Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques

Shivanee Pandey
M.Tech. C.S.E. Scholar
Department of Computer Science & Engineering
Dr. C.V. Raman Institute of Science & Technology
Bilaspur, India

Rohit Miri
Asst. Professor
Department of Computer Science & Engineering
Dr. C.V. Raman Institute of Science & Technology
Bilaspur, India

S. R. Tandan
Asst. Professor
Department of Computer Science & Engineering
Dr. C.V. Raman Institute of Science & Technology
Bilaspur, India

Abstract

Thyroid is one of the most common disease found in human being which cause many other side effects in human body it is of two type : Hyperthyroid and Hypothyroid. Prior diagnosis of Thyroid is very important and beneficial for the betterment of human life ,an early diagnosis and detection of this disease can help human being to fight against this disease . In medical science traditional techniques are being used to diagnose various types of disease, doctor's experience plays very important role in this context, an experienced doctor can have better capabilities to diagnose disease as compare to less experienced doctors. Data mining is a technique which can be used to develop expert system for the classification of medical data and are widely accepted by the researcher to develop intelligent tools .In this paper various data mining techniques like Bayes net, Multilayer perceptron, RBF network ,C4.5, CART, REP tree, decision stump are used to develop classifier for diagnosis of hypothyroid 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 7 environment. K-fold validation is also performed for each technique. Results reflected that a model with k=6 is performing better than others, accuracy in this case is obtained as 99.60% which is acceptable in range for diagnosis thyroid disease.

1. Introduction

Most crucial and challenging task in the field of medical science is to identify or diagnose disease at correct time. Disease can be cured if diagnosed correctly on time. It helps doctors in proper treatment of patient. Only doctors and physicians can identify the disease using their experience, medical reports. During Earlier day's diseases are determined by the symptoms exhibited by the patients and various medical tests but now doctors are taking help of various intelligent systems also.

Thyroid is one of the most common diseases found in human being, it is not a deadly disease, but it is chronic disease which can give rise to other diseases. Thyroid is a butterfly-shaped gland, which is located at the bottom of the throat responsible for producing two active

thyroid hormones, levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) that affect some functions of the body. These functions include stabilizing body temperature, blood pressure and regulating the heart rate. People suffering from thyroid gland tend to fall sick due to under or over production of hormones from this gland.

Hypothyroidism and hyperthyroidism are a result of an imbalance of thyroid hormone. Hypothyroidism is simply not enough thyroid hormone and hyperthyroidism is too much. Either imbalance affects the metabolism in the body. Hypothyroidism causes a reduction in stroke volume and heart rate causing lowered cardiac output with a decrease in heart sounds. Hypothyroidism is condition that underlies most chronic degenerative diseases and hormone irregularities and results in a weakened immune system.

Data mining is becoming strategically important tool for many organizations including healthcare sector having huge amount of data. Data mining, *the extraction of hidden predictive and descriptive information from large databases,* is a powerful new technology with great potential to help healthcare sector to focus on the most important information in their data warehouses. Data mining will be the cornerstone in detecting disease. Data mining is a technique which can be used to develop expert system for the classification of medical data.

A significant work is done by bindi et.al. (2012) to analyze the common features available in the two data sets using ANOVA and MANOVA they have also used several data mining based classification techniques like Naïve Bayes, C4.5,Back Propagation ,K-NN and SVM with various feature subsets obtained with the help of ranking based algorithm to classify BUPA and Indian liver disorder data sets collected from Andhra Pradesh (AP Data Set),India[1][4,5].

H.S.Hota[1],his paper explores various data mining techniques to design an ensemble model for classification of breast cancer related health care data. A testing accuracy of model show the efficiency of ensemble model. Models are also measured in terms of sensitivity and specificity. Feature subsets are obtained after applying ranking based feature selection algorithm and models are tested on these data sets.

M. R. Nazari Kousarrizi, F.Seiti, and M. Teshnehlab[2], They proposed a method which has two stages. In the first stage sequential Forward selection (SFS), sequential backward selection (SBS) and Genetic Algorithm are used as feature selection methods as a preprocessing step. In the second stage, SVM is used to classify thyroid data.

Satish N. Kulkarni, Dr. A. R. Karwankar[6],they propose modified fuzzy hyperline segment clustering neural network (MFHLSCNN) for classification of thyroid disease diagnosis. The MFHLSCNN algorithm is suitable for clustering and classification.

