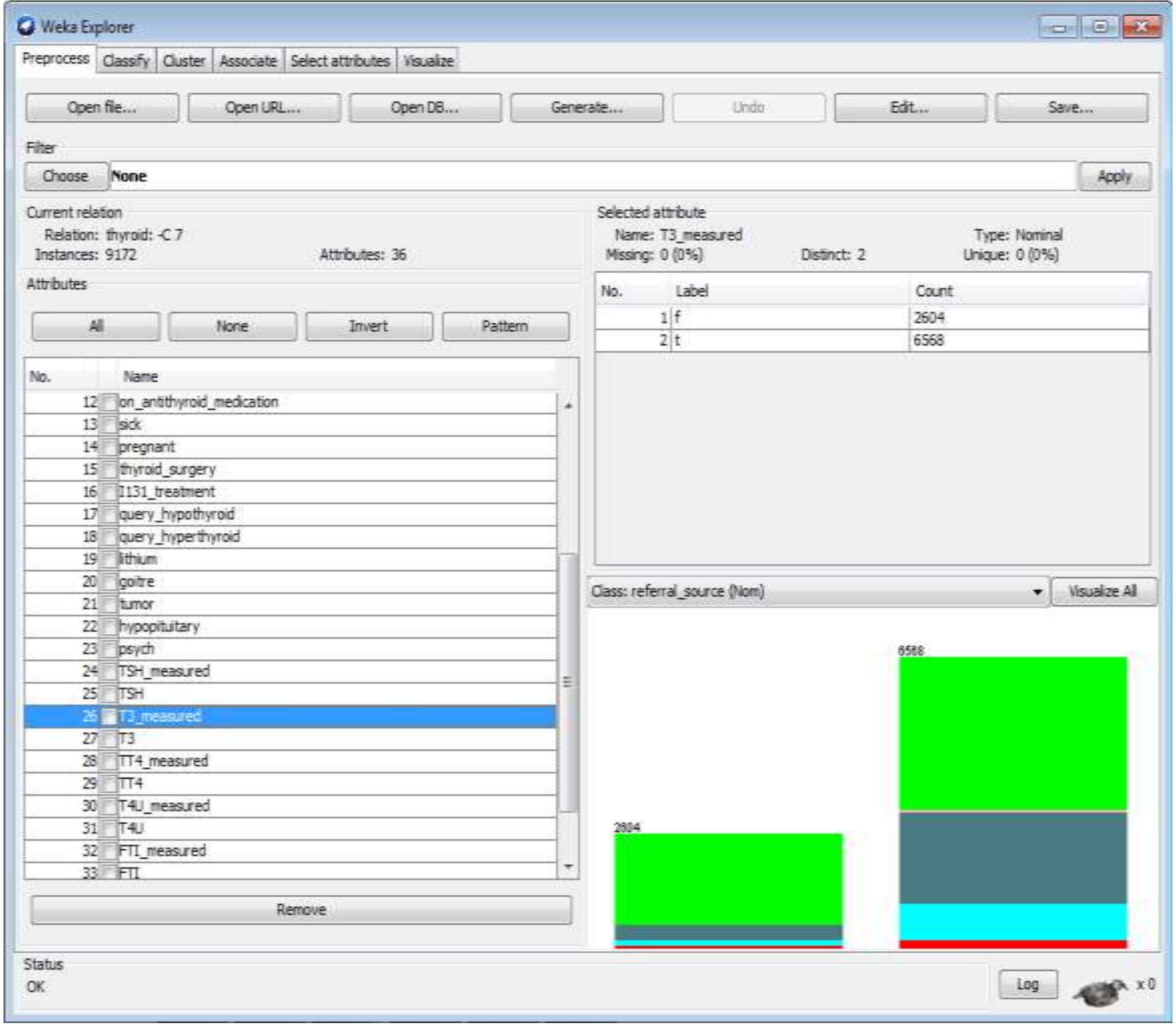
1) Its accuracy is as good as Adaboost and sometimes better.
2) It's relatively robust to outliers and noise.
3) It's faster than bagging or boosting.
4) It gives useful internal estimates of error, strength, correlation and variable importance.
5) It's simple and easily parallelized.

