DIAGNOSIS OF LUNG CANCER USING CLASSIFICATION AND REGRESSION TREE (CART)

Me Me Khaing
Faculty of Computer Science
Myanmar Institute of Information Technology (MIIT)
memekhaing2012@gmail.com

Abstract*: Many users do decisions about several diseases, and they need methods to decide and know about diseases. Lung cancer that has spread beyond the original tumor is difficult to cure. Eventually, people with lung cancer do develop symptoms. The physical examination is a crucial part of the diagnostic process for any medical problem. This decision making process classify how the extent of the disease with level for lung cancer. The audit focuses on measuring the care given to lung cancer patients from diagnosis to the primary treatment and bringing about necessary improvements. This system uses classification method CART to diagnose lung cancer. This system generates rules on the lung cancer training dataset by using classification and regression tree (CART) and then rules is used to classify the unknown dataset. The rules extracted from CART are helpful to users in diagnosing lung cancer. CART is suited to the generation of clinical decision rules.*

1. INTRODUCTION

A classification model can be used to classify the class label of unknown records. A classification model can be treated that automatically assigns a class label when presented with the attribute set of an unknown record. This system is implemented based on classification system that diagnosis the lung disease. This system supports the knowledge acquisition system and knowledge based systems. This system helps the knowledge engineer to build a new simpler, quicker and more effective system, and reusing the already existing components in similar sy stems. This system is able to assist professionals by providing accurate and timely diagnosis information for decision making, and a user-friendly tool for analyzing and selecting the suggested controls for variations in medical environment [7].

A common goal of many clinical research studies are the development of a reliable clinical decision rules, which can be used to classify new patients into clinically-important categories. There has been increasing interest in the use of classification and regression tree (CART) analysis. CART is to be quite effective for creating clinical decision rules which perform as well or better than rules developed using more traditional methods. CART analysis is a form of binary recursive partitioning. K-fold method is used for splitting original dataset into training set and test set. In this system, classification method CART builds the classification tree by analyzing the training set and test set is used to test the unknown lung cancer dataset. This system classified the stage of lung cancer by using Classification and regression tree. Entropy is used to calculate the best gain from lung cancer attributes. This system uses the lung cancer dataset in the UCI Machine Learning Data Repository. It includes 160 data records and 3 class label [12].

2. RELATED WORKS

In paper [3] implemented Text Miner and Cluster analysis to identify the claim data for Lung Cancer and to determine the category of diagnosis, treatment procedures and medication treatments for those patients. Moreover, the claims data were used to define severity level and treatment categories. Compared with using diagnosis codes directly, the combination of text mining and cluster analysis is more efficient and captures more useful information for further analysis.

In paper [6] developed “Comparison with linear discriminated analysis and with classification and regression trees for diagnosis of liver disease”. Entropy is used to select gain of liver disease attributes and jackknife validation is used.

In paper [8] developed "Classification Trees and Predicting Breast Cancer" using CART. This study is used classification and regression trees (CART), it develop the ability to fit a tree to data. This study formulated a CART model through pruning and impurity, and evaluate its predictive ability,apply this methodology to data obtained from Machine Learning Repository

3. THEORY BACKGROUND AND HISTORY OF APPLICATION

The CART decision tree is a binary recursive partitioning procedure capable of processing continuous and nominal attributes as targets and predictors. Data are handled in their raw form; no binning is required or recommended. Beginning in the root node, the data are split into two children, and each of the children is in turn split into grandchildren. Trees are grown to a maximal size without the use of a stopping rule; essentially the tree-growing process stops when no further splits are possible due to lack of data. The maximal-sized tree is then pruned back to the root (essentially split by split) via the novel method of cost-complexity pruning. The next split to be pruned is the one contributing least to the overall performance the tree on training data (and more than one split may be removed at a time).The lungs are vital organs. Working with the heart and circulatory system, they provide life-sustaining oxygen and rid the body of carbon dioxide. Normal lungs have a great reserve capacity to meet the body's need for oxygen across a wide variety of circumstances. The same is true of the heart and circulatory system. This reserve capacity permits cancerous lung tumours to grow for years without compromising lung function. Furthermore, the lungs do not have many nerves to transmit pain messages. Therefore, a cancerous lung tumour can grow for many years without causing any symptoms. Unfortunately, this means that most people are not diagnosed with lung cancer until late in the disease process. Even more unfortunate is the fact that this long period of silent growth gives the cancer time to spread before it is diagnosed.

