A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques

K. Rajam

Department of Computer Science, Bishop Heber College, Trichy,Tamilnadu
amutharajamk@gmail.com

R. Jemina Priyadarsini

Assistant Professor in Computer Science, Bishop Heber College, Trichy, Tamilnadu
jemititus@gmail.com

Abstract— The medical data having a huge amount of data set these data set classifications are highly challenging Task data mining is one of the most gratifying and challenging area of applications or method to knowledge Discovery. Thyroid is one of the unpredictable diseases in medical field. Thyroid hormones control the body's metabolic rate. Moreover, data mining techniques has been applied in various sector the classification results of the medical data set which helps the way of treatments to the patients. Most famously used techniques are Decision tree which is used to reduce noisy data, Naïve Bayse is help to get high speed and accuracy rate when applies large data base, Backpropagation gives high accuracy neural network when applied complex non-liner, Support Vector Machine is easy to extended and pattern reorganization. This paper presents the survey on Diagnosis of Thyroid Disease.

Keywords— Thyroid disease,Decision tree, Naïve bayse,SVM, Backpropagatin

I. INTRODUCTION

Thyroid Disease diagnosis is one of the very difficult and deadly tasks, because it needs lots of experience and knowledge. The traditional ways for diagnosis thyroid disease is doctor's examination or a number of blood tests. Mainly task is to provide disease diagnosis at early stages with higher accuracy. Data mining plays a vital role in medical field for disease diagnosis. It offers lot of classification techniques to predict the disease accuracy. Hospitals and clinics gathered a large amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases.

A. Data Mining

Simply stated, data mining refers to "extracting" or "mining" knowledge from large number of data[1], which is used to process the inconsistence data automatically and find the best data. Data Mining having two categories, they are Data Mining in descriptive – summarize the general properties of the data in database, Data Mining in Predictive – to predict the inference of the present data[2].

B. Image Mining

Image Mining uses different algorithms and techniques to process the digital image data. The most common image process steps are i. image preprocessing, ii. Segmentation, iii. Feature extraction, iv. Feature selection and classification. For example, thyroid US image is taken. Image preprocessing which helps to remove the noisy and inconsistence data. Segmentation process helps to segment the image used to apply object and boundaries. Each segment is characteristics by a color, texture, and intensity. Classification process use different techniques to predict the data like Decision tree, SVM, naïve bayse.

Thyroid Diagnosis is based on signs, symptoms and physical examination of patient. Predictive data mining plays a vital role in disease diagnosis. This paper shown survey on thyroid disorder from various papers and gives the idea for the future work. The rest of this paper organized as below: section II as different thyroid disorders and their symptoms, section III as Literature survey, section IV as Classification Techniques in data mining and section V as conclusion and reference.

II. DIFFERENT THYROID DISEASE AND THEIR SYMPTOMS

Dr. Sahai says about 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[3].

There are four main types of thyroid diseases hyperthyroidism (too much thyroid hormone), hypothyroidism (too little thyroid hormone), benign (noncancerous) thyroid disease and thyroid cancer(malignant). The symptoms of hypothyroidism includes fatigue, mental fogginess and forgetfulness, feeling excessively cold, constipation, dry skin, fluid retention, non specific aches and stiffness in muscles and joints, excessive or prolonged menstrual bleeding (menorrhagia), and depression. Hyperthyroidism can be observed with different signs and symptoms. Common symptoms of hyperthyroidism includes excessive sweating, heat intolerance, increased bowel movements, tremor (usually a fine shake), nervousness, agitation, rapid heart rate, weight loss, fatigue, decreased concentration and irregular and scant menstrual flow. [4]

III. LITERATURE SURVEY

Nikita Sigh and Alka Jindal [5] have concluded that SVM is better classifier as compared to KNN and Bayesian. Accuracy of SVM is about 84.62%. KNN found the nearest neighborhood automatically. It represented by the graph each vertices having object. Bayesian based on the probability classification which gives the sample data belongs to a class.

Nasrul Humaimi Mahmood and Akmal [6] have provided   a simple way to determine the thyroid lobes in the thyroid ultrasound image using a MATLAB. The thyroid have two lobes: right and left lobe. The image undergoes the contrast enhancement to suppress speckle that are present in US image. The enhanced image is used for segmentation of the thyroid region by local region-based active contour. The thyroid region is segmented into two parts, which are right and left with the active contour method separately. Transverse view of thyroid ultrasound image is used therefore the measurements only involve the width, depth and area of the thyroid region. The result of thyroid measurement is calculated in pixel unit and then it is converted in centimetre (cm) unit. It shows that from five samples, different people have different size of thyroid.

Deepika Koundal1 et al. [7] have provided the information about the existing automatic tools which are available to formulate the disease diagnosis part easier with efficient way. Also different performance evaluation metrics are studied. The future developments and trends are also investigated.

