ISSN:2249-5789
Anurag Upadhayay et al, International Journal of Computer Science & Communication Networks,Vol 3(1), 64-68

Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set.

Anurag Upadhayay (M. Tech Scholar) IFTM University, INDIA anuragbareilly@yahoo.co.in
Suneet Shukla IFTM University, INDIA sshukl1@gmail.com
Sudsanshu Kumar IFTM University, INDIA sudhanshuindian2005@gmail.com

ABSTRACT

Health information system or medical informatics is the intersection of information science, computer science, and health care. It deals with the resources, devices, and methods required optimizing the acquisition, storage, retrieval, and use of information in health and biomedicine. This paper presents a snapshot of various forces driving the e-health applications; challenges for their widespread adoption and attempts to provide a conceptual framework for successful deliverance of e-health services. In healthcare environment, Cancer is a particularly opportune disease for data mining technology for a number of reasons. First, because the mountain of data is there. Second, Cancer is a highly growing disease that costs a great deal of money, and so has attracted managers and payers .I have implemented two algorithms of Decision Tree technique, *C4.5* and C5.0 technique. ending quest for saving money and cost efficiency.

Tools & platform

C4.5 approach is implemented in java platform using Eclipse and XP operating system

Keywords

C 4.5 , C 5.0 , ID3 , KDD, DM, TSH,TRH, T3, T4

INTODUCTION

Health information system or medical informatics is the intersection of information science, computer science, and health care .E-health applications can be broadly grouped under the following:

(a)Consumer health (b)Clinical care (c)Financial and administrative transactions (d) Public health (e) Professional education (f) Biomedical research In healthcare environment, Cancer is a particularly opportune disease for data mining technology for a number of reasons. First, because the mountain of data is there. Second, Cancer is a highly growing disease that costs a great deal of money, and so has attracted managers and payers in the never ending quest for saving money and cost efficiency.

Decision rule mining techniques are used to identify relationships among different Attributes of dataset in the form of decision Rules. Decision rules are more appropriate when we are searching Rules for Highly decision making.

This has motivated me to make use of Decision Tree, a mining technique for the Thyroid cancer Database available at UCI Machine Learning Repository.

WHAT IS DATA MINING

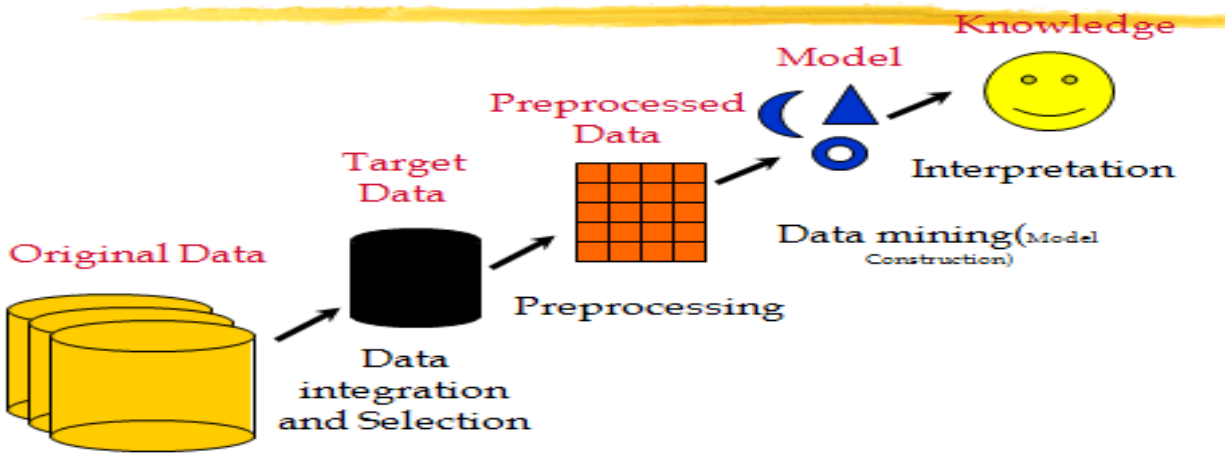
Data Mining is the process of semi-automatically analyzing large databases to find patterns.

The KDD vs. DM(DATA MINING)

KDD: is the process of finding useful information & pattern in data.

DM: is the use of algorithm to extract the information and patterns derived by the KDD process.

KDD (Knowledge discovery from Database) process



DATASET DESCRIPTION FOR THE PURPOSE OF DATA MINING

The Thyroid disease databases were obtained from Garavan Institute and J. Ross Quinlan, New South Wales Institute, Syndney, Australia. It was received by the UCI Machine learning Repository[5].The database has 2800 patients' records each with 29 attributes. I have considered 400 patients records for the research.