Eventually, people with lung cancer do develop symptoms. Approximately 95% of people diagnosed with lung cancer have symptoms related to the disease. However, they occur late in the cancerous process. The long silent growth period of lung cancer has led to great interest in lung cancer screening, especially in recent years [12].

Common lung cancer symptoms include [5]:

- Constant chest pain,
- Chronic cough that worsens over time,
- Coughing up blood (hemoptysis),
- Dyspnea (difficulty breathing),
- Fatigue,
- Lung infection (pneumonia, bronchitis),
- Shortness of breath,
- Swollen lymph nodes,
- Loss of appetite and weight loss, and Wheezing.

4. CLASSIFICATION AND REGRESSION TREE

The CART mechanism is intended to produce not one tree, but a sequence of nested pruned trees, each of which is a candidate to be the optimal tree. The "right sized" or "honest" tree is identified by evaluating the predictive performance of every tree in the pruning sequence on independent test data. Unlike C4.5, CART does not use an internal (training-data-based) performance measure for tree selection. Instead, tree performance is always measured on independent test data (or via cross-validation) and tree selection proceeds only after test-data-based evaluation. If testing or cross-validation has not been performed, CART remains agnostic regarding which tree in the sequence is best. This is in sharp contrast to methods such as C4.5 or classical statistics that generate preferred models on the basis of training data measures [2].

The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items. "Best" is defined by how well the variable splits the set into homogeneous subsets that have the same value of the target variable. Different algorithms use different formulae for measuring "best". This section presents a few of the most common formulae. These formulae are applied to each candidate subset, and the resulting values are combined (e.g., averaged) to provide a measure of the quality of the split[1].

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness to "impurity" in these Partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.. The expected information needed to classify a given sample is given by

$$I(S_1, S_2, ..., S_3) = -\sum_{i=1}^{m} P_i \log_2(P_i) \qquad (4.1)$$

where p_i is the probability that an arbitrary sample belongs to class and is estimated by $\frac{S_i}{S}$. Note that a log function to the base 2 is used since the information is encoded in bits.

$$I(S_{1j}, ..., S_{mj}) = -\sum_{i=1}^{m} P_{ij} \log_2(P_{ij}) \qquad (4.2)$$

where $Pij = \frac{sij}{|sj|}$ and is the probability that a sample in Sj belongs to class .

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + ... + S_{mj}}{S} I(S_{1j}, ..., S_{mj}) \qquad (4.3)$$

The term $\frac{s_1 j + ... + s_m j}{s}$ acts as the weight of the jth subset and is the number of samples in the subset (i.e., having value aj of A) divided by the total number of samples in S.

$$Gain(A) = I(S_1, S_2, ...S_m) - E(A) \qquad (4.4)$$

In the other words, Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly [11].

4.1 Extracting Classification Rules from Decision Trees

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). The IF-THAN rules may be easier for humans to understand[4].

