Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules

J. Jacqulin Margret
PG Scholar
Department of C.S.E.
Anna University of Technology, Coimbatore.

B. Lakshmipathi
Assistant Professor
Department of C.S.E.
Anna University of Technology, Coimbatore.

S. Aswani Kumar
Asst. Systems Engineer
Tata Consultancy Services, Chennai.

ABSTRACT
Thyroid gland secretes thyroid hormones to control the body's metabolic rate. The malfunction of thyroid hormone will leads to thyroid disorders. The under-activity and over-activity of thyroid hormone causes hypothyroidism and hyperthyroidism. This paper describes the diagnosis of thyroid disorders using decision tree attribute splitting rules. Since, decision tree attempts to follow one decision, it helps to classify the data in dataset according to aforesaid disorders. This method provides five different splitting criteria for the construction of decision tree. The splitting criteria are Information Gain, Gain Ratio, Gini Index, Likelihood Ratio Chi-Squared Statistics, Distance Measure. Among, the aforementioned splitting rules three rules belong to Impurity based splitting criteria and other two are Normalized Impurity based splitting criteria. As a result, the decision tree classifies the thyroid data-set into three classes of thyroid disorders.

Keywords
Thyroid Disorders, Decision Trees, Information Gain, Gain Ratio, Gini Index, Likelihood Ratio Chi-Squared Statistics, Distance Measure.

1. INTRODUCTION
The disease diagnosis can be done by attaining the correct classification. For this, the techniques used should be able to handle a large volume of data. Moreover, it is vital to make effective and reliable decision. In medical science, decision making plays a vital role to classify the diseases. This helps the physician to make decisions easily. For each task of decision making models, decision trees are significant to make reliable decision.

The thyroid or the thyroid gland is an endocrine gland. The thyroid gland releases thyroxine (T_4) and triiodothyronine (T_3) 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[1][2].

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The top most node in a tree is the root node. The construction of decision tree classifiers does not require any domain knowledge and appropriate for exploratory knowledge discovery. Decision tree induction can be used for simple and fast classification. Decision tree algorithms can be used for classification in many application areas such as medicine, astronomy, financial analysis, molecular biology [3][4].

This work reveals that the splitting criteria play a vital role in constructing decision trees. The five splitting criteria used in this work can be characterized as Impurity based and normalized impurity based. When the data-set is small, splitting criteria do not matter. But for large data-set, it is difficult to generate more accurate decision tree.

Thyroid dataset for this work was collected from UCI machine learning repository. It consists of 3 classes, 3772 training instances, 3428 testing instances and 21 attributes[5].

2. RELATED WORK
ANUPAM SKUKLA et al. [6] proposed the diagnosis of thyroid disorders using Artificial Neural Networks (ANNs). Three ANN algorithms; the Back propagation algorithm (BPA), the Radial Basis Function (RBFN), and the Learning Vector Quantization (LVQ) Networks have been used for the diagnosis.

LALE OZYILMAZ et al. [7] focused on proper interpretation of the thyroid data besides the clinical examination and complementary investigation is an important issue on the diagnosis of thyroid disease. Various neural network methods such as Multi Layer Perceptron (MLP) with back-propagation (BP) and fast back-propagation (FBP), Radial Basis Function (RBF), and adaptive Conic Section Function Neural Network (CSFNN) have been used and compared for the diagnosis of thyroid disease.

FATEMEH SAITI et al. [8] proposed Support Vector Machines and Probabilistic Neural Network for classification. These methods depend on powerful classification algorithms, in order to deal with redundant and irrelevant features. Genetic Algorithm is an optimization technique. It operates iteratively on a population of structures each which represent a candidate solution to the problem properly encoded as a string of symbols (e.g. Binary). Here, three basic genetic operators guide this search : selection, crossover and mutation. This is an iterative search. The PNN is a supervised neural network and it can be widely used in the area of pattern recognition, nonlinear mapping and likelihood ratios.

