ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Classification of Thyroid Disease Using ACO-MST Techniques.

Poornima M, Sumathi A, and Meganathan S*.

Department of CSE, SRC, SASTRA UNIVERSITY, KUMBAKONAM, TN, INDIA

ABSTRACT

Thyroid hormones are essential for the function of every cell in the body. Abnormality of thyroid hormones produces thyroid disorders. Thyroid glands that makes and stores thyroid hormones that help regulate the growth rate of metabolism in the human body. The under-activity and over-activity of thyroid hormone causes hypothyroidism and hyperthyroidism. Although the effects can be unpleasant or uncomfortable, most thyroid problems can be managed well if properly diagnosed and treated. This model mainly focused for the diagnosis of thyroid disease based on various decision tree classification algorithms. Different types of decision trees are formed by applying ACO algorithm, J48 and REP tree for diagnosis (hypothyroidism and hyperthyroidism) of thyroid disease accurately with the given data set. From the experimental results, the proposed model gives better accuracy.

Keywords: Decision Tree, Classification, ACO, Minimum Spanning Tree

Corresponding author

ISSN: 0975-8585

INTRODUCTION

The **thyroid gland** is one of the largest endocrine glands in the body, and consists of two connected lobes. It is found in the anterior neck, below the laryngeal prominence. The thyroid gland controls rate of use of energy sources, protein synthesis, and controls the body's sensitivity to other hormones. It participates in these processes by producing thyroid hormones, the principal ones being thyroxine (T_4) and triiodothyronine (T_3), which is more active. These hormones regulate the growth and rate of function of many other systems in the body. T_3 and T_4 are synthesized from iodine and tyrosine. The thyroid also produces calcitonin, which plays a role in calcium homeostasis. The thyroid may be affected by some frequent thyroid diseases. In addition, the thyroid gland may also develop several types of nodules and cancer. [1][2]

Classification is an important data mining technique where huge data are classified into clusters and is used to retrieve relevant information. Classification leads to clustering of datasets with rules. These rules are mined appropriately based on the features which are extracted from the datasets. Unfortunately the features may be relevant or irrelevant depending on the users or the nature of request. To classify or cluster the valid or certain data, there are different approaches like DTL, Rule based Classification, Naive Bayes Classification and many more techniques. The study takes a modified genetic version which is a descendant of the Ant Colony Optimization algorithm and J48 algorithm and uses Decision Trees like REP Tree, C4.5, and J48. This is used to test the effectiveness of the data and its impact in the dimensionality purpose. The data obtained is uncertain in that uncertainty occurs in the data because of the imprecise measurement of the results, like all scientific results. The main task is to handle the uncertainty of the data in order to classify or cluster it. As for the thyroid dataset has been used to shows Type I and Type II.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules.[3]

PROPOSED METHODOLOGY

J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. This algorithm has a few base cases.

1. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
2. None of the features provide any information gain. In this case,J48 creates a decision node higher up the tree using the expected value of the class.
3. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. The basic of pruning of this algorithm is it used REP with back over fitting. It kindly sorts values for numerical attribute once and it handling the missing values with embedded method by C4.5 in fractional instances. In this algorithm we can see it used the method from C4.5 and the basic REP also count.

In graph theoretic clustering methods to extract the appropriate features from given dataset. In particular, the minimum spanning tree (MST) based clustering algorithms used, because they do not assume

that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice.

Steps to be followed for constructing decision trees

1. Remove the irrelevant features.
2. Construct a minimum spanning tree from relative ones.
3. Features are divided into clusters by using graph-theoretic clustering methods
4. Most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features
5. Partition the **Minimum Spanning Tree [MST]** and select the representative features.

A cluster consists of all features and each cluster is treated as a single feature and dimensionality is drastically reduced. This system describes the decision tree attempts to follow one decision, it helps to classify the data in thyroid dataset. The rule mining and classification process consists of training set that are analyzed by a classification rules.

Feature Removal:

It is composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones. The latter removes redundant features from relevant ones via choosing representatives from different feature clusters. The main task is to handle the uncertainty of the data in order to classify or cluster it. Tree partitions classify the dataset and select the representative feature. This produces the final subset.