In this paper various data mining techniques like Bayes net, Multilayer perceptron, RBF network, C4.5, CART, REP tree, decision stump, are used to develop classifier for diagnosis and classification of hypothyroid 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 7 environment. K-fold validation is also performed with each technique..

2. Data and Methodology

The data set used for experimental purpose is downloaded from university of California of Iravin (UCI) repository site (web source http://www.archive.ics.uci.edu/ml/datasets.html). Details of data set is given in Section 2.1

2.1. Data set Description

The data set used for experimental purpose is downloaded from university of California of Iravin (UCI) repository site (web source http://www.archive.ics.uci.edu/ml/datasets.html). 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 in table 1.

Table 1: Hypothyroid Data Set

Attribute	Value
Age	Integer
Sex	Male(M),Female(F)
On thyroxine	False(f),true(t)
Query on thyroxine	False(f),true(t)
On antythyroid	False(f),true(t)

Sick	False(f),true(t)
Pregnant	False(f),true(t)
Thyroid surgery	False(f),true(t)
T131 treatment	False(f),true(t)
Query Hypothyroid	False(f),true(t)
Query Hyperthyroid	False(f),true(t)
Lithium	False(f),true(t)
Goiter	False(f),true(t)
Tumor	False(f),true(t)
Hypopitutory	False(f),true(t)
Psych	False(f),true(t)
Tsh measured	False(f),true(t)
TSH	Real
T3 measured	False(f),true(t)
T3	Real
TT4 measured	False(f),true(t)
TT4	Real
T4U measured	False(f),true(t)
T4U	Real
FTI Measured	False(f),true(t)
FTI	Real
TBG Measured	False(f),true(t)
TBG	Real
Referal source	SVHC, other, SVI, STMW, SVHD
Class	negative, compensated hypothyroid, primary hypothyroid, secondary hypothyroid

2.2Algorithm description

Our developed model can assist doctors to take proper decision in the treatment of patients. Various classification algorithms are used to build classifiers, these classifiers are decision stump, Bayes net, Multilayer perceptron, C4.5, CART, REP tree**.** These techniques are discussed in more detail as below:

Multi Layer perceptron (MLP) – it is a feed forward neural network with one or more layers between input and output layer. Feed forward means that data flows in one direction from input to output layer (forward). This type of network is trained with the back propagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi Layer Perceptron can solve problems which are not linearly separable [13].

Radial basis function (RBF) networks are feed-forward networks trained using a supervised training algorithm. They are typically configured with a single hidden layer of units whose activation function is selected from a class of functions called basis functions. While similar to back propagation in many respects, radial basis function networks have several advantages. They usually train much faster than back propagation networks [12].

Decision stumps are basically decision trees with a single layer. As opposed to a tree which has multiple layers, a stump basically stops after the first split. Decision stumps are usually used in population segmentation for large data. Decision stumps can be useful for a variety of purpose and by linking them together; we can build a full decision tree for modelling purposes [8].

Reduced error pruning Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The dual goal of pruning is reduced complexity of the final classifier as well as better predictive accuracy by the reduction of over fitting and removal of sections of a classifier that may be based on noisy or erroneous data. One of the simplest forms of pruning is reduced error pruning. Starting at the leaves, each node is replaced with its most popular class. If the prediction accuracy is not affected then the change is kept. While somewhat naive, reduced error pruning has the advantage of simplicity and speed [5].

C4.5 Algorithm Just like Classification and Regression Tree, the C4.5 algorithms recursively visits each node, selecting the optimal split, until no further splits are possible. The steps of C4.5 algorithm for growing a decision tree is given below

1. Choose attribute for root node by using attribute selection measure Gain Ratio.
2. Create branch for each value of that attribute.
3. Split cases according to branches.
4. Repeat process for each branch until all cases in the branch have the same class or all attributes are processed [3].

CART (Classification and Regression Tree) :CART classification algorithm which is based on decision tree

induction (Jiwai H. and Micheline Kamber,2009) which is a learning of decision trees from class label training tuples.The Classification and Regression (CART) tree method uses recursive partitioning to split the training records into segments with similar output field values. The CART tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. CART uses Gini index splitting records measures in selecting the splitting attribute. Pruning is done in CART by using a training data set. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary (only two subgroups)[1].