Steps involved in the random forest algorithm are as follow

- Select number of trees (Tn) to grow.
- Select Vm number of variables which is used to split each node. Vm is the number of input variables.
- Make trees to grow(decisions), for each tree the following is done:
  - Build a sample of size S obtained from the N training cases and allow it to grow.
  - While growing a tree at each node, select Vm variables in random from M which is used to find the best split.
  - Grow the tree to a maximal extent where there is no pruning.
- The classification point K collects the votes from every tree in the forest and then use majority voting to decide on class label.

IV. Experimental result

Weka is an open source data mining tool and it is developed by University of Waikato in New Zealand where the data mining algorithms are designed using the java language. Weka performs data mining task which contains number of machine learning algorithm. It includes process such as data processing and visualization, attribute selection, classifications, prediction (nearest neighbour), model evaluation, clustering, association rules.



Filter
None

Current relation
Relation: thyroid:-C 7
Instances: 9172
Attributes: 36

Selected attribute
Name: T3_measured
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count
1	f	2604
2	t	6568

Attributes

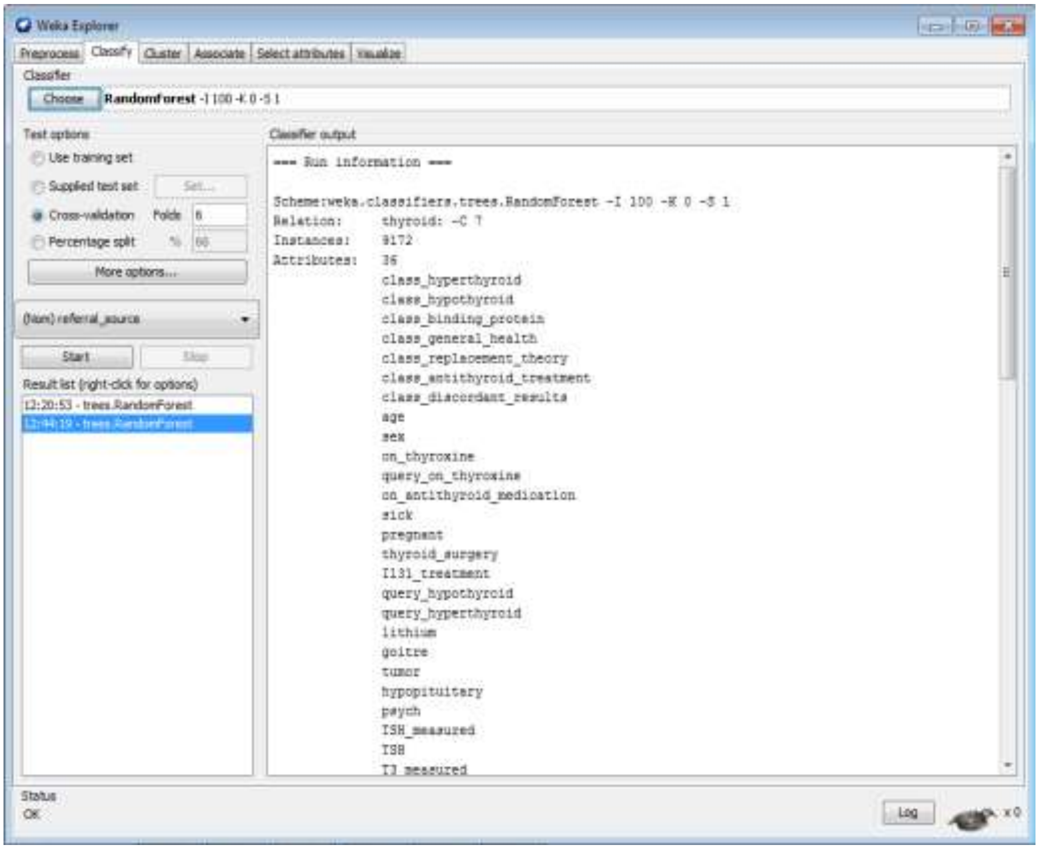
No.	Name
12	on_antithyroid_medication
13	sick
14	pregnant
15	thyroid_surgery
16	I131_treatment
17	query_hypothyroid
18	query_hyperthyroid
19	lithium
20	goitre
21	tumor
22	hypopituitary
23	psych
24	TSH_measured
25	TSH
26	T3_measured
27	T3
28	TT4_measured
29	TT4
30	T4U_measured
31	T4U
32	FTI_measured
33	FTI

Class: referral_source (Nom)

Status
OK

Fig. 1: Represent the preprocessing stage of the thyroid dataset

The result is obtained from the weka tool itself where the accuracy, precision, recall and F-measures are calculated in it.