C.ORDONEZ et al. [9] compared association rules and decision tree rules. Association rules look for hidden patterns making them suitable for discovering predictive rules involving subsets of the medical data set. The results are association rules are irrelevant, appear at low support and discovered rules are large in number. A decision tree represents a predictive model of the data set. In fact, the large number of discovered association rules may acquire rule summarization. A decision tree can have at least 50% prediction accuracy and generally above 80% accuracy for binary target variables.

KALLE et al.[10] used Logical similarity measure for the detection of liver, thyroid, diabetes and cancer.

Comparison among neural network and comparison between support vector machine and artificial neural network

algorithms have been used for diagnosis of thyroid diseases[11][12].
Selvathi et al. [13] used ultrasonic images for the classification thyroid disease.

3. METHODOLOGY USED

This section explains about the algorithm and formulae used for this work. The block diagram for this work is shown in figure 1.

3.1 Description of data-set

The dataset consists of 21 attributes and 3 classes. They are class 1, class 2 and class 3[5].

3.2 Algorithm used for the Construction of Decision Tree

The following algorithm has been used for the construction of the decision tree[3].

Input:
Data partition, D (training dataset)
Attribute list
Attribute selection method
Output:
Decision tree
Method:
Create a node N;
If tuples in D are all of the same class, C then
Return N as a leaf node labeled with the class C;
If attribute list is empty then
Return N as a leaf node labeled with the majority class in D;
Apply attribute selection method to find the best splitting rule;
Label node N with splitting criterion;
Attribute list = attribute list – splitting attribute;
For each outcome j of splitting criterion
Let D_j be the set of data tuples in D satisfying outcome j;
If D_j is empty then
Attach a leaf labeled with the majority class in D to node N;
Else attach the node returned by generating decision tree to node N;
Endfor
Return N;

3.3 Impurity Based Splitting

It is a measure of impurity using probability function. If the probability vector has a component of one, then it is said as pure. The impurity level reaches the maximum, when all the components are equal[14].

3.3.1 Information Gain

This splitting criteria minimizes the information needed to classify the attributes and also minimizes the expected number of tests needed to classify the attributes. It is calculated using the formula,

$$Gain(A) = Info(D) - Info_A(D)$$

where Info(D) and $Info_A(D)$ are given as,

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$Info_A(D) = \sum_{i=1}^{v} \left(\frac{|D_i|}{|D|} \right) \times Info(D_i)$$

The attribute which have high information gain is chosen as the splitting attribute and it is the root node. ID3 uses IG as its attribute selection measure. But Information Gain bias towards the attributes on large number of values. Thus, overfitting occurs.

3.3.2 Gini Index

This kind of attribute selection can be used in CART decision tree. Gini index measures the impurity as,

$$Gini(D) = 1 - \sum_{i=1}^{m} (p_i)^2$$

Where p_i is the probability that a tuple in D belongs to class C_i. It is computed by $\frac{|C_{i,D}|}{|D|}$.

3.3.3 Likelihood Ratio Chi-Squared Statistics

The statistical significance of information gain criteria can be measured using this rule,

$$G^2(a_i, S) = 2\ln(2)|S|.InformationGain(a_i, S)$$

3.4 Normalized Impurity Based Criteria

The impurity based criteria is biased towards attributes with larger domain values. So, this is used to "normalize" the impurity-based criteria[14].

3.4.1 Gain Ratio

Gain ratio can be defined as a modification of the information gain, that reduces its bias on large attributes i.e., it uses normalized information gain. Gain ratio is large when data can be evenly spread. It corrects the information gain, by taking the number of information that helps to branch an instance (intrinsic information). It is measured using,

$$Gainratio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

where

$$SplitInfo(D) = -\sum_{i=1}^{v} \left(\frac{|D_i|}{|D|} \right) \times \log_2 \left(\frac{|D_i|}{|D|} \right)$$

But it may choose an attribute just because its intrinsic information is very low.