Bayesian Net: Bayesian net (Han, J., & Micheline, K., 2006) is statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Let X is a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P (H|X), the probability that the hypothesis H holds given the observed data sample X.P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X. Bayesian classifier is very popular and applied for health care domain by many authors (Bendi V.R ,2011) (Alaa M.Elsayad,2010)[1]

K-fold cross-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. 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 once [11].

3. System Implementation

Building a predictive model for diagnosis of hypothyroid is a quaternary classification as the data set contains 4 classes i.e. negative, compensated hypothyroid, primary hypothyroid or secondary hypothyroid. Hypothyroid dataset was downloaded from UCI data repository site and 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 10 fold cross validation method. Evaluation is performed using certain performance measures and calculated results are presented in table 2 in terms of accuracy, precision, recall, TP-rate, FP-rate, F-measure and ROC area.We have also performed k-fold cross validation for different values of k for C4.5 algorithm.

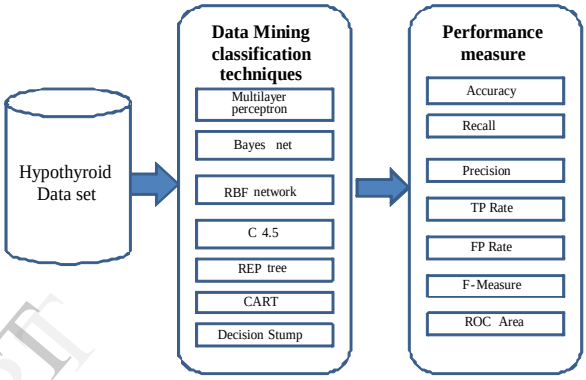


Figure1.System architechture

4. Experiments with Weka

Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, feature selection and visualization. Weka can be downloaded from the website http://www.cs.waikato.ac.nz/ml/weka/. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement.[10]

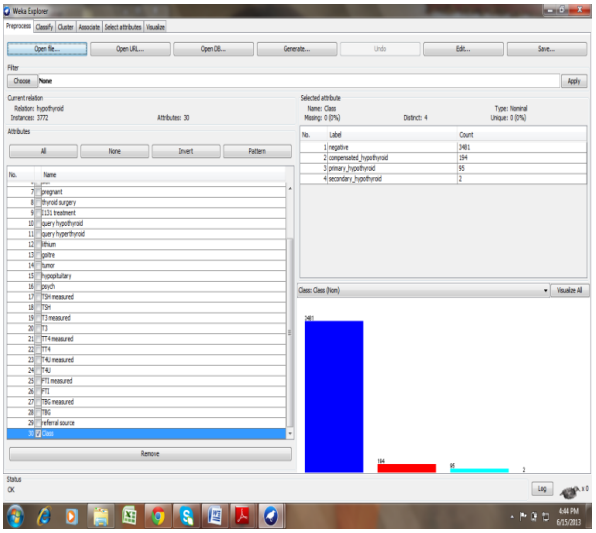


Figure 2. Screenshot of weka during preprocessing stage.

5. Result and Analysis

There are in total 3772 records in the hypothyroid dataset. All the records are classified as negative, compensated hypothyroid, primary hypothyroid or secondary hypothyroid. In our experiment data is supplied to each classifier one by one and a list of detailed accuracies are obtained as depicted in table 2 for all the classifiers .Each row of the table represents detailed accuracy for respective classifier. From this table it is clear that REP tree and C4.5 are performing well as compared to others. Table 3 represents confusion matrix for C 4.5 algorithm for 6 folds cross validation. Table 4 depicts detailed accuracy for different k-folds for C4.5 algorithm.

Table 2. Detailed Accuracy by class for various classifiers

Algorithm	Accuracy	Precision	Recall	F-Measure	ROC-Area	TP-rate	FP-rate
Mutlilayer perceptron	94.035	0.937	0.94	0.938	0.891	0.94	0.398
RBF Network	95.228	0.945	0.952	0.946	0.898	0.952	0.407
Bayes net	98.59	0.986	0.986	0.986	0.997	0.986	0.08
C4.5	99.57	0.995	0.996	0.995	0.993	0.996	0.019
CART	99.54	0.995	0.995	0.995	0.993	0.995	0.007
Decision stump	95.38	0.95	0.954	0.948	0.981	0.954	0.009
REP tree	99.57	0.995	0.996	0.996	0.993	0.996	0.007

