DIAGNOSE ABNORMAL NASAL BASED ON DATA MINING TECHNIQUES

WASIN SRISAWAT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2013

Thesis
entitled

DIAGNOSE ABNORMAL NASAL BASED ON DATA MINING TECHNIQUES

…….....
Mr. Wasin Srisawat
Candidate

.....
Asst. Prof Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Major advisor

.....
Asst. Prof Adisorn Leelasantitham,
Ph.D. (Electrical and Computer Engineering)
Co-advisor

.....
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Co-advisor

…….....
Prof. Banchong Mahaisavariya,
M.D., Dip. (Thai Board of Orthopedics)
Dean
Faculty of Graduate Studies,
Mahidol University

…….
Asst. Prof Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Program Director
Master of Science Program in Technology of Information System Management
Faculty of Engineering
Mahidol University

Thesis
entitled

DIAGNOSE ABNORMAL NASAL BASED ON DATA MINING TECHNIQUES

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
December 26, 2013

…….....
Mr. Wasin Srisawat
Candidate

.....
Asst.Prof.Bunlur Emaruechi,
Ph.D. (Environment Systems Engineering)
Chair

…….....
Asst. Prof Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Member

…….
Asst. Prof Adisorn Leelasantitham,
Ph.D. (Electrical and Computer Engineering)
Member

…….....
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Member

…….
Asst. Prof Werapon Chiracharit,
Ph.D. (Electrical and Computer Engineering)
Member

…….....
Prof. Banchong Mahaisavariya,
M.D., Dip. (Thai Board of Orthopedics)
Dean
Faculty of Graduate Studies,
Mahidol University

…….
Lect. Worawit Israngkul,
M.S. (Technical Management))
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

The success of this thesis can be succeeded by the attentive support from my major advisor Asst.Prof. Supaporn Kiattisin for her valuable advice in choosing thesis topic which were so much beneficial for successful of this thesis.

I am grateful to Asst.Prof. Visan mahasithiwat MD. from Otolaryngplogy head and neck surgery Department, HRH Princess Maha Chakri Sirindhorn medical center, Srinakhanarinwirot University for his assistance on capturing nasal cross sectional area data. We are grateful to Co-advisor who provided helpful comments on a previous draft of this thesis.

Finally, I am very grateful to my family members for attending to my study, care and love and great encouragement. This successfulness I dedicate to my parents and all the teachers who's the main source of inspiration to push me to future success. And I would like to thank all of those who I have not listed above.

Wasin Srisawat

DIAGNOSE ABNORMAL NASAL BASED ON DATA MINING TECHNIQUES

WASIN SRISAWAT 5438245 EGTI/M

M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THESIS ADVISORY COMMITTEE: SUPAPORN KIATTISIN, Ph.D., ADISORN LEELASANTITHAM, Ph.D., WARANYU WONGSEREE, Ph.D.

ABSTRACT

This thesis proposes methods to classify the pattern of an unusual nasal cavity using Ripper Rule, C4.5 decision tree and K-Nearest neighbor. It aims to help physicians classify an abnormal nasal cavity from an acoustic rhinometry signal. The experiments showed that the algorithm with the most effective classification was the C4.5 decision tree, which has an ROC of 0.99 (sensitivity 0.99, specificity 0.99, and standard deviation 0.1). The results showed that abnormalities of the nasal cavity are about 0 – 4.24 cm and the nasal cross sectional area is less than 0.55 cm^2. Therefore, this study suggests that the C4.5 decision tree algorithm could be applied for screening abnormal nasal cavities. It can lead to an application or tool development in medical devices in the future.

KEY WORDS: ACOUSTIC RHINOMETRY / CLASSIFICATION / NASAL CROSS SECTIONAL AREA / C4.5 DECISION TREE / DATA MINING

36 pages

การคัดแยกความผิดปกติของโครงสร้างโพรงจมูกด้วยวิธีการทำเหมืองข้อมูล
DIAGNOSE ABNORMAL NASAL BASED ON DATA MINING TECHNIQUES

วศิน ศรีสวัสดิ์ 5438245 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : สุภาภรณ์ เกียรติสิน, Ph.D., อดิศร ลีลาสันติธรรม, Ph.D., วรัญญู วงษ์เสรี, Ph.D.

บทคัดย่อ

การวิจัยนี้ได้นำเสนอวิธีการคัดแยกความผิดปกติของโครงสร้างของโพรงจมูกด้วยใช้อัลกอริทึม Ripper Rule, C4.5 decision tree, K-Nearest neighbor เพื่อวิเคราะห์พื้นที่หน้าตัดของโพรงจมูกจากเครื่อง RhinoScan ซึ่งจากการทดลองแสดงให้เห็นว่าอัลกอริทึมที่มีประสิทธิภาพในการคัดแยกดีที่สุดคือ C4.5 decision tree โดยมีค่า ROC 0.99 (sensitivity 0.99, specificity 0.99 และ standard deviation 0.1).จะเห็นว่าค่าของระยะที่ผิดปกติของโพรงจมูกจะอยู่ที่ช่วงระยะประมาณ 0.3 – 5 ซม. และมีพื้นที่หน้าตัดของโพรงจมูกน้อย 0.55 ตารางเซนติเมตร ดังนั้นจากการวิจัยนี้เราสามารถนำอัลกอริทึม C4.5 decision tree มาประยุกต์ใช้ในการพัฒนาโปรแกรมเพื่อคัดยากความผิดปกติของโครงสร้างโพรงจมูกบนอุปกรณ์การแพทย์ได้

36 หน้า

CONTENTS

CONTENTS(cont.)

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background and Problem Statement

Nasal and sinus disease are common diseases and public health problems in the country. These diseases, which cause blockages in the nasal cavity because of the nasal mucosa swelling by inflammation. Patients often have a stuffed nose and other symptoms such as running nose, the nose does not smell, pain, nasal congestion, sneezing, etc.

Stuffy nose are common symptoms that occur normally such as stuffy nose caused by running opposite sides naturally or stuffy nose caused by changing the posture. These are symptoms that causes patients to suffer and bother. In the U.S., there have been estimates that the cost to treat stuffy nose up to 5 billion U.S. dollars per year. Moreover, cost up to 60 million U.S. dollars per year in the surgical treatment of nasal congestion.

The secondary prevention is the principle of prevention and control disease. The objective are to diagnose or discover the disease as early as possible for reducing sick time, cost of treatment and also, make patients a better quality of life. However, the diagnosis often has several limitations such as there are not enough experts, take a long time, complicated, costly and sometimes harmful to the patient.

This thesis proposes a method of data mining for analyze, build and identify the abnormal pattern of nasal cavity which will as guidelines for develop software that is used to help Physician analyze Thai people's disease of nasal cavity.