3.4.2 Distance Measure

It is similar to Gain ratio. This normalizes in a different way.

$$DM(D) = \frac{Gini(D)}{\left(-\sum_{j=1}^{v} \sum_{i=1}^{m} p_{ij} . \log_2 (p_{ij}) \right)}$$

The above mentioned splitting criteria were used for decision tree construction and rules were extracted. Then, comparison was made among these five splitting rules.

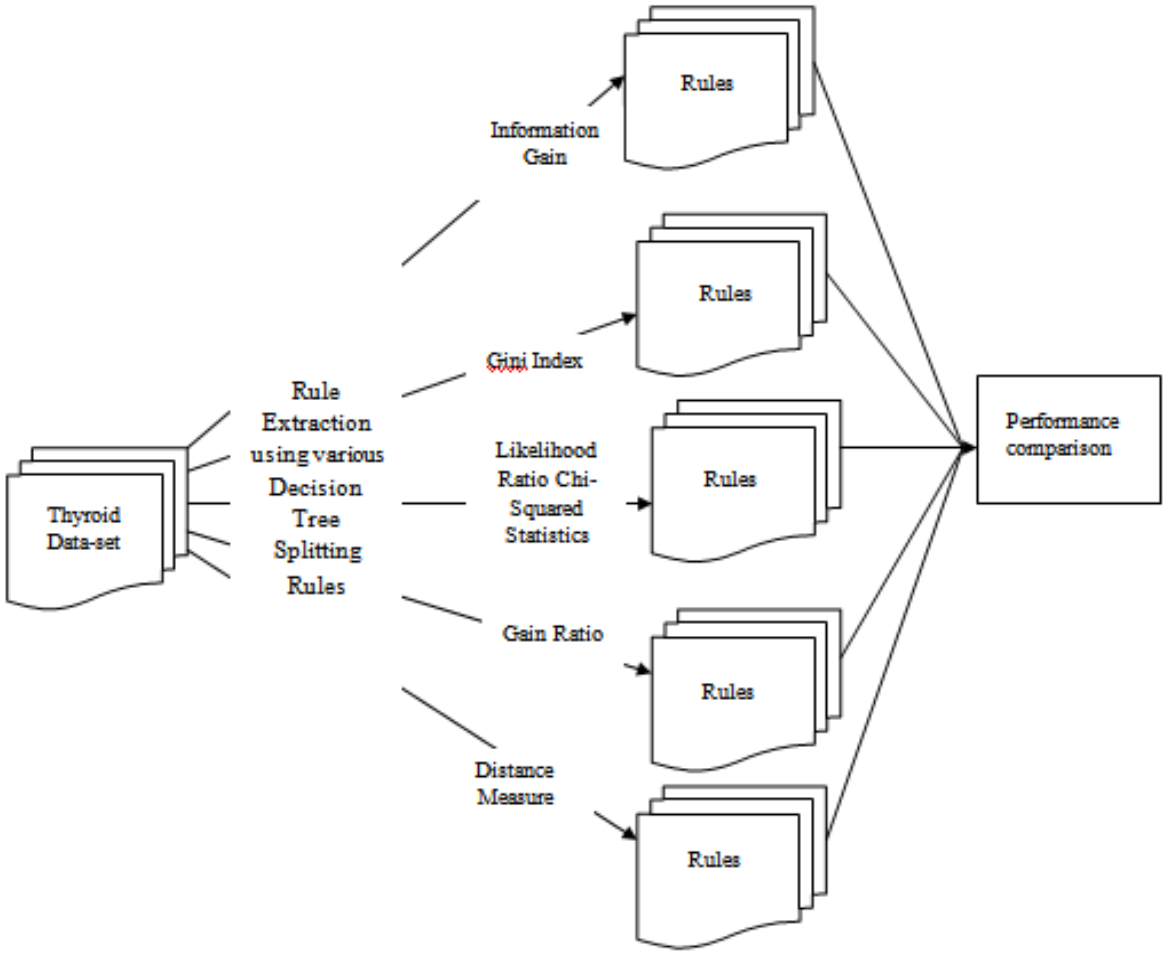

Figure 1: Block Diagram of the methodology used in this work

4. SIMULATED RESULTS

This section shows the rules and graph obtained from this work. The thyroid data-set named "thyroid disease" obtained from UCI machine learning repository[4] have been used for the experiment. It was trained by developing code in JAVA. The following results were obtained.

