Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4– No12, January 2016 – www.caeaccess.org

Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique

G. Rasitha Banu, PhD
Assistant Professor,
Department of HIT & HIM,
Public Health and Tropical Medicine,
Jazan University, Saudi Arabia.

ABSTRACT
Thyroid disease is very common disease in human. Nowadays most of the women suffering from thyroid disease than male. There are two types in thyroid disease like hypothyroid and hyperthyroid disease. These diseases giving many side effects such as weight gain, weight loss, stress and so on to our human body .If this disease is detected in earlier stage, then physician can give proper treatment to the patients .Data Mining is playing important role in predicting many diseases. Classification is one the most significant Technique in Data Mining. It is a supervised learning. It is used to classify predefined data sets. Health care data are having exponential growth in volume and complexity. Data mining Technique is mainly used in healthcare organizations for decision making, diagnosing disease and giving better treatment to the patients. In this paper hypothyroid disease is to be predicted using data mining. The dataset used for the study on hypothyroid is taken from UCI repository. Classification of this thyroid disease is a considerable task. An experimental study is carried out using Linear Discriminant Analysis (LDA) to achieve better accuracy. There are many data mining classification Algorithms such as CART, REP Tree, and J48 and so on. The LDA Algorithm gives accuracy is 99.62% with cross validation k=6.

Keywords
Thyroid disease, classification, Linear Discriminant Analysis, Decision tree

1. INTRODUCTION
Hypothyroidism is a relatively common problem worldwide often with insidious onset and is relatively asymptomatic. Lung volumes are usually normal, but few studies have shown findings suggestive of restrictive pattern of impairment. Weight gain independently of physical activity is frequently associated with hypothyroidism. Hypothyroidism can have numerous effects on the respiratory system. Lung volumes are usually normal or mildly reduced, but maximal breathing capacity and diffusing capacity are usually reduced. Hyperthyroidism and hypothyroidism are common conditions that have lifelong effects on health. Hypothyroidism is a condition in which the body lacks sufficient thyroid hormone. Since the main purpose of thyroid hormone is to "run the body's metabolism," it is understandable that people with this condition will have symptoms associated with a slow metabolism. Hypothyroidism (underactive thyroid or low thyroid) is a condition in which your thyroid gland doesn't produce enough of certain important hormones. Women, especially those older than age 60, are more likely to have hypothyroidism. Hypothyroidism upsets the normal balance of chemical reactions in your body. It seldom causes symptoms in the early stages, but, over time, untreated hypothyroidism can cause a number of health problems, such as obesity, joint pain, infertility and heart disease. Hyperthyroidism (overactive thyroid) is a condition in which your thyroid gland produces too much of the hormone thyroxin. Hyperthyroidism can accelerate your body's metabolism significantly, causing sudden weight loss, a rapid or irregular heartbeat, sweating, and nervousness or irritability. Data Mining is the process of semi-automatically analyzing large databases to find patterns. Classification is a data mining (machine learning) technique used to predict group membership for data instances [5, 6]. In this research, LDA Algorithm is used to predicate thyroid disease. A data set with 29 features downloaded from UCI repository site is used for the experimental purpose, entire work is carried out with WEKA open source software under Windows 7environment. K-fold validation is also performed.

2. DATA SET DESCRIPTION
The data set used for experimental purpose is down loaded from the website (http://repository. seasr.org/Datasets/UCI/arff). The data set has 3772 instances from which 3481 belongs to category negative,194 belongs to category compensated hypothyroid , 95 belongs to primary hypothyroid category while 2 belongs to category secondary hypothyroid. The last attribute is the class, hence there are 29 features in all, which will be used to classify the data .The detail of data set is shown below.

Table 1: Hypothyroid Data Set

Data Description SN	Attribute Name	Value Type
1	age	continuous
2	sex	M,F
3	on thyroxin	F,T
4	query on thyroxin	F,T
5	on anti thyroid medication	F,T
Data Description SN	**Attribute Name**	**Value Type**
6	sick	F,T
7	pregnant	F,T
8	thyroid surgery	F,T
9	i131treatment	F,T
10	query hypothyroid	F,T

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 4– No12, January 2016 – www.caeaccess.org

11	query hyperthyroid	F,T
12	lithium	F,T
13	goiter	F,T
14	tumor	F,T
15	hypo pituitary	F,T
16	psych	F,T
17	TSH measured	F,T
18	TSH	continuous
19	T3 measured	F,T
20	T3	continuous
21	TT4 measured	F,T
22	TT4	continuous
23	T4U measured	F,T
24	T4U	continuous
25	FTI measured	F,T
26	FTI	continuous
27	TBG measured	F,T
28	TBG	continuous
29	referral source	WEST, STMW, SVHC, SVI,SVHD 0ther

3. METHODOLOGY

3.1 Algorithm Description

The following developed model can assist doctors to take proper decision to give better treatment to the patients. The LDA Algorithm is described below.