1.2 Objectives

1) To analyze the patterns of nasal cavity's abnormalities.
2) To create the model for abnormalities classification of nasal cavity.

1.3 Scope of Work

1) The data comes from the patients (both male and female) with nasal cavity troubles, age between 18-66 years, from HRH Princess Maha ChaKri Sirindhorn Medical Center.

2) To study the result from medical measurement by using Acoustic Rhinometry.

1.4 Expected Result

1) To analyze the patterns of nasal cavity's abnormalities with high accuracy and performance.

2) The model supported for abnormalities classification of nasal cavity with efficiency and minimum time.

3) To publish the knowledge that obtained from analysis and creating a model to the researchers in related fields.

CHAPTER 2
LITERATURE REVIEW

Currently, Medical could diagnose the congested nose by objective measurement (measuring nasal patency) with two tools [1] are:

- Active Anterior Rhinomanometry (AAR) which is a measure of resistance's nasal cavity.
- Acoustic Rhinometry which is a measure of cross sectional area and nasal volume.

The nasal cavity with no choke should has low nasal resistance, proper nasal volume and minimal cross sectional area.

Rhinomanometry is considered as a standard diagnosis (Gold standard) [2] that assessment of nasal patency is widely used in clinical. But in actual practice, Rhinomanometry is a measure with high costs, expensive equipment, requiring staff with specialists and cooperation from patient particularly young patients. The simple way to evaluate nasal patency is using acoustic rhinometry to measure cross section area of nasal cavity.

Acoustic Rhinometry measures the structure inside nose that easily in order to track the treatment and evaluate blockages in the nasal cavity [3-8] by the reflection of sound waves [9-10] for cross-sectional area of nasal passages and distance from the nostril into the measured position. This tool cannot analyze primary abnormalities of the nasal cavity. Physician required information of cross sectional area and pattern of nasal cavity. Then data were compared with the pattern of the nasal cavity that has a statistical study in Chinese, Malay, Indian [12] and Iraq [13] for examination of position and areas where there is a blockage in nasal passages.

Present, medical have been used data mining techniques to assist in the examination or analysis of diagnose anomaly [14]. Such as Riper rule, using a decision tree to identify group of abnormality that are risk factors cardiovascular disease (metabolic syndrome) [2] or predict Parkinson's symptoms, using the k-nearest

neighbor to classify MRI brain image. The technical or knowledge gained from doing data mining, which has been validated or accepted in the medical community. Then the manufacturers of medical devices will develop the technical to help increase the ability of the tool for analysis.

2.1 Basic Knowledge of Acoustic Rhinometry

Acoustic rhinometry is measurement of the nasal internal structure by reflection of sound wave signal. The principle of tool is to generate an acoustic sound then pass the nasal cavity both two sides and measure the sound ware that echoes. when signals pass through the nasal structure at different area, a computer will calculate cross section area from the concentration of signal the echoes.in addition, it also calculate the distance from the nostril into position area where measuring.

2.2 Ripper Rule

The rules of the Ripper Rule [15] was created by Rich Cohen in 1995 is comprised of two phases. The first phase will identify the initial rules and the second phase will identify the post-process rule optimization data for the learning (Training data) is divided into a growing set and pruning set.

Therefore, the algorithm will create a relationship in greedy fashion rule as creating the Ripper Rule to find the best value for the growing set of rule space, which will be explained from the BNF. After growing set, it will be pruning the data immediately. When finished, the sample is common to the coverage rule of the training set, then it will be removed, the remaining training data, which is divided again. After learning the rules and to resolve problems that arise from segmentation errors [16] that this process is done until the results are satisfied.

2.3 C4.5 decision tree

Algorithm J48 decision tree or C4.5 algorithm is used to create a decision tree developed by Ross Quinlan [17]. C.45 extension that is added to the algorithm ID3 decision tree. This structure could be used for C4.5 classification and this reason it is called frequently for the statistical classifier in the C4.5 algorithm to build decision trees from the same training data ID3 principle of information entropy [18]. The C4.5 uses an accuracy of each list of attributes data for decision to split the data into sub-groups which will review the C4.5 normalized information gain (the difference in entropy). Results from the selected distribution list for the group data by feature with the highest normalized information gain is one of a decision.

Decision tree is created by done recursively in splitting the data set on the independent variables. Each possible split is evaluated by calculating the resulting purity gain if it was used to divide the data set D into the new subsets $\{D_1,...,D_n\}$. The purity gain Δ is the difference in impurity between the original data set and the subsets as defined as follow

$$\Delta = I(D) - \sum_{i=1}^{n} P(D_i) \cdot I(D_i)$$

Where $I(\cdot)$ is impurity measure of a given node and $P(D_i)$ is the proportion of D that is placed in D_i. The split resulting in the highest purity gain is selected, and repeat recursively for each subset in this split.

Different decision tree algorithms apply different impurity measures. C4.5 uses entropy as impurity measures as follow.

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) log_2 p(i|t)$$

Where, c is the number of classes and $p(i|t)$ is the the fraction of instances belonging to class i at the current node t.

2.4 K-Nearest neighbor

K-nearest neighbor technique is suitable for the classification problem which is algorithm in a group of supervised learning by setting data that close to the same group. This technique will imply that which class will replace the new conditions by examining the number of K if the terms of the decision is complicated. This approach can generate effective model and different from other techniques in that it does not use the training data to build the model but will use the data as a model. In the K-NN algorithm, we have to specify a positive integer to k which is defined as the number of cases to find the predicted new cases. K-NN algorithm such as 1-NN, 2-NN, 3-NN, …. K-NN where k instead of a positive integer such as 4-NN means this algorithm will find four cases are similar to new case (4 nearest cases) to predict new cases.

Classification of abnormal uses Euclidean metric to measure the dissimilarities between examples represented as vector inputs [19]. Euclidean distance is defined as the following formula.

$$d(x_i, y_j) = \sqrt{\sum_{r=1}^{n} w_r \left(a_r(x_i) - a_r(x_j)\right)^2}$$

Define x is vector input (a_1 , a_2 , a_3 , . . . , a_n), n is number of attributes of the vector input, a_r is the attribute, w_r is the weight of the attribute, r is from 1 to n

2.5 10 fold Cross-validation

Cross-validation is a method to evaluate performance of classifier by divide data into two groups for use in training and testing are independent. This method is a standard procedure in the experiments of classification with small number of sample. In the case of K - fold cross-validation, the data is divided into K set equally and calculates error values K round. Each round is calculated for a set of K data set and is selected for testing and the other K - 1 set is used as input for learning.

In this research used 10-fold cross-validation [20], Dataset is divided into 10 subsets (fold) for each subset data were the same. Dataset are divided into 10 subsets (Training data are 9 subsets and keep another for testing). The experiment is repeated 10 times, but changed the data set for training and testing.