Classifier: **RandomForest** -I 100 -K 0 -S 1

Test options
○ Use training set
○ Supplied test set   Set...
● Cross-validation   Folds 6
○ Percentage split   % 66

(Nom) referral_source

Result list (right-click for options)
12:20:53 - trees.RandomForest
12:44:19 - trees.RandomForest

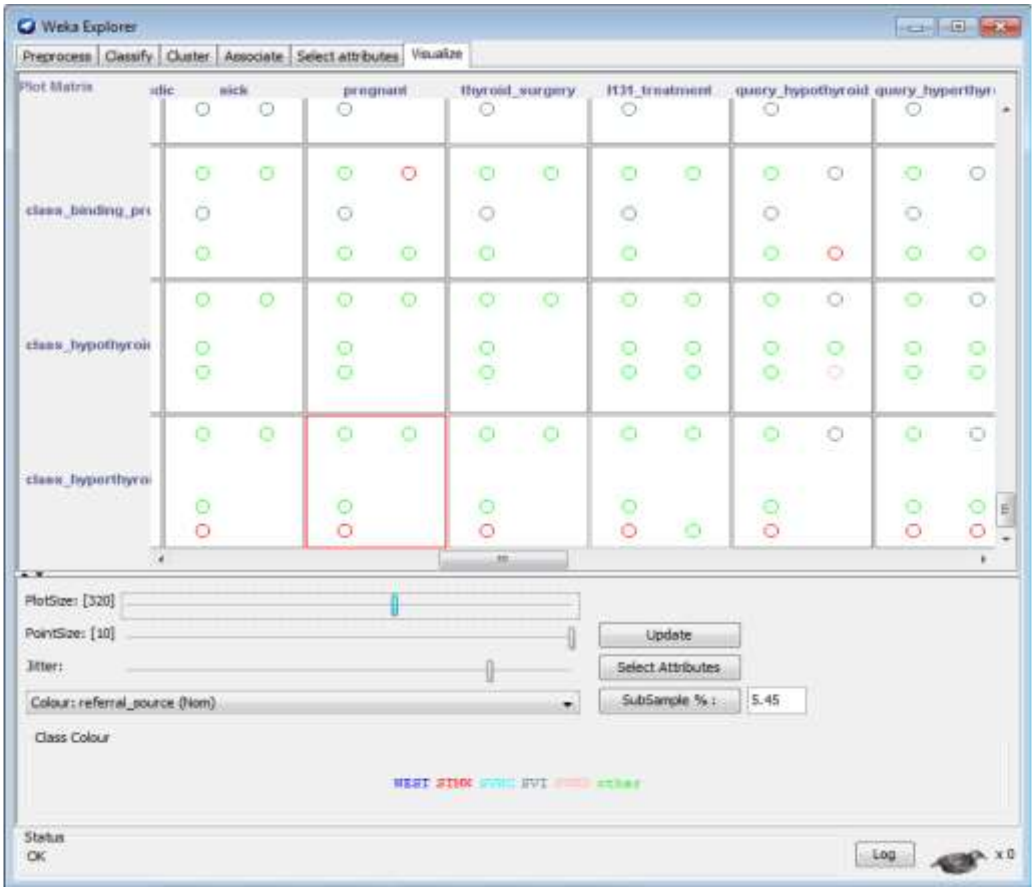
Classifier output

```
=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 100 -K 0 -S 1
Relation:     thyroid: -C 7
Instances:    9172
Attributes:   36
              class_hyperthyroid
              class_hypothyroid
              class_binding_protein
              class_general_health
              class_replacement_theory
              class_antithyroid_treatment
              class_discordant_results
              age
              sex
              on_thyroxine
              query_on_thyroxine
              on_antithyroid_medication
              sick
              pregnant
              thyroid_surgery
              I131_treatment
              query_hypothyroid
              query_hyperthyroid
              lithium
              goitre
              tumor
              hypopituitary
              psych
              TSH_measured
              TSH
              T3_measured
```

Status
OK

Fig. 2: Represent the classification of dataset using the random forest algorithm



PlotSize: [320]
PointSize: [10]
Jitter:
Colour: referral_source (Nom)
SubSample % : 5.45