Several rules corresponding to the thyroid disorders were achieved using decision tree with five splitting criteria. The following are the sample rules obtained using Distance Measure splitting criteria.

4.1 Observed Rules

4.1.1 Sample Training rules

hypopituitary >= 0.0,qthyroxine = 0.0,sick = 0.0,pregnant = 0.0 then class = 3.0

hypopituitary >= 0.0,age = 0.3,sex = 1.0,thyroidsurgery = 0.0,i131 = 0.0 then class = 3.0

hypopituitary >= 0.0,lithium = 0.0,tumor = 0.0,t3 = 0.011,tt4 = 0.113,t4u = 0.113 then class = 3.0

hypopituitary >= 0.0,qthyroxine = 0.0,thyroidsurgery = 0.0,i131 = 0.0,hyperthyroid = 0.0 then class = 2.0

hypopituitary >= 0.0,age = 0.71,sex = 0.0,thyroidsurgery = 0.0,i131 = 0.0 then class = 3.0

hypopituitary >= 0.0,lithium = 0.0,tumor = 0.0,t3 = 0.026,tt4 = 0.108,t4u = 0.096 then class = 3.0

hypopituitary >= 0.0,age = 0.3,sex = 1.0,thyroidsurgery = 0.0,i131 = 0.0 then class = 3.0

hypopituitary >= 0.0,qthyroxine = 0.0,sick = 1.0,pregnant = 0.0 then class = 2.0

hypopituitary >= 0.0,qthyroxine = 0.0,sick = 0.0,pregnant = 0.0 then class = 3.0

hypopituitary >= 0.0,lithium = 0.0,tumor = 0.0,t3 = 0.028,tt4 = 0.111,t4u = 0.131 then class = 2.0

hypopituitary >= 0.0,lithium = 0.0,tumor = 0.0,t3 = 0.011,tt4 = 0.113,t4u = 0.113 then class = 3.0

hypopituitary >= 0.0,qthyroxine = 0.0,sick = 0.0,pregnant = 0.0 then class = 2.0

4.1.2 Sample Testing rules

Maximum Distance Measure attribute is 1.4440392797448176-thyroidsurgery

Rule

thyroidsurgery >= 0.0,age = 0.85,sex = 0.0,i131 = 0.0 then class = 1.0

thyroidsurgery >= 0.0,age = 0.52,sex = 1.0,i131 = 0.0 then class = 3.0

thyroidsurgery >= 0.0,qthyroxine = 0.0,sick = 0.0,pregnant = 0.0 then class = 2.0

thyroidsurgery >= 0.0,qthyroxine = 0.0,i131 = 0.0,hyperthyroid = 0.0 then class = 3.0

thyroidsurgery >= 0.0,age = 0.57,sex = 1.0,i131 = 0.0 then class = 3.0

hypopituitary >= 0.0,age = 0.29,sex = 0.0,thyroidsurgery = 0.0,i131 = 0.0 then class = 2.0

4.2 Performance Comparison

The following graph shows the comparison of accuracy (%) among the splitting rules.

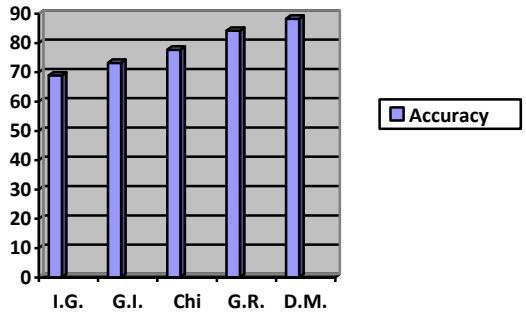

Figure 2 : Accuracy comparison graph for the splitting rules

While comparing the accuracy of splitting rules, Distance Measure has more accuracy.

Figure 3 represents the true positive and false positive values (%).