Attributes Name: (1) Age: continuous.(2) Sex: M, F. (3) On thyroxine: f, t. (4) query on thyroxine: f, t. (5) on antithyroid medication: f, t. (6) sick: f, t.(7) pregnant: f, t. (8) thyroid surgery: f, t.(9) I131treatment: f, t.

(10) query hypothyroid: f, t. (11) query hyperthyroid: f, t.(12) lithium: f, t.(13) goitre: f, t.(14) tumor: f, t.

(15) hypopituitary: 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, other.

Overview of Thyroid Disease:

In Thyroid disease, there are lumps which commonly arise within an otherwise normal thyroid gland also known as Thyroid Nodules which represent a common problem brought to medical attention. Four to Seven percent of the United States adult population (10–18 million people) has a palpable thyroid nodule.The thyroid gland is one of the largest endocrine glands. The thyroid gland is found in the neck, below the thyroid cartilage. The thyroid gland controls how quickly the body uses energy, makes proteins, and controls how sensitive the body is to other hormones. It participates in these processes by producing thyroid hormones, the principal ones being triiodothyronine (T3) and thyroxine which can sometimes be referred to as tetraiodothyronine (T4).Hormonal output from the thyroid is regulated by thyroid-stimulating hormone (TSH) produced by the anterior pituitary, which itself is regulated by thyrotrophic-releasing hormone (TRH) produced by the hypothalamus. Below is the figure of thyroid [7]

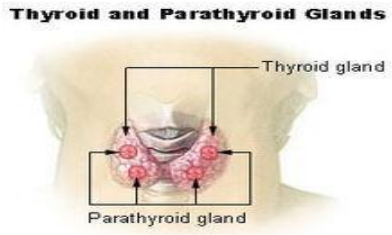
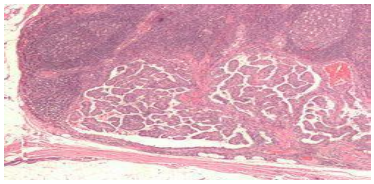


Figure: Thyroids and Parathyroid Gland

Thyroid cancer is a thyroid neoplasm that is (benign tumor in fig). It can be treated with radioactive iodine or surgical resection of the thyroid gland. Chemotherapy or radiotherapy 11
may also be used.Most often the first symptom of thyroid cancer is a nodule in the thyroid region of the neck.However, many adults have small nodules in their thyroids, but typically fewer than 5% of these nodules are found to be malignant [8].



Micrograph of A Lymph Node With Papillary Thyroid Carcinoma

C4.5 & C 5.0 INTRODUCTION

C 4.5: This algorithm is a successor to ID3 (Iterative Dichotomies 3)developed by Quinlan Ross. It is also based on Hunt's algorithm. C4 .5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values.

In pseudo code the algorithm is:

1. Check for base cases
2. For each attribute *a* (. Find the normalized information gain from splitting on *a)*
3. Let *a_best* be the attribute with the highest normalized information gain

4. Create a decision *node* that splits on *a_best*
5. Recur on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node*

C 5.0 algorithm

C 5.0 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = (s1. S2, …s n) of already classified samples. Each sample s_i consists of a p-dimensional vector (x1, x2,….xn) where the xn represent attributes or features of the sample, as well as the class in which si .

Entropy calculation

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where *pi* is the proportion of S belonging to class *i*. Note the logarithm is still base 2 because entropy is a measure of the expected encoding length measured in bits. The maximum possible entropy is log2c.

Information Gain calculation

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where *Values(A)* is the set of all possible values for attribute *A*, and *Sv* is the subset of *S* for which attribute *A* has value *v* (i.e., *Sv* = {s □ S |*A(s)* = *v*}).

Measurement factor for C 4.5 & C 5.0

Accuracy: The C5.0 rulesets have noticeably lowers error rates on unseen cases for the sleep and forest datasets. The C4.5 and C5.0 rulesets have the same predictive accuracy for the income dataset, but the C5.0 rule set is smaller. **Speed**: The times

are almost not comparable. For instance, C4.5 required nearly 15 hours finding the rule set for forest, but C5.0 completed the task in 2.5 minutes.

Memory: C5.0 commonly uses an order of magnitude less memory than C4.5 during rule set construction. For the forest dataset, C4.5 needs more than 3GB (the job would not complete on earlier 32-bit systems), but C5.0 requires less than 200MB.

Smaller decision trees - C5.0 gets similar results to C4.5 with considerably smaller decision trees.

Support for Boosting - Boosting improves the trees and gives them more accuracy.

Weighting - C5.0 allows you to weight different attributes and misclassification types.

Winnowing - C5.0 automatically winnows the data to help reduce noise.