5. SYSTEM DESIGN AND IMPLEMENTATION

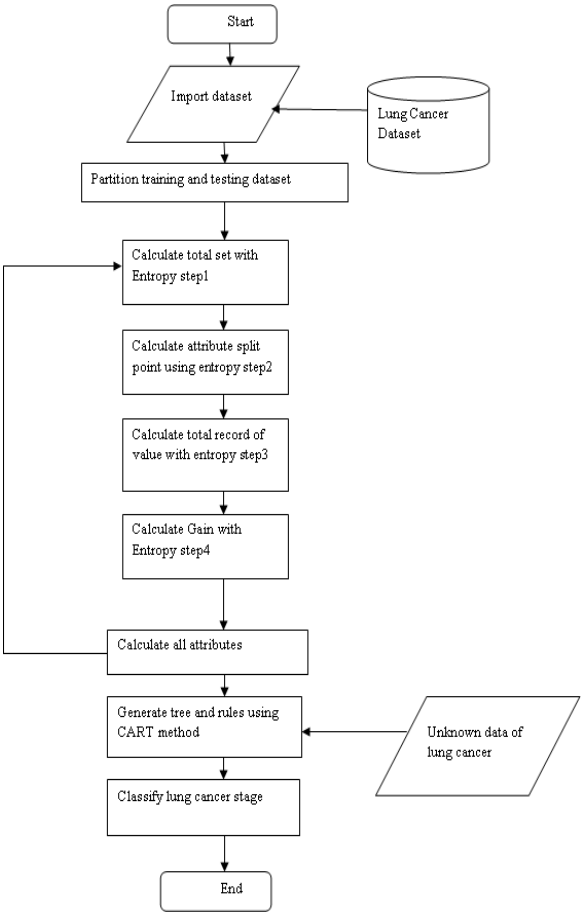

Figure 1. System Flow Diagram

In Figure 1, first user import lung cancer dataset from lung cancer database. User can partition lung cancer dataset as training set and test set. The system constructs with the entropy steps to get classification tree (CART). The system generates the evaluated on a database of lung cancer, whose experimental results show that the system correctly diagnostic rules. And then the system accepts unknown data lung cancer , and displays diagnosis result of lung cancer stage.

5.1. About Lung cancer dataset and Case study

This system used lung cancer dataset from UCI machine learning website [2].This dataset include 160 records, 10 attributes and 3 stages. Attributes names are Fever, Hemoptysis, Chronic Cough, Dyspnea, Breathlessness, Stridor, Hoarseness, Tiredness, Lung Infection and Swollen Lymph

Nodes. By using this data set, each lung cancer attribute is analyzed by entropy steps to get the best gain. Table 1 is organized with some patients and ten attributes.

Table 1. Lung cancer dataset for some patients

ID	Fever	Hemoptysis	Chronic Cough	Dyspnea	Breathlessness	Stridor	Hoarseness	Tiredness	Lung Infection	Swollen Lymph Nodes	Cancer Stage
1	Yes	Present	Yes	Normal	Yes	Normal	Severe	Severe	Mild	Moderate	2
2	Yes	Increased	Severe	Variable	No	Normal	Severe	Mild	Mild	Mild	3
3	Yes	Increased	Moderate	Variable	Yes	Normal	Severe	Severe	Moderate	Moderate	1
4	Yes	Present	Severe	Increased	Yes	Increased	Moderate	Severe	Mild	Mild	3
5	No	Increased	Yes	Normal	Yes	Variable	Moderate	Moderate	Mild	Mild	2
6	Yes	Present	Severe	Increased	No	Variable	Moderate	Moderate	Mild	Moderate	3
7	Yes	Increased	Severe	Increased	No	Normal	Severe	Severe	Mild	Moderate	1
8	Yes	Increased	Moderate	Increased	Yes	Increased	Moderate	Moderate	Mild	Mild	2
9	Yes	Present	Moderate	Increased	No	Increased	Mild	Moderate	Mild	Mild	3
10	Yes	Increased	Moderate	Variable	No	Normal	Severe	Severe	Mild	Moderate	1
11	Yes	Increased	Moderate	Increased	No	Variable	Mild	Severe	Mild	Mild	2

5.2. Result Table using Entropy

Table2 described the best gain of lung cancer attribute from lung cancer dataset for some patients using entropy steps using equation 1 to 4. Table2 showed the best gain Stridor is Normal condition with highest gain from all of ten attributes.

Table 2. Gain values calculate with Entropy

Attributes	Conditions	Gain value (Entropy)
Fever		0.144867603
Hemoptysis		0.35563985
Chronic Cough	Yes	0.320402072
	Severe	**0.40020576**
	Moderate	0.08493930248
Dyspnea	**Normal**	**0.3204020719**
	Variable	0.255592462
	Increased	0.084933248
Breathlessness		0.15356439
Stridor	**Normal**	**0.4040097576**
	Increased	0.1867043463
	Variable	0.1867043463
Hoarseness	**Severe**	**0.4040097576**
	Moderate	0.2183875778
	Mild	0.09401798138
Tiredness	**Severe**	**0.3353836209**
	Mild	0.1444867604
	Moderate	0.2183875778
Lung Infection		0.1890526685
Swollen Lymph Nodes		0.404009757

5.3. Extracts Lung Cancer Dataset

Table 3 showed Stridor is normal condition and this condition; attribute is again calculated for maximum gain using entropy equation 1 to 4.