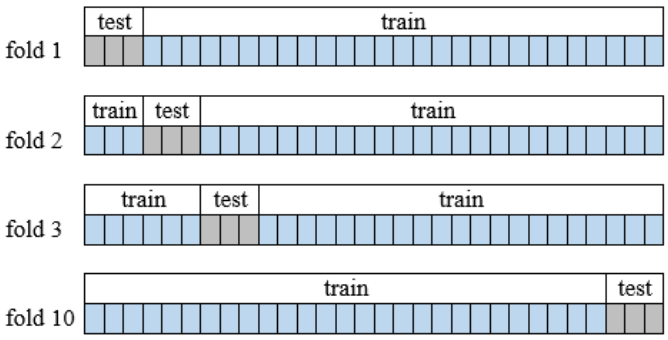


Figure 2.1 10-fold cross validation on a data

2.6 Paired T-test

Paired samples t-tests typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" t-test). Data of each pair is stored under the same condition but between pairs may not be the same condition. This is a data control (Treat) to illustrate the difference clearly. The analysis of the differences actions to different levels of each pair directly.

A typical example of the repeated measures t-test would be where subjects are tested prior to a treatment, say for needle hardness testing. There are two kinds of plastic components to find that both types of needle test, There are differences in performance testing or not. The evaluation indicates that such performance will measure the depth of the needle that press down on the plastic piece. Every time by pressing will set in the same pressure. If we choose a piece of plastic to test at random and every time of the experiment is changed work piece, when hypothesis testing, we need Two sample t-test to test. The experiments like this may be error results significantly because the differences such as using plastic in the different group, production lot or manufacturing date which as a result of a vary of plastic’s hardness. When measure the depth of the indentation, the results are difference. However, when using the two sample t-test, it will include the effects of various factors caused by the test needle only.

The way to define such differences is eliminating the effect of other factors as possible. The experiment will use the same piece for testing (Each pair of two types of the needle). It means first pair uses the same piece, second pair uses the next piece, so third pair and others pair will use the same way.

When the experiment are as follows, these called “Paired” and the data will be paired. As the equation:

$$\bar{d} = \sum_{i=1}^{n} \frac{d_i}{n}$$

Given $\bar{d}$ is the difference between the means of the samples and n is the number of trials.

2.7 Evaluation

2.7.1 Sensitivity and specificity

Sensitivity and specificity are statistical methods to test the efficiency of the categories classification. The Statistical function, which is commonly used in the medical community.

Sensitivity values of detection are the ratio of patients which are testing results were positive divide by all patients. In practice, we should choose the detection with high sensitivity in screening patients for diseases that are more serious but it can be treated. If the patient has not been diagnosed with the diseases, they will lose the benefits. It is also suitable for initial screening processes to reduce the number of patients who were determined specifically for the diagnosis further. The results in a way that has a high sensitivity value will more meaningful in case of result is negative. Because it means that patients has less likely to be diseased with this method.

The detection with high specificity value means that patients with a positive test result has the high opportunity to be a disease. Therefore, it is useful to confirm the diagnosis in case of the data from other detection that the patient likely to be sick with the disease. This feature is very useful in case of positive results to cause effect to

patients both the mind and the treatment of vulnerable. Therefore, detection with high specificity is very useful in case of a positive result.

In general, we usually expect the diagnostic method that developed with highest sensitivity and specificity but it is impossible. When increasing higher sensitivity, the specificity value of detection usually decreased. On the contrary, if the specificity value is higher, sensitivity value will lower typically. The cut - off point appropriate to classify between normal and abnormal in case of result with continuous data depends on the appropriate of high sensitivity or high specificity.

Screening for classifying and diagnosis in experiments, the test result is positive means predicted to be diseased or the test result is negative means predicted to normal. The results of the experiment might not be correspond to results of the medical diagnosis.

In that setting:

- True positive: Sick people correctly diagnosed as sick
- False positive: Healthy people incorrectly identified as sick
- True negative: Healthy people correctly identified as healthy
- False negative: Sick people incorrectly identified as healthy

In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Let us define an experiment from P positive instances and N negative instances for some condition. The four outcomes can be formulated in a 2×2 contingency table or confusion matrix, as follows:

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	True positive	False positive (Type I error)	Precision = Σ True positive / Σ Test outcome positive
	Test outcome negative	False negative (Type II error)	True negative	Negative predictive value = Σ True negative / Σ Test outcome negative
		Sensitivity = Σ True positive / Σ Condition positive	Specificity = Σ True negative / Σ Condition negative	Accuracy

Figure 2.2 Contingency table or confusion matrix

Sensitivity

Sensitivity relates to the test's ability to identify positive results. The sensitivity of a test is the proportion of people that are known to have the disease who test positive for it. This can also be written as:

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

Sensitivity is not the same as the precision or positive predictive value (ratio of true positives to combined true and false positives), which is as much a statement about the proportion of actual positives in the population being tested as it is about the test.

Specificity

Specificity relates to the test's ability to identify negative results. Consider the example of the medical test used to identify a disease. The specificity of a test is defined as the proportion of patients that are known not to have the disease who will test negative for it. This can also be written as:

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}$$

However, highly specific tests rarely miss negative outcomes, so they can be considered reliable when their result is positive. Therefore, a positive result from a test with high specificity means a high probability of the presence of disease.

Another method that can be used to select the appropriate the cut-off point is creating a Receiver Operator Characteristic (ROC) curve. To create a relationship graph between the true positive rate (Sensitivity) and false positive rate (1 - Specificity) by changing the cut - off point. In addition, creating a ROC curve also helps in comparison the efficiency of the diagnosis by comparing the area under the lines of each test. The area under curve represents the higher performance.

2.7.2 Receiver Operating Characteristic (ROC)

Receiver operating characteristic (ROC) or ROC curve is creating a graph that shows the performance of the algorithm which will change according to the Threshold setting. It is created from the plot of the Sensitivity and Specificity values so it shows the tradeoff between the sensitivity and specificity.

If a sensitivity value is increased, a specificity value will be increased respectively. Therefore, if the algorithm is better, the error will rise as well. We can see that the researchers around the world are trying to solve this problem and seek to maximize or minimize values of specificity.