Edgar Gabriel et al. [8] have proposed two parallel versions of a code that are used for texture-based segmentation of thyroid FNAC images which is a critical first step in realizing a fully automated CAD solution. An MPI version of the code is Developed to exploit distributed memory compute resources such as PC clusters

M. Savelonas et al. [9] have proposed active contour model named Variable Background Active Contour model. It is used for the detection of thyroid nodules in ultrasound images. The new model offers edge independency, smoothingless operation, ability for topological changes. It gives more accuracy when compared to the Active Contour Without Edges model. Accuracy can be improved by introducing as background a limited image subset which appropriately changes shape to reduce the effects of background inhomogeneity.

Preeti Aggarwal et al[10] they suggested an automatic segmentation method. They have provided a summary of all the results obtained either by automatic tools as well as by applying specific algorithm (automatic) segmentation on both lung CT as well as on thyroid US .There are two tools: Analyze 10.0 and Mazda for segmentation of thyroid US images.

Eystratios .[11] have suggested a computer-aided diagnosis (CAD) system prototype named as TND (Thyroid Nodule Detector).It is used for the detection of nodular tissue in ultrasound (US) thyroid images and videos acquired during thyroid US examinations.

Mary C. Frates et al. [12] have provided in her paper US features associated with thyroid cancer. They have also suggested that which nodules should be subjected to US-guided fine needle aspiration and which thyroid nodules need not be subjected to fine-needle aspiration. Different image classifiers are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Fuzzy measures, Genetic Algorithms (GA), Fuzzy support Vector Machines (FSVM). The textural features in ANN helps to resolve misclassification. SVM is the best available machine learning algorithms in classifying high-dimensional data sets.

Won-Jin Moon et al. [13] have done in her paper the evaluation on the diagnostic accuracy of ultrasonographic (US) criteria for the depiction of benign and malignant thyroid nodules. It is done by using tissue diagnosis as the reference standard. They concluded that shape, margin, echogenicity and presence of calcification are important criteria for the discrimination of malignant from benign nodules.

Singh1 and Mrs Alka Jindal [14] in their paper have used the groups Benign (non-cancerous) and Malignant (cancerous) Thyroid Nodules images. The texture feature method like GLCM are very useful for classifying texture of images. These texture features are used to train the classifiers such as SVM, KNN and Bayesian. The experimental result shows the performance of the various classifiers and shows the best predictive value and efficiently identify the percentage of the non-cancerous or cancerous people.It also shows the best performance accuracy using the SVM classifier as compare to the KNN and Bayesian classifier. It uses the segmentation based algorithm localized based active contour (region based) method. It is basically to segment the local area of the images and to segment the nodule which gives the information about the type of nodule exist that is either benign or malignant. The classification accuracy is calculated in term of true positive (TP), true negative (TN), false positive (FP), false negative (FN) with respect to the ground truth.

IV. CLASSIFICATION TECHNIQUES IN DATAMINING

(i) Decision Tree

A Decision tree has three types of node, internal node denotes the test on an attribute, leaf node denotes the classes or class distributions, root node is the top most node in a tree.C4.5 and ID3 these are the two main algorithms used to construct the decision tree.

- ➤ **Advantages**
- Decision tree uses "tree pruning" approach to identify and remove noisy data from the branch and to improve the classification accuracy.
- The attribute with the highest normalized data is chosen to make a decision.
- Algorithm used continuous and discrete values.

- ➤ **Disadvantages**
- Efficiency and scalability are low when applied to mining of very large data bases.
- Decision tree construction inefficient when swapping sample data from main memory to cache memory.
- C4.5 algorithm can contain empty attributes, over fitting.

(ii) Naïve Bayesian Classification

Bayesian classification is derived from the Bayes theorem. It is also known as "Simple Bayesian Classifier". In this classifier each data sample is represented by an n- dimensional vector and measurements samples are formed by n- attributes. Suppose there are m classes, C_1, C_2 …..C_3 having a unknown data sample, X, Naïve Bayesian classifier will predict that X belongs to class having the highest probability conditioned on X. an unknown sample X to the class C_i.

$P(C_i| X) > P(C_j|X)$ for $1 \leq j \leq m$, $j \neq i$.

➢ **Advantages**

- Comparing other classifier naïve classifier gives high accuracy and speed when applied large data base.
- minimum error rate
- Suitable for large amount of data bases.

➢ **Disadvantages**

- It considering each attribute in each class separately.
- Zero conditional probability problem.

(iii) Backpropagation Neural Network

Backpropagation is based on the neural network algorithm. Here we are having the three kinds of layer, input layer- the inputs are fed into this layer, hidden layer – weighted outputs can be input to another hidden layer, number of hidden layer's arbitrary, but only one is used, output layer- hidden layers are give the input to the output layer, which emits the network's prediction. This network model helps to classifying new data.