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

- 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 Dj be the set of data tuples in D satisfying outcome j;
- If Dj 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;
- End for Return

MST-ACO Optimization

The minimum spanning tree (MST) based clustering algorithms. They do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. This step involves partitioning the MST and select representative features

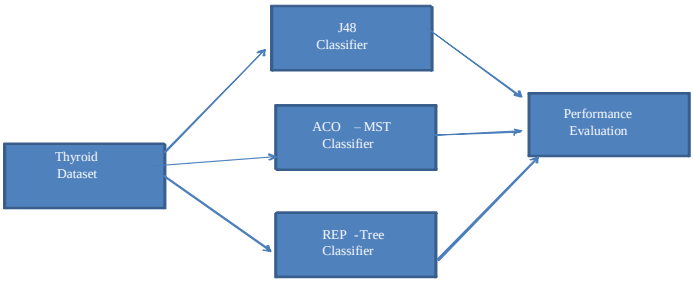
ACO Algorithm

1. Set k for each variable.
2. Set C evap, Cinc and q0.
3. Generate a random initial ant=> X(best)
4. Generate a random initial matrix F with the condition that all fij are the same
5. Calculate PC following the equation (6).
6. For j=0 to j = (Iter_max-1) do
7. For i=1 : Z do

ISSN: 0975-8585

8. Generate an ant (based on equations (9-12))
9. Xj+1(i)
10. End for
11. Update X (best)
12. Update the matrix F (based on equation (7)) and matrix PC (based on equation (6)).
13. Verify the stopping criteria
14. End for

Flow Diagram 1:



THYROID DATASET

ARFF MULTIVARIATE DATASET

ATTRIBUTE: **CATEGORICAL**
Discordant, negative | classes

age:	continuous.
sex:	M, F.
on thyroxine:	f, t.
query on thyroxine:	f, t.
on antithyroid medication:	f, t.
sick:	f, t.
pregnant:	f, t.
thyroid surgery:	f, t.
I131 treatment:	f, t.
query hypothyroid:	f, t.
query hyperthyroid:	f, t.
lithium:	f, t.
goitre:	f, t.
tumor:	f, t.
hypopituitary:	f, t.
psych:	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.

ISSN: 0975-8585

TBG measured:	f, t.
TBG:	continuous.

EXPERIMENTAL RESULTS

Thyroid dataset for this work was collected from UCI machine learning repository. The following results were obtained using various decision tree techniques.

Output Design:

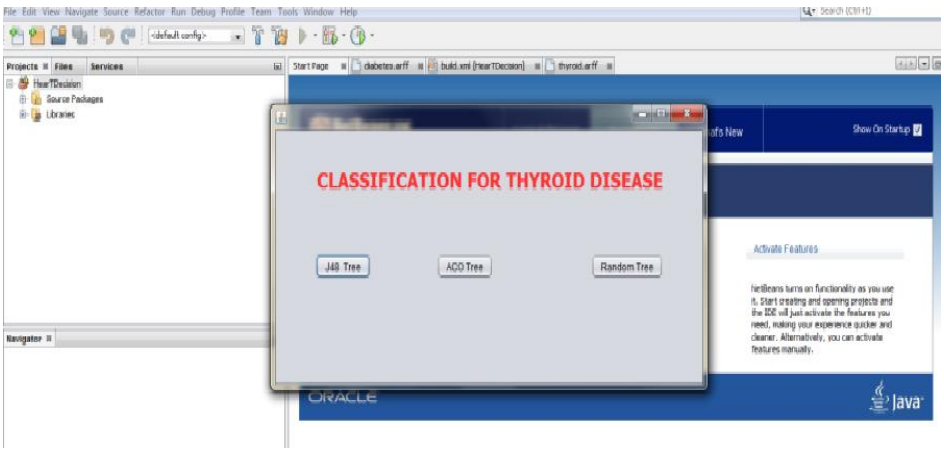


Fig 1: Select the Option for constructing decision trees

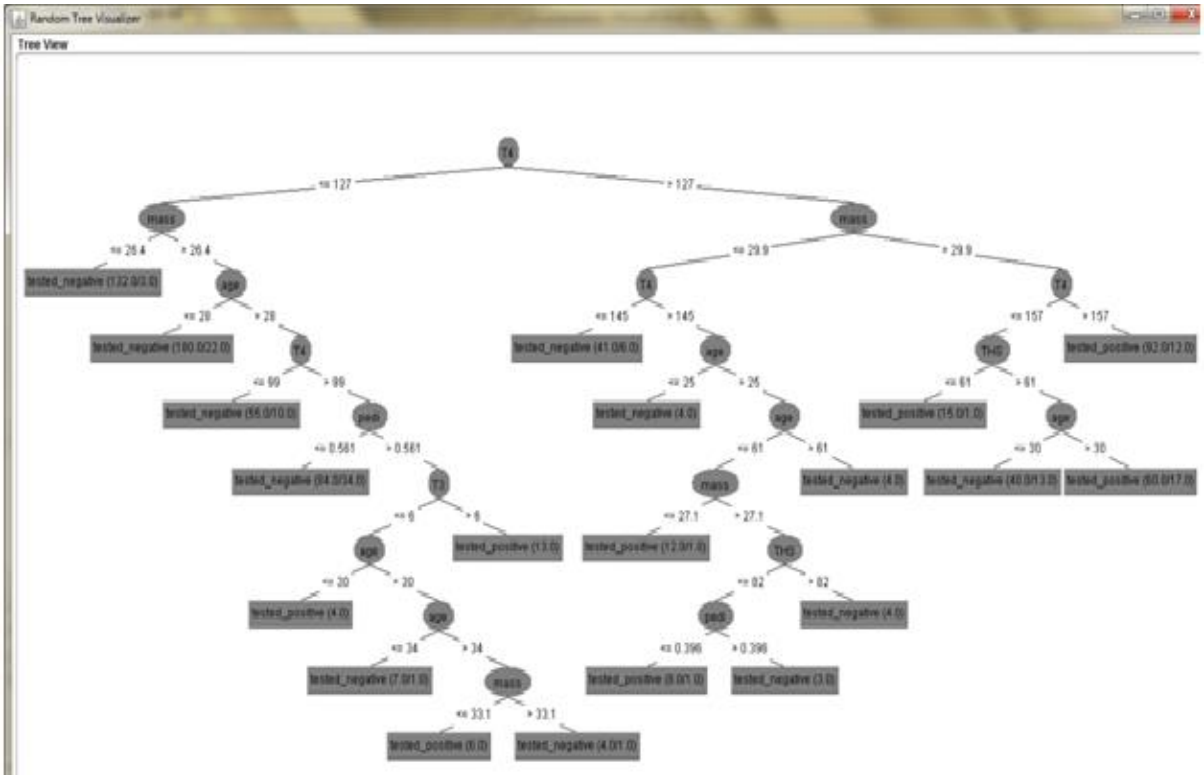


Fig 2: J48

ISSN: 0975-8585

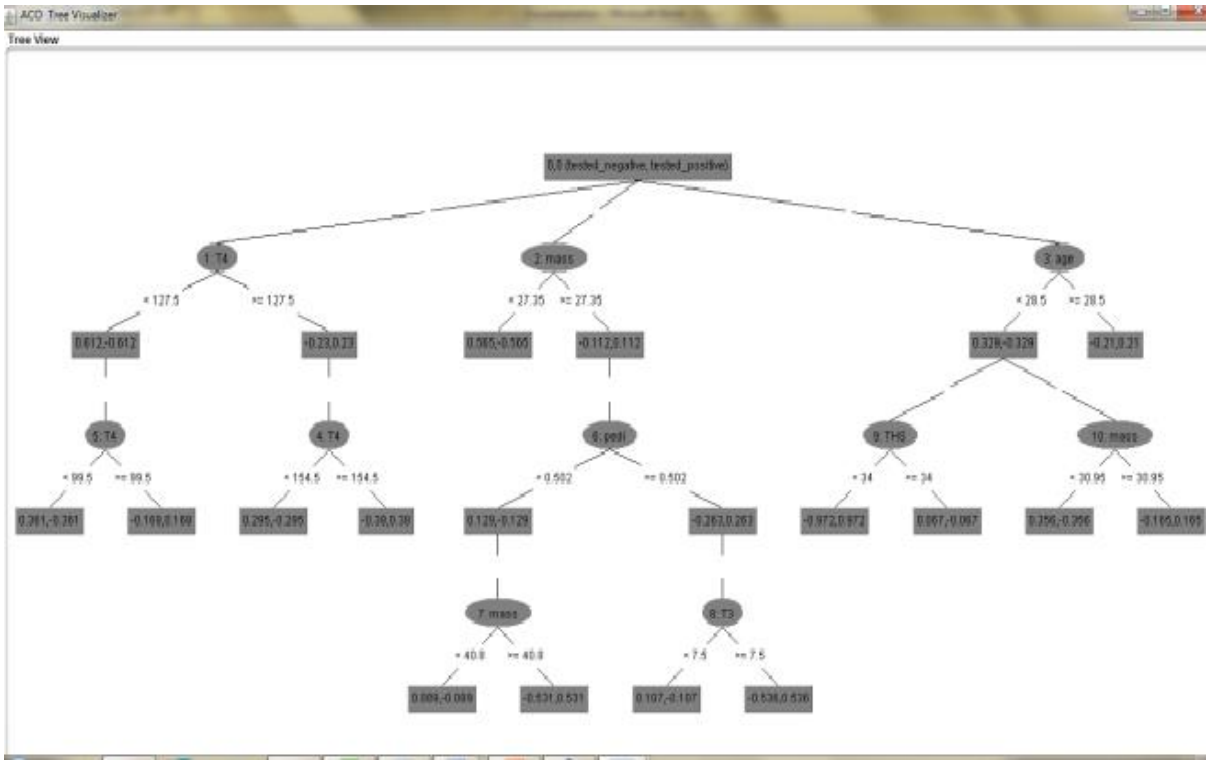


Fig 3: ACO-MST

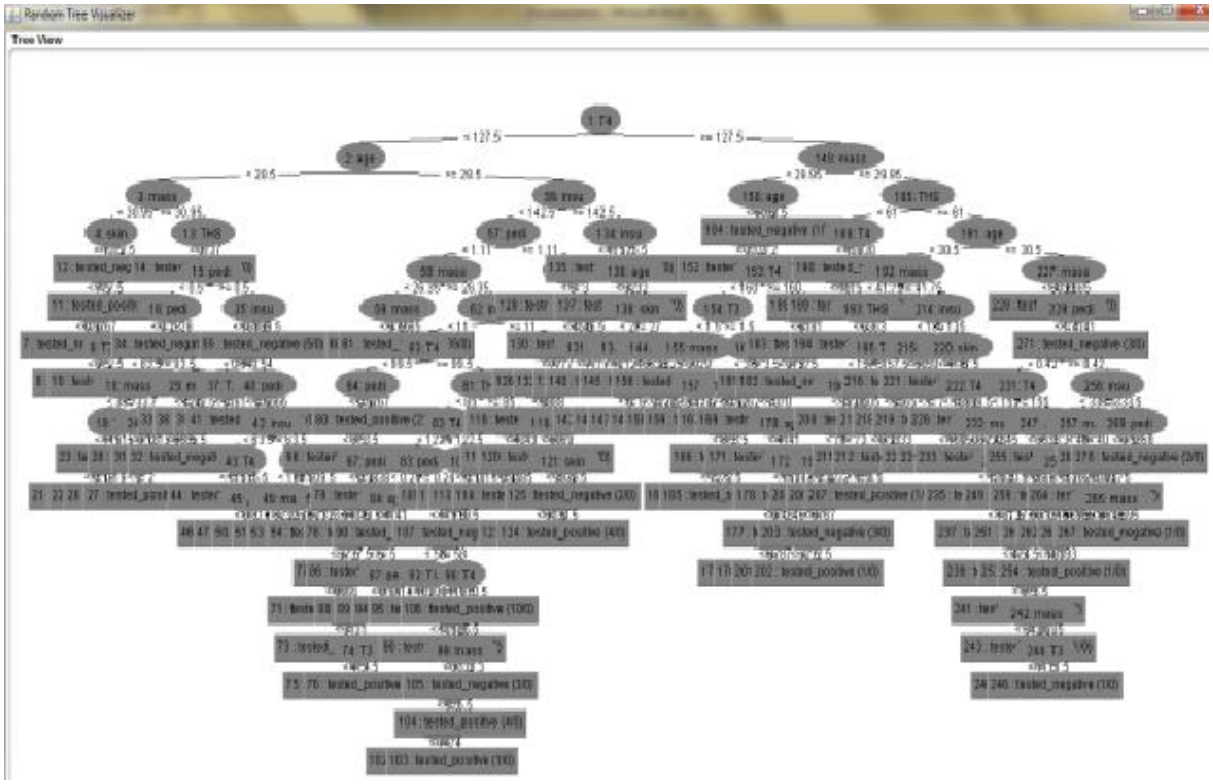


Fig 4: REP Tree

The result shown that the decision tree model applied to any data test is of course the norm for classifying it. The model uses ACO optimization and decision tree classification for classifying thyroid datasets accurately. The main task is to handle the uncertainty of the data in order to classify or cluster it. This is used

to test the effectiveness of the data and its impact in the dimensionality purpose. The data obtained is uncertain in that uncertainty occurs in the data because of the imprecise measurement of the results, like all scientific results. Tree partitions classify the dataset and select the representative feature and remove the redundant feature effectively.
The Accuracy (AC) of the data points pi is calculated using equation 1.
Accuracy (AC) is exact predictions of the entire number ratio. It is determined using Equation 1