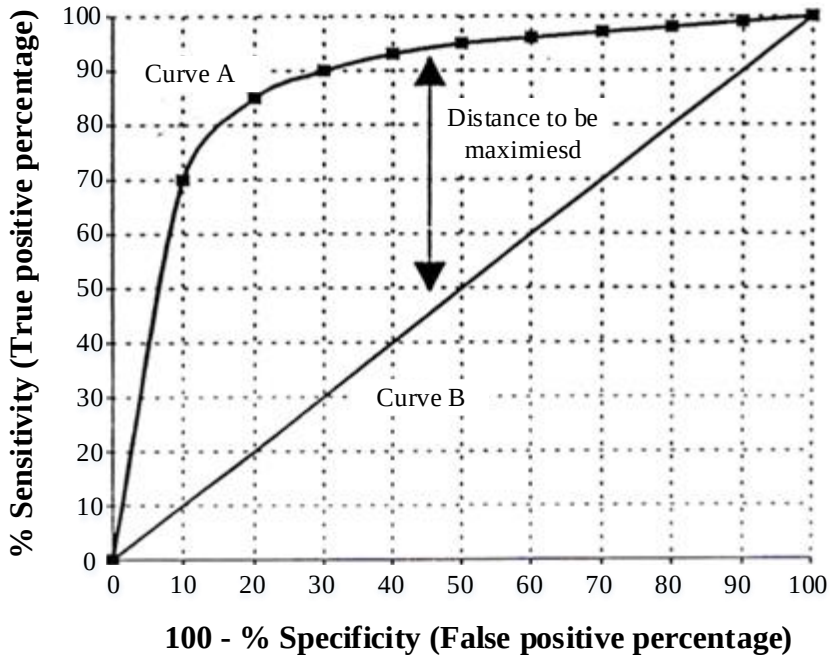


Figure 2.3 ROC curve: Illustrates the relationship between Sensitivity and Specificity

Figure 2.3 if a graph curve approaches the point that is on the left corner, the slope graph is high and area under the curve is greater. It means that algorithm has a good performance.

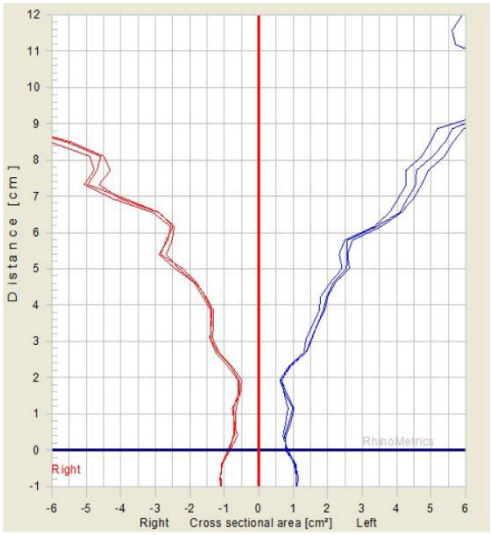
CHAPTER 3
RESEARCH METHODOLOGY

In this research, we use data mining for creating a model to classify the nasal cavity's abnormalities. Data come from the measurement by Acoustic Rhinometry of the patients with nasal cavity troubles. We do the experiments for measuring the efficiency in abnormalities classification of nasal cavity. Moreover, to compare each of the methods to find the best way for classification. The methods and related knowledge are following.

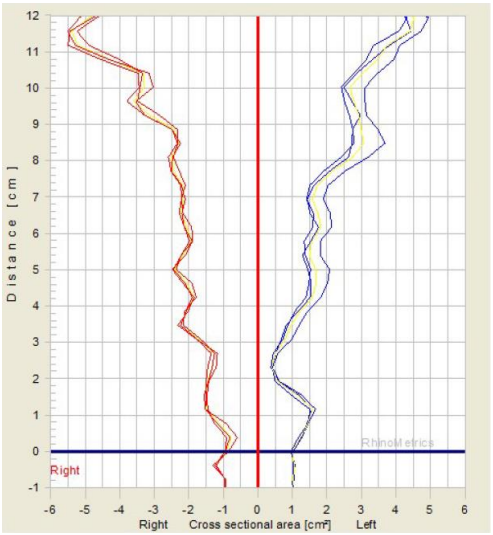
3.1 Nasal cross sectional area dataset

The data used in this research comes from a study in patients with nasal and sinus that using Acoustic Rhinometry 114 people (68 male, 46 female), age between 18-66 years from Otolaryngplogy head and neck surgery Department, HRH Princess Maha Chakri Sirindhorn medical center, Srinakhanarinwirot University, Thailand. This study was during 2011-2012.

As the nasal cross sectional area in several distance of the nasal cavity 2904 records (Left 1452 record, Right 1452 record). There are consists of following attributes: patient, gender, age and nose side. The distance of the nasal cross section area from -3.85 cm. to 20.43 cm. (with phase increased by 0.38 cm.) total of 64 attributes. The last attribute is the result from diagnosed by given 0 is 'normal' and 1 is 'abnormal', details are as follows.



a) Normal nasal cavity



b) Abnormal nasal cavity

Figure 3.1 Shows sample graphs of cross sectional area (Red is the right, Blue is the left)

Table 3.1 Attributes used in the experiments

Attribute	Attribute name	Description
1	Gender	Patient gender
2	Age	Patient age
3	Side	Nose side
4	D1	Distance -3.85 cm.
5	D2	Distance -3.47 cm.
….	….	….
63	D63	Distance 20.04 cm.
67	D64	Distance 20.43 cm.
68	Class	Diagnose

3.2 Experimental process

The objective of this experiment is to model abnormalities classification and comparison of the performance with others models which has 3 steps as shown.

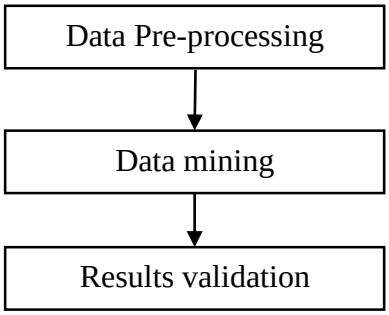


Figure 3.2 Experimental process

First, pre-processing is to perform data cleaning by removing the irrelevant information based on data from acoustic rhinometry such as date, time, patient name, patient ID, the type of equipment, software version and incomplete data. Second, use data that passed pre-processing to model abnormalities classification of nasal cavity. Finally, compare models to find the most effective by using paired t-test.

3.2.1 Data Pre-processing

From collected data which have some attributes cannot be used so in this process need to be cut and transform the data into a format that can be used. As following table:

Table 3.2 Show Attribute that used for the Data Pre-processing.

Old Attribute name	New Attribute name	Description
Patient Name	-	Patient Name
Gender	Gender	Patient gender
Pat.Born	Age	Patient age

Table 3.3 Show Attribute that used for the Data Pre-processing. (cont.)