Rule set Generated By C5.0 algorithm:

Rule 1: TSH > 6 TT4 <= 41 -> class primary hypothyroid [0.900]

Rule 2: thyroid surgery = t TSH > 6 -> class primary hypothyroid [0.500]

Rule 3: on thyroxine = f thyroid surgery = f TSH > 6 51

TT4 > 41 -> class compensated hypothyroid [0.885]

Rule 4: TSH <= 6 -> class negative [0.997]

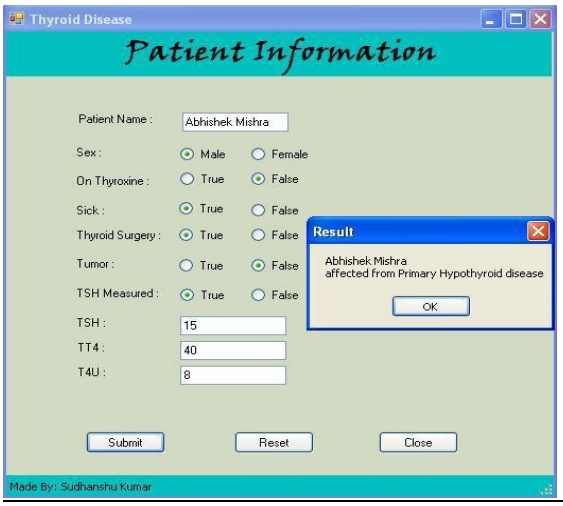
Rule 5: on thyroxine = t -> class negative [0.978]

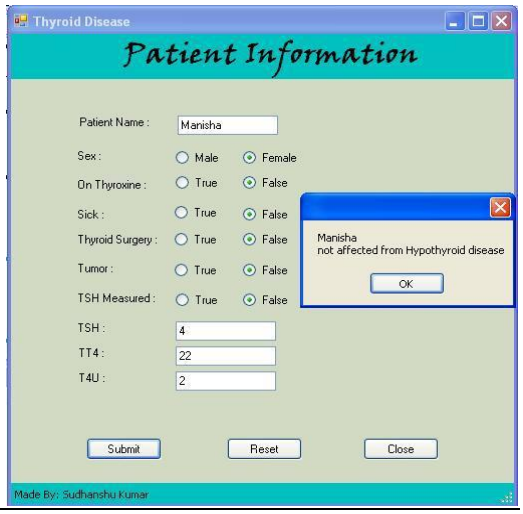
Rule 6: TSH measured = f -> class negative [0.977]

Comparison Between two Algorithms on 400 cases

Algorithm	**Tree Size Before Pruning**	**Train Error**	**Tree Size After Pruning**	**Train Error**	**Rules**	**Rule Confidence**
C 4.5	25	5	26	5	> 6	More then 90%
C 5.0	6	3	6	6	6	More then 95%

User Interface of Prediction Model





CONCLUSION

Our goal was to observe the impact of data mining technique on Thyroid Cancer. In this study, I have implemented two algorithms of Decision Tree technique, viz *C4.5* and C5.0 technique. And the following observation is noted down.

1.Tree Size of C4.5 was very large in compare to C5.0.

2. Rules Generated by both algorithms was different.

3.After Pruning C5.0 Tree generated more accurate rule set.

4. Running Time of C5.0 was Small as compare to C4.5.

5. Train error in case of c5.0 was small in compare to the C4.5

6. Rule set Generated by the C5.0 algorithm is 6 and the confidence level of rules was more than 95%.

REFERENCES

1. G.K. Gupta . Introduction to Data Mining with Case Studies,2006,Prentice Hall India.

2. Marisa S. Viveros, John P Nearhos, Michael J. Rothman," Applying Data Mining Techniques to a Health Insurance Information System",2003.

3. SellappanPalaniappan, RafiahAwang," Intelligent Heart Disease prediction System Using Data Mining Techniques",2008.

4. Klaus Vilstrup Pedersen1, Aarhus School of Business, Denmark University of Technology, Sydney," Using Data Mining Techniques for Optimizing Medical Treatment Processes",2003.

5. UCI Repository of machine learning databases.

6. Development of a clinical decision model for thyroid nodules, Alexander Stojadinovic*1,3, George E Peoples2,3, Steven K Libutti4,Leonard R Henry3,5, John Eberhardt6, Robin S Howard7, David Gur8,9, Eric A Elster5 and Aviram Nissan3,10

7. http://en.wikipedia.org/wiki/Thyroid

8. http://en.wikipedia.org/wiki/Thyroid_cancer

9. FahadShahbaz Khan, Rao Muhammad Anwer, OlofTorgersson, and GöranFalkman," Data Mining in Oral Medicine Using Decision Trees", 2008.

10. Surjeet Kumar Yadav, Saurabh Paul "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification" WCSIT,ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

11. Nevena Stolba and A Min Tjoa, "The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making",2003