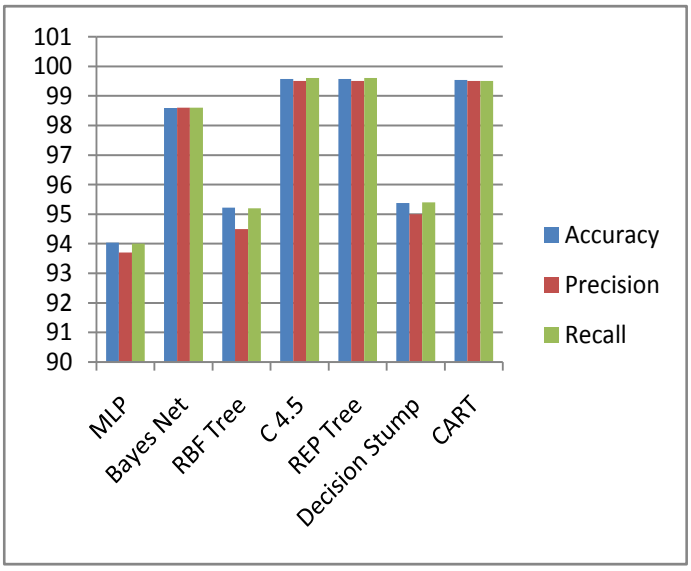
Table3. Confusion matrix for C4.5 algorithm for k=6 fold (Best performance among all classifiers)

Target class	Negative	Compensated hypothyroid	Primary hypothyroid	Secondary hypothyroid
Negative	**3476**	**2**	**3**	**0**
Compensated hypothyroid	**1**	**192**	**1**	**0**
Primary hypothyroid	**3**	**3**	**89**	**0**
Secondary hypothyroid	**2**	**0**	**0**	**0**

Table 4. Detailed Accuracy for different k-folds for C4.5 algorithm

K=n	Accuracy	TP-rate	FP-rate	Precision	Recall	ROC area
K=2	99.469	0.995	0.01	0.994	0.995	0.996
K=4	99.57	0.996	0.013	0.995	0.996	0.994
K=6	99.60	0.996	0.019	0.995	0.996	0.994
K=8	99.54	0.995	0.019	0.995	0.995	0.993
K=10	99.575	0.996	0.019	0.995	0.996	0.993

Figure 3. Bar graph showing Accuracy,Precision and Recall for different predictive model.



6. Conclusion and Future Scope

Diagnosis of disease is a very challenging and crucial task in the field of health care. Data mining techniques will be the gem stone in healthcare sector. Various data mining techniques has proven to be very helpful in decision making. In this paper we have applied various data mining techniques to develop classifier for diagnosis and classification of hypothyroid disease. k-fold cross validation is also performed. As a future work Dimensionality reduction can be applied to the data set so that it will reduce number of test and time required to diagnose the disease.

7. References

[1]. H.S.Hota, Diagnosis of Breast Cancer Using Intelligent Techniques, International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-3, January 2013.

[2]. M. R. Nazari Kousarrizi, F.Seiti, and M. Teshnehlab, "An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification," International Journal of Electrical & Computer Sciences IJECS-IJENS, Vol. 12 No. 01, February 2012, pp. 13-20.

[3]. Poonam Gupta, Rohit Miri, S.R.Tandan, Decision Tree Applied For Detecting Intrusion, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May - 2013 ISSN: 2278-0181.

[4].Bendi V.R.Prasad, M.S.Babu and Venkateswarlu N. B.(2012),*A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis*, International Journal of Computer Science Issues, Vol.9. Issue 3, No. 2 ,PP 506-516.

[5]. Bendi V. R., Prasad M. S. Babu and Venkateswarlu N. B.(2011) ,*A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis*, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, PP 101-114.

[6]Satish N. Kulkarni, Dr. A. R. Karwankar, *Thyroid disease detection using modified fuzzy hyperline segment clustering neural network*, International Journal of Computers & Technology, Volume 3 No. 3, Nov-Dec, 2012

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

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

[9]. "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/ .

[10]. http://www.cs.waikato.ac.nz/ml/weka/

[11]. http://en.wikipedia.org

[12].http://www.eee.metu.edu.tr/~halici/courses/543LectureNotes/lecturenotes-pdf/ch9.pdf

[13].http://neuroph.sourceforge.net/tutorials/MultiLayerPerceptron.html