Old Attribute name	New Attribute name	Description
Diagnose Date	-	Diagnose Date
Side	Side	Nose side
D1	D1	Distance -3.85 cm.
D2	D2	Distance -3.47 cm.
….	….	….
D63	D63	Distance 20.04 cm.
D64	D64	Distance 20.43 cm.
Class	Class	Diagnose

3.2.2 Data mining and Results validation

Data mining technique that use in this study is classification by using C4.5 Decision trees, Ripper Rule, K-Nearest neighbor. In this thesis, the modeling for nasal cavity classification in case of abnormal, we create modeling and validation simultaneously. The experiment is modeled by 30 times by using 10 fold Cross-validation in order to divide data sets. Then we use Paired T-test technique which to make the same set of data for creating a model and validation. The last step is to compare the performance of model by finding the ROC curve.

CHAPTER 4
RESULTS AND DISCUSSION

In classification of abnormal divided data into three sets are right nasal cavity, left nasal cavity and both sides of nasal cavity by using algorithm C4.5 decision tree, Ripper Rule, K-Nearest neighbor. Then compere performance of classifier by paired t-test which is used 10-fold cross-validation to run data. Each of the test, new data is generated based on cross-validation and tested 30 times. The result will be an average of all the tests by compare with a baseline classifier (here it is the C4.5 decision tree), the output uses the annotation v or * to indicate that a specific result is statistically better (v) or worse (*) than the baseline scheme at the significance level specified (currently 0.05)

Table 4.1 Summary the performance of the classification from the right nasal cavity 1,452 recodes (numbers in parentheses are standard deviations.)

Algorithm	Accuracy	ROC	Sensitivity	Specificity
Ripper Rule	96.07 (1.92)	0.97 (0.02)	0.96 (0.02)	0.96 (0.03)
C4.5 decision tree	99.48 (0.60)	0.99^{V} (0.01)	0.99 (0.01)	0.99 (0.01)
K-Nearest neighbor	93.62 (1.62)	0.94 (0.01)	0.93 (0.03)	0.94 (0.03)

Table 4.2 Summary the performance of the classification from the left nasal cavity 1,452 recodes (Numbers in parentheses are standard deviations.)

Algorithm	Accuracy	ROC	Sensitivity	Specificity
Ripper Rule	95.64 (2.16)	0.96 (0.02)	0.96 (0.02)	0.94 (0.04)
C4.5 decision tree	98.90 (0.91)	0.99^{V} (0.01)	0.99 (0.01)	0.99 (0.02)
K-Nearest neighbor	92.31 (1.83)	0.92 (0.02)	0.93 (0.03)	0.91 (0.04)

Table 4.3 Summary the performance of the classification from the both sides of nasal cavity 1,452 recodes (Numbers in parentheses are standard deviations.)

Algorithm	Accuracy	ROC	Sensitivity	Specificity
Ripper Rule	96.07 (1.92)	0.97 (0.02)	0.96 (0.02)	0.96 (0.03)
C4.5 decision tree	99.48 (0.60)	0.99^{V} (0.01)	0.99 (0.01)	0.99 (0.01)
K-Nearest neighbor	93.62 (1.62)	0.94 (0.01)	0.93 (0.03)	0.94 (0.03)

Table 4.1 - Table 4.2 show that the best efficiency classification algorithm is C4.5 decision tree and when test data of both sides of nasal cavity from Table 4.3 has ROC 0.99 (standard deviations 0.1), Sensitivity 0.99 and Specificity 0.99 with Tree of classification from Table 4.3

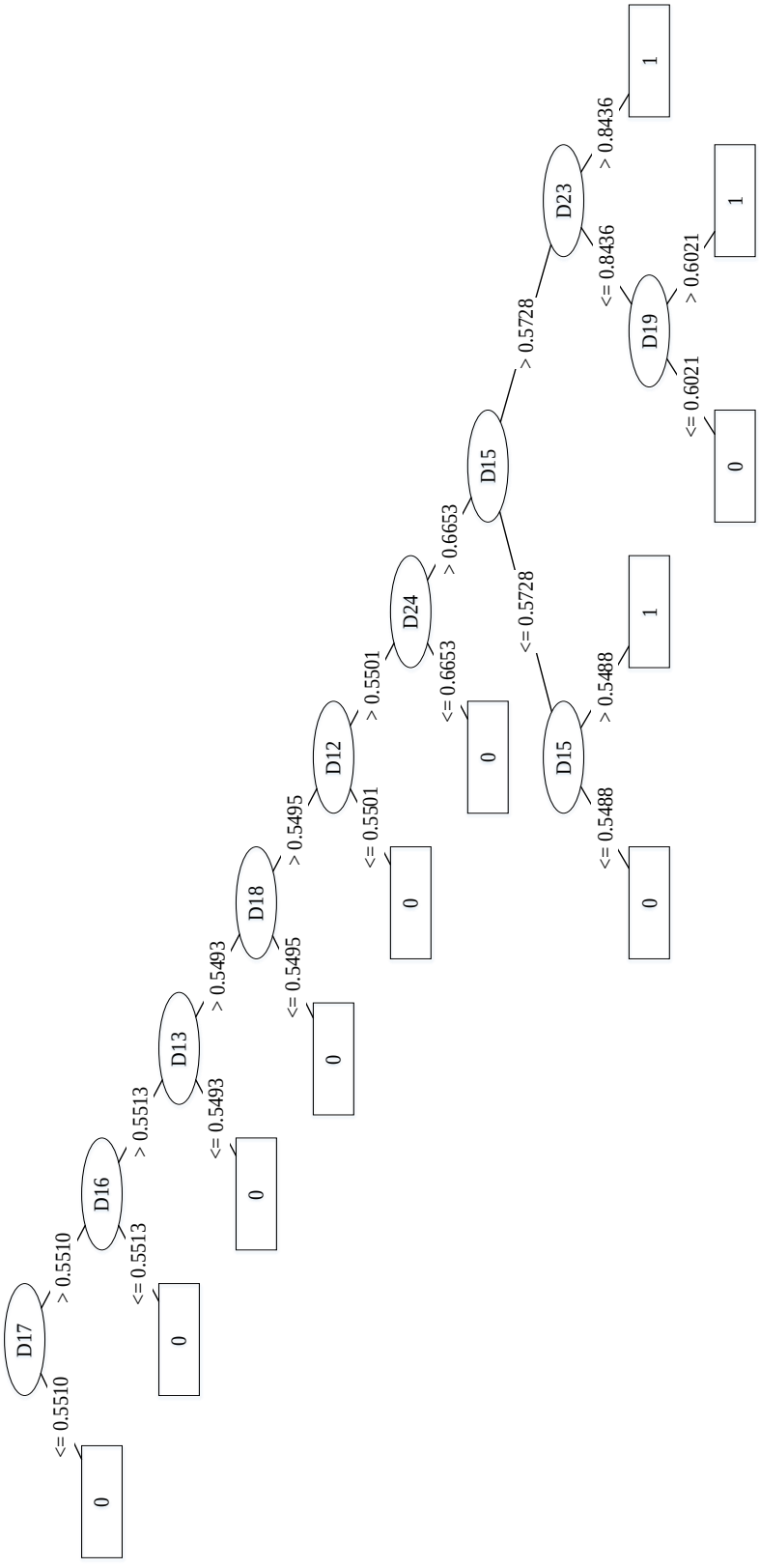


Figure 4.1 Shows the tree from the classification of the abnormal nasal cavity with the C4.5 decision tree. Define 0 is abnormal, 1 is normal