$$AC = \sum_{i=1}^{N-1} pi * (p1+p2\ldots) \text{--------equ(1)}$$

Table 1: Classification of Thyroid Disease Accuracy Details

Techniques	Accuracy
J48	94.4336
REP TREE	96.3430
ACO-MST	97.1260

While comparing the accuracy of J48,REP-Tree, ACO-MST has more accuracy.

CONCLUSION

The proposed clustering is very fast and accurate. The redundant feature removes from relevant ones via choosing representatives from different feature clusters. The main task is to handle the uncertainty of the data in order to classify or cluster it. Tree partitions classify the dataset and select the representative feature. This method has the capacity to handle high dimensional data's. It does not buckle under dimensionality curse. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. The problem can be further improved by including genetic algorithms. This will lead to hybrid solutions with more accuracy. Further the hybrid solutions may trigger redundant features even earlier. Dimensionality Curse problem Computational efficiency is high. Its can work equally with small training sets and huge high dimensional data sets. Various decision tree attribute selection had been analysed and compared. This helps to diagnosis the thyroid diseases through the extracted rules.

REFERENCES

[1] Dr.Sahni BS, Thyroid Disorders [online]. Available : http://www.homoeopathyclinic.com /articles/diseases/tyro id.pdf
[2] thyroid :www.wikipedia.org/thyroid
[3] https://en.wikipedia.org/wiki/Decision_tree
[4] Y. Su, T.M. Murali, V. Pavlovic, M. Schaffer, S. Kasif, RankGene: identification of diagnostic genes based on expression data, Bioinformatics 19 (12) (2003) 1578–1579.
[5] A.C. Tan, D. Gilbert, Ensemble machine learning on gene expression data for Thyroid classification, Appl. Bioinform. 2 (3) (2003) S75–S83.
[6] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San-Mateo, CA, 2005.
[7] 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.
[8] S.J. Xiao, C. Zhang, Q. Zou, Z.L. Ji, TiSGeD: a database for tissue-specific genes, Bioinformatics 26(9) (2010) 1273–1275.
[9] K.saravana kumar,DR.R. Manichachezian Analysis on suspicious Thyroid recognition using association rule mining, JGRCS, VOL3, NO.9, 2012.
[10] Gayana .H .B, Nanda. s :Classfication of Thyroid nodules in ultrasound images using KNN of decision rule, IRJET, VOL 2 issue 05, Aug 2015