➢ **Advantages**
- High accuracy neural networks are able to approximate complex non-liner mappings.
- Very flexible with the noisy and inconsistence data
- Easily updated with the present data

➢ **Disadvantages**
- Major disadvantage is knowledge representation
- Knowledge interpret is difficult for humans.
- Removing weighted links that decreases the accuracy of the given network.
- General method is not available.
- Difficult to select training dataset.

(iv) Support Vector Machine

Support Vector Machine is one type of learning system algorithm, which is used to perform classification more accurately. SVM used for two class classifier. The essence of SVM is hyper plane also known as "decision boundary or decision surface". This hyper plane separates the positive and negative of training data sample.

➢ **Advantages**
- SVM is easy to extended, useful to pattern reorganization, formulated quadratic optimization problem.

➢ **Disadvantages**
- It is suitable for real valued space
- For binary classification, it allow only two classes and for multiple class classification, it apply several strategies
- Hyper plane is very hard to use by the user

V. CONCLUSION

Thus the survey helps to identify the how data mining techniques to predict the thyroid disorder at earlier stage. Different Researchers have proposed different techniques to predict the thyroid disorder and different kinds of accuracy level as per used techniques. These techniques help to minimize the noisy data of the patient's data from the data bases. Algorithms such as Naïve bayes, decision tree, back propagation, Support vector machine are considered for the study. These algorithm gave the various result based on speed, accuracy, performance and cost. Also these effective classification data are helps to find the treatment of the patient. In future a better method to diagnose thyroid disease disorder can be found out with improvements in the existing methods.

References

[1] **Jiewai Han and Micheline Kamber** – Data Mining Concepts and Techniques, second edition

[2] **Bendi Venkata Ramana, M. Rendra Prasad Babu and N.B. Venkaeswarlu,** " A Critical Study of Selected Classification Algorithms for Liver Disease", International Journal of Database Management System(IJDMS), Vol.3, No.2, May 2011.

[3] **Dr. Sahai BS,** Thyroid Disorders[online]. Available : Http://www.homoeopathyclinic.com/articles/diseases/tyroid.pdf

[4] http://www.nlm.nih.gov/medlinep/us/ency/article/000371/htm

[5] **Nikita Singh, Alka Jindal** , "A Segmentation Method and Comparison of Classification Methods for Thyroid Ultrasound Images", International Journal of Computer Applications (0975 – 8887) Volume 50 – No.11, July 2012

[6] **Nasrul Humaimi Mahmood and Akmal**, "Segmentation and Area Measurement for Thyroid Ultrasound Image", International Journal of Scientific & Engineering Research Volume 2, Issue 12, December 2011

[7] **Deepika Koundle, Savita Gupta, Sukhwinder**., "Computer-Aided Diagnosis of Tyroid Nodule: A Review", International Journal of Computer science & Engineering Survey(IJCSES) Vol.3,No.4,August 2012

[8] **Edgar Gabriel, Viswanath Venkatesan and Shishir Shah,** "Towards high performance cell segmentation in multispectral fine needle aspiration cytology of thyroid lesions", computer methods and programs in biomedicine.

[9] **M. Savelonas, D. Maroulis and D. Likouvidis,** "Variable Background Active Contour Model for Automatic Detection of Thyroid Nodules in Ultrasound Images", Dept. of Informatics and Telecommunication

[10] **Preeti Aggarwal, Renu Vig, Sonali Bhadoria and C.G. Dethe,** "Role of Segmentation in Medical Imaging: A Comparative Study", International Journal of Computer Applications (0975 – 8887) Volume 29– No.1, September 2011

[11] **Eystraints G,**"TND: A thyroid Nodule Detection System for analysis of Ultrasound Image and Videos", Springer Science and Business Media, LLC 2010.

[12] **Mary C. Frates, Carol B. Benson, J.William Charboneau and Edmund S**. "Management of Thyroid Nodules Detected at US: Society of Radiologists in US consensus", conference statement management of thyroid nodules detected at US Volume 237, Number3.

[13]**Won-Jin Moon**." Benign and Malignant Thyroid Nodules: US Differentiation", Multicenter Retrospective Study Radiology: Volume 247: Number 3- June 2008.

[14] **Singh1 and Mrs. Alka Jindal**," A segmentation method and classification of diagnosis for thyroid nodules", IOSR Journal of Computer Engineering (IOSRJCE)ISSN:2278-0661 Volume1, issue 6, 2012.

[15] **Prerana, Parveen Sehgal, Khushboo Taneja**, "Predictive Data Mining for Diagnosis of thyroid Disease using Neural Network", International Jurnal of Resaearch in Management, Scienc & Technology(E- ISSN: 2321- 3264) Volume 3, No.2 April 2015

[16] **Ms. Priti Dhaygude, Mrs. S.M. Handore,** " A review of thyroid disorder detection using medical images", International Journal on Recent and Innovation Trends in Computing and Communication, Volume 2, Issues:12, December 2014