CHAPTER 5
CONCLUSION

The results showed that C4.5 decision tree is an algorithm which is suitable for classification of abnormal nasal cavity detected by Acoustic Rhinometry. From Figure 4.1 when analyzed Tree found that abnormalities of nasal cavity is approximately 0.0 – 4.24 cm and nasal cross sectional area less than 0.55 cm^2 which is close to medical research related to the average of Nasal Cross sectional area of Thailand [21].

REFERENCES

1 CS. Kim, BK. Moon, DH. Jung, and YG. Min. (1998). Correlation between nasal obstruction symptoms and objective parameters of acoustic rhinometry and rhinomanometry. Auris Nasus Larynx, vol, 25, 45–48.

2 Kern EB. (1981). *Committee report on standardization of rhinomanometry*. Rhinology, 231-236.

3 Roithmann R, Cole P, and Chapnik J. (1995). *Acoustic rhinometry in the evaluation of nasal obstruction*. Laryngoscope, vol, 105, 275–281.

4 Fisher EW, Scadding GK, and Lund VJ. (1993). *The role of acoustic rhinometry in studying the nasal cycle*. Rhinology, vol, 31, 57–61.

5 Mostafa BE. (1997). *Detection of adenoidal hypertrophy using acoustic rhinomanometry*. Eur Arch Otorhinolaryngol, vol, 254, suppl, 1, S27–S29.

6 Lenders H, and Pirsig W. (1992). *Acoustic rhinometry: A diagnostic tool for patients with chronic rhonchopathies*. Rhinol Suppl 14, 101–105.

7 Djupesland P, Kaastad E, and Franzen D. (1997). *Acoustic rhinomety in the evaluation of congenital choanal malformations*. Int J Pediatr Otorhinolaryngol, vol, 41, 319–337.

8 Marais J, Murray JAM, Marshall I. (1994). *Minimal crosssectional area, nasal peak flow and patients' satisfaction in septoplasty and inferior turbinectomy*. Rhinology, vol, 32, 145–147.

9 Hilberg O. (2002). *Objective measurement of nasal airway dimensions using acoustic rhinometry: methodological and clinical aspects*. Allergy, vol, 57, suppl, 70, 5-39,

10 Hilberg O. (1989). *Jackson AC, Swift Dl, and Pedersen OF. Acoustic rhinometry: Evaluation of nasal cavity geometry by acoustic reflection*. Journal of Applied Physiology, vol, 66, 295–303.

11 Clement PA, and Gordts F. (2005). *Standardization committee on objective assessment of the nasal airway, IRS and ERS*. Rhinology, vol, 43, 169–179.

12 Huang ZL, Wang DY, Zhang PC, Dong F, and Yeoh KH. (2001). *Evaluation of Nasal Cavity by Acoustic Rhinometry in Chinese, Malay and Indian Ethnic Groups*. Acta Otolaryngol, vol, 121, 844–848.

13 Alireza M, Mohammad F, and Artemis E. (2008). *Assessment of Nasal Volume and Cross-Sectional Area by Acoustic Rhinometry in a Sample of Normal Adult Iranians*. Archives of Iranian Medicine, vol, 11, 555 – 558.

14 Abe H., Yokoi H., Ohsaki M., Yamaguchi T. (2007). *Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining*. Seventh IEEE International Conference on Data Mining - Workshops, 127 – 132.

15 W. Lee, K. W. Mok, and S. J. Stolfo. (1998). *Mining sequential patterns: Techniques, visualization, and applications*. submitted for publication.

16 W. W. Cohen. (1995). *Fast effective rule induction*. Machine Learning: the 12th International Conference, Lake Taho, CA.

17 JR. Quinlan. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

18 JR. Quinlan. (1996). *Improved use of continuous attributes in c4.5*. Journal of Artificial Intelligence Research, vol, 4, 77-90.

19 K.Q. Weinberger and L.K. Saul. (2009). *Distance metric learning for large margin nearest neighbor classification*. The Journal of Machine Learning Research, vol, 10, 207-244.

20 R. Kohavi. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1137–1143 ,

21 P Tantilipikorn, P Jareoncharsri, S Voraprayoon,C Bunnag, Clement, and Peter A. (2008). *Acoustic rhinometry of Asian noses*. American Journal of Rhinology, Vol, 22, 617-620(4).

22 Worachartcheewan A,Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P and Prachayasittikul V. (2010). *Identification of metabolic syndrome using decision tree analysis*. Diabetes Research and Clinical Practice, vol, 90, e15-e18.

APPENDICES

APPENDIX A
EXPERIMENTAL OUTPUT

C4.5 decision tree output

Number of Leaves : 11

Size of the tree : 21

```
2.3129251 <= 0.551067: 0 (1217.0)
2.3129251 > 0.551067
|  1.92743757 <= 0.5513: 0 (183.0)
|  1.92743757 > 0.5513
|  |  0.77097499 <= 0.5493: 0 (131.0)
|  |  0.77097499 > 0.5493
|  |  |  2.69841263 <= 0.5495: 0 (52.0)
|  |  |  2.69841263 > 0.5495
|  |  |  |  0.38548747 <= 0.5501: 0 (31.0)
|  |  |  |  0.38548747 > 0.5501
|  |  |  |  |  5.01133779 <= 0.6653: 0 (27.0)
|  |  |  |  |  5.01133779 > 0.6653
|  |  |  |  |  |  1.54195005 <= 0.5728
|  |  |  |  |  |  |  1.54195005 <= 0.5488: 0 (14.0)
|  |  |  |  |  |  |  1.54195005 > 0.5488: 1 (11.0)
|  |  |  |  |  |  1.54195005 > 0.5728
|  |  |  |  |  |  |  4.62585026 <= 0.8436
|  |  |  |  |  |  |  |  3.08390015 <= 0.6021: 0 (3.0)
|  |  |  |  |  |  |  |  3.08390015 > 0.6021: 1 (22.0)
|  |  |  |  |  |  |  4.62585026 > 0.8436: 1 (1155.0/3.0)
```