Class Colour
WEST STMW SVHC SVI SVHD other

Status
OK

Fig. 3: Visualize the classification of the thyroid dataset

Table - 2
Confusion matrix Result for Different K value

K=n	*10*	*8*	*6*	*3*
Accuracy	*70.519 %*	*71.086 %*	*71.160 %*	*71.162 %*
TP rate	*0.705*	*0.711*	*0.712*	*0.712*
FP rate	*0.318*	*0.312*	*0.312*	*0.318*
Precision	*0.698*	*0.701*	*0.701*	*0.705*
Recall	*0.705*	*0.711*	*0.712*	*0.712*
F-Measure	*0.690*	*0.696*	*0.696*	*0.695*

V. Conclusions

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. In this paper, hypothyroid disorder is predicted using the random forest approach from data mining technique. The experimental result provides improved accuracy, precision, recall and F-measure by comparing the random forest with LDA algorithm. Future work is applied on validating the multiple disease dataset simultaneously like heart disease, diabetics, and many.

References

[1] Prerana, Parveen Sehgal, Khushboo Taneja, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", published in International Journal of Research in Management, Science & Technology, Vol. 3, No. 2, April 2015.

[2] Parneet Kaur, Deepak Aggarwal, "Classification of Medical Dataset using Hybrid Feature Selection & Enhanced Decision Table Classification Approach", International Journal of Science and Research, Volume 6 Issue 3, March 2017.

[3] Ling Chen, Xue Li, Quan Z. Sheng, Wen-Chih Peng, "Mining Health Examination Records - A Graph-based Approach", IEEE Transactions On Knowledge Discovery And Engineering, 2016.

[4] I Md. Dendi Maysanjaya, Hanung Adi Nugroho, Noor Akhmad Setiawan, "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases", in International Seminar on Intelligent Technology and Its Applications, IEEE, 2015.

[5] Sudesh Kumar, Nancy, "Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, 2014.

[6] M. Z. Islam, "EXPLORE: A Novel Decision Tree Classification Algorithm," Data Security and Security Data, Lecture Notes in Computer Science, vol. 6121, pp. 55-71, 2012.

[7] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," Boston, U.S.A.: Pearson Education Inc., 2006, ch. 4, pp. 151-154.

[8] R. Polikar, "Ensemble Based Systems in Decision Making," IEEE Circuits and Systems Magazine, vol. 6, pp. 21-45, Third Quarter, 2006.

[9] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp 832-844, August 1998

[10] Hossam M. Zawbaa, Maryam Hazman, Mona Abbass, Aboul Ella Hassanien, "Automatic fruit classification using random forest algorithm", 14th International Conference on Hybrid Intelligent Systems,pp. 164 – 168, 2014.

[11] Jiawei Hanl, Yanheng Liu, Xin Sun, "A Scalable Random Forest Algorithm Based on MapReduce", IEEE 4th International Conference on Software Engineering and Service Science, pp. 849 – 852, 2013.

[12] B.V.S Dheeraj Reddy, Mounika Booreddy, "Classification And Clustering Medical Datasets By Using Artificial Neural Network Models", Publications Of Problems & Application In Engineering Research – Paper, Vol 04, Special Issue 01, 2013.

[13] Dr. G. Rasitha Banu, M.Baviya, "Predicting Thyroid Disease Using Datamining Technique", International Journal of Modern Trends in Engineering and Research, 2014.

[14] S. Anto, Dr.S.Chandramathi, "Supervised Machine Learning Approaches for Medical Data Set Classifcation - A Review", InternatIonal Journal of Computer Science & Technology, Vol. 2, Issue 4, Oct. - Dec. 2011.

[15] Sudesh Kumar, Nancy, "Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, 2014.

[16] K.Saravana Kumar, Dr. R. Manicka Chezian, "Support Vector Machine And K- Nearest Neighbor Based Analysis For The Prediction Of Hypothyroid", International Journal of Pharma and Bio Sciences, pp. 447 – 453, 2014.

[17] Noor Azah Samsudin ; Aida Mustapha ; Mohd Helmy Abd Wahab, "Ensemble classification of cyber space users tendency in blog writing using random forest", Innovations in Information Technology (IIT), 2016 12th International Conference on 28-30 Nov. 2016

[18] R. Shreyas, D.M Akshata, B.S Mahanand, B. Shagun, C.M Abhishek, "Predicting Popularity of Online Articles using Random Forest Regression", Second International Conference on Cognitive Computing and Information Processing (CCIP), 2016.

[19] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html