Table 3. Extracts dataset after calculating the best gain attribute

ID	Fever	Hemoptysis	Chronic Cough	Dyspnea	Breathlessness	Stridor	Hoarseness	Tiredness	Lung Infection	Swollen Lymph Nodes	Cancer Stage
1	Yes	Present	Yes	Normal	Yes	Normal	Severe	Severe	Mild	Moderate	2
2	Yes	Increased	Severe	Variable	No	Normal	Severe	Mild	Mild	Mild	3
3	Yes	Increased	Moderate	Variable	Yes	Normal	Severe	Severe	Moderate	Moderate	1
7	Yes	Increased	Severe	Increased	No	Normal	Severe	Severe	Mild	Moderate	1
10	Yes	Increased	Moderate	Variable	No	Normal	Severe	Severe	Mild	Moderate	1

Table 4. Gain values calculate with entropy

Attributes	Conditions	Gain value (Entropy)
Fever		0
Hemoptysis	**Present**	**0.7219280944**
Chronic Cough	**Yes**	**0.7219280944**
	Severe	0.4199730935
	Moderate	0.0199730934
Dyspnea	**Normal**	**0.7219280944**
	Variable	0.4199730935
	Increased	0.170950594
Breathlessness		0.419973093
Stridor		
Hoarseness		0
Tiredness		0.7219280944
Lung Infection		0.70590594
Swollen Lymph Nodes		0.7219280944

Table 4 described the best gain of Hemoptysis condition is present. So, Stridor is normal and Hemoptysis is present, lung cancer stage is 2 according to ID 1 from Table3.

6. CLASSIFICATION RULES

IF Stridor = "Variable" AND Hemoptysis = "NOT Absent" AND Hoarseness = "Mild" THEN "Stage2"

2. ELSE IF Stridor = "Variable" AND Hemoptysis = "NOT Absent" AND Hoarseness = "NOT Mild" AND Chronic_Cough = "Yes" THEN "Stage2"

3. ELSE IF Stridor = "Variable" AND Hemoptysis = "NOT Absent" AND Hoarseness = "NOT Mild" AND Chronic_Cough = "NOT Yes" AND Fever = "Yes" AND Dyspnea = "NOT Normal" AND Tiredness = "Mild" AND Breathlessness = "No" THEN "Stage2"

4. ELSE IF Stridor = "Variable" AND Hemoptysis = "NOT Absent" AND Hoarseness = "NOT Mild" AND Chronic_Cough = "NOT Yes" AND Fever = "Yes" AND Dyspnea = "NOT Normal" AND Tiredness = "Mild" AND Breathlessness = "NOT No" THEN "Stage3"

5. ELSE IF Stridor = "Variable" AND Hemoptysis = "NOT Absent" AND Hoarseness = "NOT Mild" AND Chronic_Cough = "NOT Yes" AND Fever = "Yes" AND Dyspnea = "NOT Normal" AND Tiredness = "NOT Mild" AND Breathlessness = "No" AND Lung_Infection = "Mild" AND Swollen_Lymph_Nodes = "NOT Mild" THEN "Stage2"

6. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "Mild" THEN "Stage3"

7. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "NOT Mild" AND Hemoptysis = "Absent" THEN "Stage2"

8. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "NOT Mild" AND Hemoptysis = "NOT Absent" AND Swollen_Lymph_Nodes = "Mild" AND Lung_Infection = "Mild" AND Chronic_Cough = "Yes" THEN "Stage1"

9. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "NOT Mild" AND Hemoptysis = "NOT Absent" AND Swollen_Lymph_Nodes = "Mild" AND Lung_Infection = "Mild" AND Chronic_Cough = "NOT Yes" AND Hoarseness = "Mild" AND Dyspnea = "NOT Normal" AND Breathlessness = "No" THEN "Stage2"

10. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "NOT Mild" AND Hemoptysis = "NOT Absent" AND Swollen_Lymph_Nodes = "Mild" AND Lung_Infection = "Mild" AND Chronic_Cough = "NOT Yes" AND Hoarseness = "NOT Mild" AND Dyspnea = "Normal" THEN "Stage2"

11. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "NOT Mild" AND Hemoptysis = "NOT Absent" AND Swollen_Lymph_Nodes = "Mild" AND Lung_Infection = "Mild" AND Chronic_Cough = "NOT Yes" AND Hoarseness = "NOT Mild" AND Dyspnea = "NOT Normal" AND Breathlessness = "No" THEN "Stage3"

12. ELSE IF Stridor = "NOT Variable" AND Fever = "Yes" AND Tiredness = "NOT Mild" AND Hemoptysis = "NOT Absent" AND Swollen_Lymph_Nodes = "Mild" AND Lung_Infection = "Mild" AND Chronic_Cough = "NOT Yes" AND Hoarseness = "NOT Mild" AND Dyspnea = "NOT Normal" AND Breathlessness = "NOT No" THEN "Stage2"

In Figure 2 shows the classification result for lung cancer stage according to classification rules.

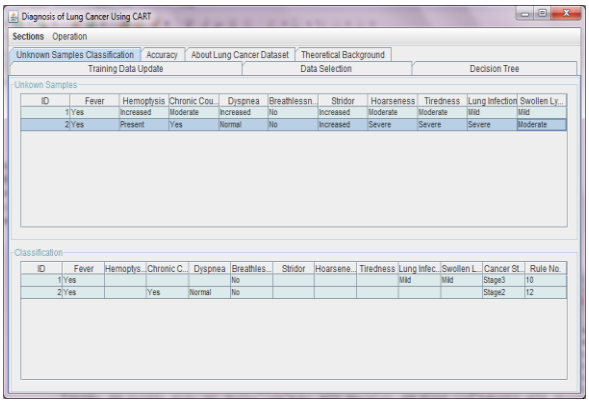

Figure 2.Unknown data classification

7. CONCLUSION

This system is interactive information systems that assist a decision marker in solving lung's medical problems. This system will take the truth decision on not only diagnosis but also the analyzing previous disease condition of lung cancer. The system takes patient history and disease symptoms. Disease symptoms are analyzed by using decision tree induction for diagnosis rule based structure and then provide the correct diagnosis. The system aids the doctors in decision making process during the patient's attendance, supplying the most correct diagnosis and treatment in relation to the types of lung symptoms. Thus, computer-based diagnosis systems will play an increasingly important role in health care. System can also be used when an expert is unavailable or can be served as a consult.

This system describes lung cancer classification system that uses a tree-based method Classification and Regression Tree (CART) method to classify the unknown samples for lung cancer symptoms. CART analysis is a powerful technique with significant potential and clinical utility. This system will present generating of classification tree and rules effectively in detailing with lung cancer data.

REFERENCES

[1] Alin Dobra, "Classification and Regression Tree Construction", Department of Computer Science, Cornell University, Ithaca , November 2002.

[2] Dataset, "UCI Machine Learning Repository".

[3] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann , Los Altos , California , 1993.

[4] J. R. Quinlan, "Simplifying Decision Tree", International Journal of Man-Machine Studied, 27:221-234, 1987.

[5] Keith Ballinger, "Lung Cancer for Patients", Scottish Intercollegiate Guidelines Network.

[6] Leon N Cooper, Institute for Medical Chemistry and Biochemistry, and Department of Internal Medicine, University of Innsbruck, Fritz Pregl Strasse 3, September 16, 1991

[7] Mark A. Hall, "Correlation-based Feature Selection for Machine Learning" , Department of Computer Science , Hamilton , New Zealand , March 2009.

[8] Megan Howard, Department of Mathematics, Computer Science and Statistics
Spring 2010.

[9] R. R. Quinlan, "Introduction of Decision Trees" , Machine Learning , 1:81-106 , 1986.

[10] Roger J. Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis" , Department of Emergency Medicine , Harbor , UCLA Medical Center , Torrance , California.

[11] T. M. Mitchell, "Machine Learning. McGraw-Hill", New York, 1997.

[12] Timothy Finin , Charles Nicholas , "Lung Cancer Diagnosis and Staging" , University of Maryland Baltimore County.