C4.5 decision tree Stratified cross-validation

Correctly Classified Instances	2833	99.5432 %
Incorrectly Classified Instances	13	0.4568 %
Kappa statistic	0.9906	
Mean absolute error	0.0064	
Root mean squared error	0.0676	
Relative absolute error	1.3106 %	
Root relative squared error	13.7036 %	
Total Number of Instances	2846	

JRIP rules output

Number of Rules : 7

Rule 1 = (2.3129251 >= 0.5612) and (1.15646252 >= 0.7933) and (6.55328789 >= 3.0544) and (1.15646252 >= 0.9777) and (2.69841263 >= 0.579867) => ans=1 (408.0/3.0)

Rule 2 = (2.3129251 >= 0.5512) and (1.92743757 >= 0.5746) and (0.77097499 >= 0.5516) and (3.08390015 >= 0.8886) and (-0.00000006 >= 0.8102) => ans=1 (571.0/11.0)

Rule 3 = (2.3129251 >= 0.5512) and (0.77097499 >= 0.687567) and (1.92743757 >= 0.56) and (1.92743757 <= 0.8956) and (4.24036273 >= 0.7165) => ans=1 (129.0/12.0)

Rule 4 = (1.92743757 >= 0.9134) and (2.69841263 >= 0.556067) and (1.54195005 >= 1.1608) => ans=1 (42.0/4.0)

Rule 5 = (2.3129251 >= 0.591) and (0.77097499 >= 0.5539) and (1.92743757 >= 0.5525) and (0.38548747 <= 0.7738) and (0.38548747 >= 0.5833) and (3.08390015 >= 0.7327) => ans=1 (43.0/2.0)

Rule 6 = (2.3129251 >= 0.5556) and (1.92743757 >= 0.5604) and (0.77097499 >= 0.5506) and (2.69841263 >= 0.6185) and (17.34693864 >= 3.917233) and (-0.00000006 >= 0.7882) => ans=1 (19.0/2.0)

Rule 7 = => ans=0 (1634.0/7.0)

JRIP rules Stratified cross-validation

Correctly Classified Instances	2801	98.4188 %
Incorrectly Classified Instances	45	1.5812 %
Kappa statistic	0.9675	
Mean absolute error	0.0229	
Root mean squared error	0.1232	
Relative absolute error	4.7121 %	
Root relative squared error	24.9894 %	
Total Number of Instances	2846	

K-Nearest neighbor Stratified cross-validation

Correctly Classified Instances	2620	92.059 %
Incorrectly Classified Instances	226	7.941 %
Kappa statistic	0.8353	
Mean absolute error	0.1153	
Root mean squared error	0.2497	
Relative absolute error	23.7291 %	
Root relative squared error	50.6546 %	
Total Number of Instances	2846	

APPENDIX B
ATTRIBUTES USED IN THE EXPERIMENTS

Attributes used in the experiments

Attribute	Attribute name	Description
1	Gender	Patient gender
2	Age	Patient age
3	Side	Nose side
4	D1	Distance -3.85 cm.
5	D2	Distance -3.47 cm.
6	D6	Distance -3.08 cm.
7	D7	Distance -2.70 cm.
8	D8	Distance -2.31 cm.
9	D9	Distance -1.93 cm.
10	D10	Distance -1.54 cm.
11	D11	Distance -1.16 cm.
12	D12	Distance -0.77 cm.
13	D13	Distance -0.39 cm.
14	D14	Distance 0.00 cm.
15	D15	Distance 0.39 cm.
16	D16	Distance 0.77 cm.

Attribute	Attribute name	Description
17	D17	Distance 1.16 cm.
18	D18	Distance 1.54 cm.
19	D19	Distance 1.93 cm.
20	D20	Distance 2.31 cm.
21	D21	Distance 2.70 cm.
22	D22	Distance 3.08 cm.
23	D23	Distance 3.47 cm.
24	D24	Distance 3.85 cm.
25	D25	Distance 4.24 cm.
26	D26	Distance 4.63 cm.
27	D27	Distance 5.01 cm.
28	D28	Distance 5.40 cm.
29	D29	Distance 5.78 cm.
30	D30	Distance 6.17 cm.
31	D31	Distance 6.55 cm.
32	D32	Distance 6.94 cm.
33	D33	Distance 7.32 cm.
34	D34	Distance 7.71 cm.
35	D35	Distance 8.10 cm.
36	D36	Distance 8.48 cm.
37	D37	Distance 8.87 cm.
38	D38	Distance 9.25 cm.

Attribute	Attribute name	Description
39	D39	Distance 9.64 cm.
40	D40	Distance 10.02 cm.
41	D41	Distance 10.41 cm.
42	D42	Distance 10.79 cm.
43	D43	Distance 11.18 cm.
44	D44	Distance 11.56 cm.
45	D45	Distance 11.95 cm.
46	D46	Distance 12.34 cm.
47	D47	Distance 12.72 cm.
48	D48	Distance 13.11 cm.
49	D49	Distance 13.49 cm.
50	D50	Distance 13.88 cm.
51	D51	Distance 14.26 cm.
52	D52	Distance 14.65 cm.
53	D53	Distance 15.03 cm.
54	D54	Distance 15.42 cm.
55	D55	Distance 15.80 cm.
56	D56	Distance 16.19 cm.
57	D57	Distance 16.58 cm.
58	D58	Distance 16.96 cm.
59	D59	Distance 17.35 cm.
60	D60	Distance 17.73 cm.

Attribute	Attribute name	Description
61	D61	Distance 18.12 cm.
62	D62	Distance 18.50 cm.
63	D63	Distance 18.89 cm.
64	D64	Distance 19.27 cm.
65	D65	Distance 19.66 cm.
66	D63	Distance 20.05 cm.
67	D64	Distance 20.43 cm.
68	Class	Diagnose

APPENDIX C

DIAGNOSE ABNORMAL NASAL BASED ON THE C4.5 MODELING USING CROSS SECTION AREA CURVE FROM ACOUSTIC RHINOMETRY

Wasin Srisawat, Adisorn Leelasantitham, Waranyu Wongseree and Supaporn Kiattisin
Technology of Information System Management Program, Faculty of Engineering, Mahidol University
25/25 Puttamonthon, Nakorn Pathom 73170, Thailand
tong.wasin@gmail.com, adisorn.lee@mahidol.ac.th, waranyu.won@mahidol.ac.th,
supaporn.kit@mahidol.ac.th

***Abstract*—This research proposes methods to classify the pattern of unusual nasal cavity using Ripper Rule, C4.5 decision tree, K-Nearest neighbor which aims to help physicians classify abnormal nasal cavity from acoustic rhinometry signal. The experiments showed that the algorithm was best effective classification is C4.5 decision tree has ROC 0.99 (sensitivity 0.99, specificity 0.99 and standard deviation 0.1). The result showed that abnormalities of the nasal cavity are about 0.3 – 5 cm. and nasal cross sectional area is less than 0.55 cm.^{2}. Therefore, this study suggests that the C4.5 decision tree algorithm could apply for screening abnormal nasal cavity. It led to application or tool development on medical devices in the future.**

Keywords— Nasal cross sectional area; Acoustic Rhinometry; Classification; Ripper Rule; C4.5 decision tree; K-Nearest neighbor;

INTRODUCTION

Nasal and sinus disease are common diseases and public health problems in the country. These diseases, which cause blockages in the nasal cavity because of the nasal mucosa swelling by inflammation. Patients often have a stuffed nose and other symptoms such as running nose, the nose does not smell, pain, nasal congestion, sneezing, etc.