3.1.2 Lda Algorithm

Linear Discriminant analysis (LDA) is a generalization of Fisher's linear Discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before classification. [4, Wikipedia]

3.1.2 K – Foldcross -Validation

In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data [4]. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly the data is loaded into weka software. After pre-processing, various data mining classification techniques are applied on the data set to develop the predictive models. And the system is trained using the training set. After the system is trained it is tested using up to 10 fold cross validation method. Evaluation is performed using certain performance measures.

4. EXPERIMENTS WITH WEKA

Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, feature selection and visualization. Weka can downloaded from the website. The preprocessing stage of weka is shown in below fig.1

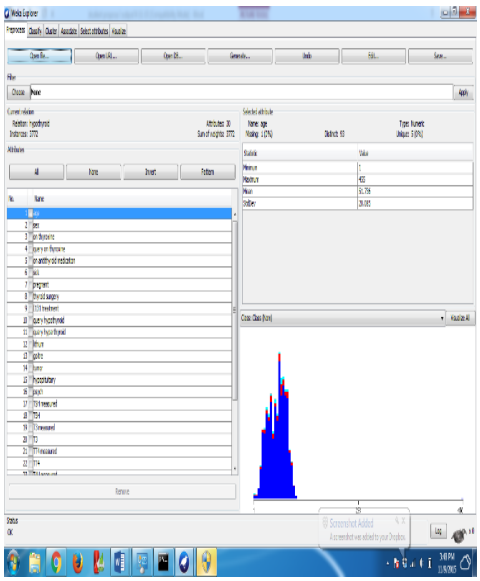

Figure 1. preprocessing stage.

5. RESULT AND ANALYSIS

There are in total 3772 records in the hypo thyroid data set. All the records are classified as negative, Compensated hypothyroid, primary hypothyroid or secondary hypothyroid. In our experiment data is supplied to classifier of LDA Algorithm. The following Table 2 represents confusion matrix for LDA Algorithm with k=6 fold.

Table2. Confusion matrix for LDA algorithm fork=6 fold

Target class	**Negative**	**Compensated hypothyroid**	**Primary hypothyroid**	**Secondary hypothyroid**
Negative	3475	4	2	0
Compensated Hypothyroid	0	191	3	0
Primary Hypothyroid	2	1	92	0
Secondary Hypothyroid	2	0	0	0

The following Table 3 depicts detailed accuracy for different k-folds for LDA algorithm.

Table 3: accuracy for different k-folds for LDA algorithm.

K=n	Accuracy	TP-rate	FP-rate	Precision	Recall	ROCarea
K=10	99.49	0.995	0.013	0.995	0.995	0.997
K=8	99.57	0.996	0.013	0.995	0.996	0.996
K=n	**Accuracy**	**TP-rate**	**FP-rate**	**Precision**	**Recall**	**ROCarea**
K=6	99.62	0.996	0.013	0.996	0.996	0.996
K=4	99.60	0.996	0.006	0.996	0.996	0.996
K=2	99.39	0.994	0.016	0.993	0.994	0.996

6. CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is a very challenging task in the field of health care. Various data mining techniques has proven to be very helpful in decision making. In this paper we have applied LDA data mining classification techniques is used to classify the hypothyroid disease. K-fold cross validation is also performed. The LDA Algorithm gives 99.62% accuracy with k=6 folds cross validation. As a future work the same technique is used to apply for other disease datasets such as heart disease, diabetes and so on.

7. REFERENCES

[1] Jiawei Han, Kamber Micheline (2009). Data mining: Concepts and Techniques, Morgan Kaufmann Publisher.

[2] Paper 094-2010 Building Decision Trees from Decision Stumps Murphy Choy, University College Dublin Peter Flom, Peter Flom Consulting

[3] "UCI Machine Learning Repository of machine learning database", University of California, school of Information and Computer Science, Irvine. C.A. http://www.ics.uci.edu/.

[4] www.wikipedia.org.

[5] Dr. G. Rasitha Banu, Baviya "A study on Thyroid disease using Data Mining Technique", IJTRA Journal, aug 2015.

[6] Dr .G .Rasitha Banu, Baviya " predicting Thyroid disease using Data Mining Technique ", IJMTER journal, March 2015.

[7] Pandey, Rohit Miri , Tandan "Diagnosis And Classification Of Hypothyroid Disease Using Data Mining", International Journal of Engineering Research & Technology (IJERT),Vol. 2 Issue 6, June – 2013

[8] D .Lavanya, Dr .Usha Rani, " Performance Evaluation of Decision Tree Classifiers on Medical Datasets", *International Journal of Computer Applications (0975 – 8887),Volume 26– No.4, July 2011*

[9] 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.,0ct.2014.www.ijpbs.net

[10] Prerana, Parveen Sehgal, Khushboo Taneja "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", International Journal of Research in Management, Science & Technology (E-ISSN: 2321-3264) Vol. 3, No. 2, April 2015.

[11] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Elsevier, Australia, 2006

[12] WWW.Academia.edu