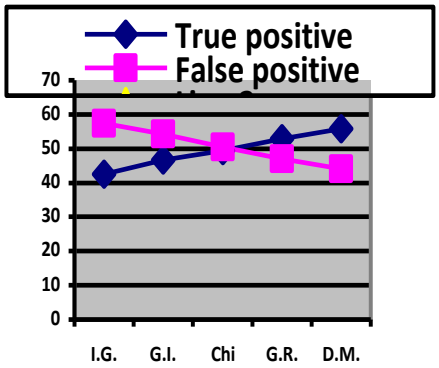

Figure 3 : True and False positive values comparison graph for the splitting rules

The following table gives the accuracy, specificity and sensitivity of the splitting rules.

Table 1 : Specifies the values of Accuracy, Specificity and sensitivity

Splitting rules	Accuracy (%)	Sensitivity (%)	Specificity (%)
Information Gain (I.G.)	68.925	42.59	57.4
Gini Index (G.I.)	73.222	45.76	53.23
Chi Square (χ^2)	77.769	49.44	49.55
Gain Ratio (G.R.)	84.190	52.92	46.08
Distance Measure (D.M.)	88.340	55.54	43.46

But while comparing the training time, Gini Index has less training time and Likelihood chi squared statistics has more training time.

5. CONCLUSION

Various splitting rule for decision tree attribute selection had been analyzed and compared. This helps to diagnosis the thyroid diseases through the extracted rules. From this work, it is clear, that normalized based splitting rules have high accuracy and sensitivity or true positive rate. This work can be extended for any medical datasets. Further enhancement can be made by using various optimization algorithms or rule extraction algorithms.

6. ACKNOWLEDGMENTS

Our thanks to Mr. Joseph Raj and Ms. Fatima who have contributed towards development of this work.

7. REFERENCES

[1] Dr. Sahni BS, Thyroid Disorders [online]. Available : http://www.homoeopathyclinic.com/articles/diseases/tyroid.pdf

[2] thyroid :www.wikipedia.org/thyroid

[3] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques. *Published by* Elsevier 2006.

[4] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, Decision trees: An overview and their use in medicine. In *Proceedings of* Journal Medical System 2002, pages 445–463.

[5] http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/

[6] Anupam Shukla, Prabhdeep Kaur, Ritu Tiwari and R.R. Janghel, Diagnosis of Thyroid disease using Artificial Neural Network. In *Proceedings of IEEE IACC 2009*, pages 1016-1020.

[7] Lale Ozyilmaz and Tulay Yildirim, Diagnosis of Thyroid disease using Artificial Neural Network Methods. In *Proceedings of ICONIP 2002*, pages 2033-2036.

[8] Fatemeh Saiti and Mahdi Aliyari, Thyroid Disease Diagnosis based on Genetic algorithms using PNN and SVM. In *Proceedings of IEEE 2009*.

[9] C. Ordonez, Comparing association rules and decision trees for disease prediction. In *Proceedings of* Int. Conf. Inf. Knowl. Manage., 2006, pages 17–24.

[10] Kalle Saastamoinen and Jaakko Ketola, Medical Data Classification using Logical Similarity Based Measures. In *Proceedings of* IEEE 2006.

[11] Saeed Shariati and Mahdi Motavalli Haghighi, Comparison of anfis neural network with several other anns and support vector machine for diagnosing hepatitis and thyroid diseases. In *Proceedings of* IEEE IACSIT 2010, pages 596-599.

[12] Modjtaba Rouhani and Kamran Mansouri, Comparison of several ANN architectures on the Thyroid diseases grades diagnosis. In P*roceedings of* IEEE IACSIT 2009, pages 526-528.

[13] D.Selvathi and V.S.Sharnitha, Thyroid Classification and Segmentation in Ultrasound Images Using Machine Learning Algorithms. In *Proceedings of* IEEE ICSCCN 2011, pages 836-841.

[14] Lior Rokach and Oded Maimon, Decision Trees.Tel-Aviv University, pages 165-192.