Currently, Medical could diagnose the congested nose by objective measurement (measuring nasal patency) with two tools [1] are:

- Active Anterior Rhinomanometry (AAR) which is a measure of resistance's nasal cavity.
- Acoustic Rhinometry which is a measure of cross sectional area and nasal volume.

The nasal cavity with no choke should has low nasal resistance, proper nasal volume and minimal cross sectional area.

Rhinomanometry is considered as a standard diagnosis (Gold standard) [2] that assessment of nasal patency is widely used in clinical. But in actual practice, Rhinomanometry is a measure with high costs, expensive equipment, requiring staff with specialists and cooperation from patient particularly young patients. The simple way to evaluate nasal patency is using acoustic rhinometry to measure cross section area of nasal cavity.

Acoustic Rhinometry measures the structure inside nose that easily in order to track the treatment and evaluate blockages in the nasal cavity [3-8] by the reflection of sound waves [9-10] for cross-sectional area of nasal passages and distance from the nostril into the measured position. This tool cannot analyze primary abnormalities of the nasal cavity. Physician required information of cross sectional area and pattern of nasal cavity. Then data were compared with the pattern of the nasal cavity that has a statistical study in Chinese, Malay,Indian [12] and Iraq [13] for examination of position and areas where there is a blockage in nasal passages.

Present, medical have been used data mining techniques to assist in the examination or analysis of diagnose anomaly[14]. Such as Riper rule, using a decision tree to identify group of abnormality that are risk factors cardiovascular disease (metabolic syndrome)[2] or predict Parkinson' s symptoms,

using the k-nearest neighbor to classify MRI brain image.The technical or knowledge gained from doing data mining, which has been validated or accepted in the medical community. Then the manufacturers of medical devices will develop the technical to help increase the ability of the tool for analysis.

This paper proposes a method of Data mining for analyze, build and identify the abnormal pattern of nasal cavity which will as guidelines for develop software that is used to help Physician analyze Thai people's disease of nasal cavity.

MATERIALS AND METHODS

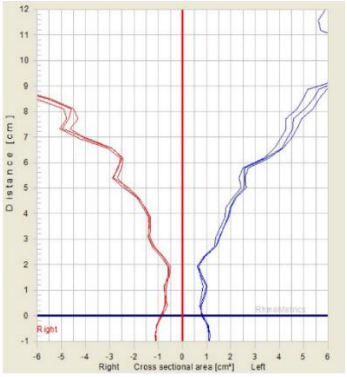
A. Basic Knowledge of Acoustic Rhinometry

Acoustic rhinometry is measurement of the nasal internal structure by reflection of sound wave signal. The principle of tool is to generate an acoustic sound then pass the nasal cavity both two sides and measure the sound ware that echoes. when signals pass through the nasal structure at different area, a computer will calculate cross section area from the concentration of signal the echoes.in addition, it also calculate the distance from the nostril into position area where measuring.

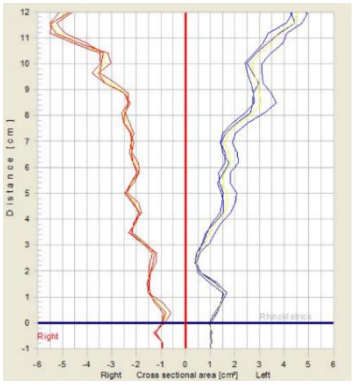
B. Nasal cross sectional area dataset

The data used in this research comes from a study in patients with nasal and sinus that using Acoustic Rhinometry 114 people (68 male, 46 female), age between 18-66 years. from Otolaryngplogy head and neck surgery Department, HRH Princess Maha Chakri Sirindhorn medical center, Srinakhanarinwirot University

As the nasal cross sectional area in several distance of the nasal cavity 2904 records (Left 1452 record, Right 1452 recorda). The distance of the nasal cross section area from -3.85 cm. to 20.43 cm. (with phase increased by 0.38 cm.) total of 64 attributes. Details are as follows.



a) Normal nasal cavity



b) Abnormal nasal cavity

Fig. 1. Shows sample graphs of cross sectional area (Red is the right, Blue is the left)

TABLE I. ATTRIBUTES USED IN THE EXPERIMENTS

Attribute	Attribute name	Description
1	D1	Distance - 3.85 cm.
2	D2	Distance - 3.46 cm.
….	….	….
63	D63	Distance 20.04 cm.
64	D64	Distance 20.43 cm.

Experimental methods

The objective of this experiment is to model abnormalities classification and comparison of the performance with others models which has 3 steps as shown.

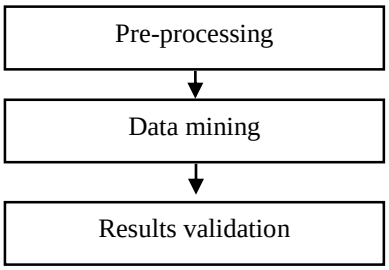


Fig. 2. Experimental process

First, pre-processing is to perform data cleaning by removing the irrelevant information based on data from acoustic rhinometry such as date, time, patient name, patient ID, the type of equipment, software version and incomplete data. Second, use data that passed pre-processing to

model abnormalities classification of nasal cavity. Finally, compare models to find the most effective by using paired t-test.

A. Ripper Rule

The rules of the Ripper Rule [15] was created by Rich Cohen in 1995 is comprised of two phases. The first phase will identify the initial rules and the second phase will identify the post-process rule optimization data for the learning (Training data) is divided into a growing set and pruning set.

Therefore, the algorithm will create a relationship in greedy fashion rule as creating the Ripper Rule to find the best value for the growing set of rule space, which will be explained from the BNF. After growing set, it will be pruning the data immediately. When finished, the sample is common to the coverage rule of the training set, then it will be removed, the remaining training data, which is divided again. After learning the rules and to resolve problems that arise from segmentation errors [16] that this process is done until the results are satisfied.

B. C4.5 decision tree

Algorithm J48 decision tree or C4.5 algorithm is used to create a decision tree developed by Ross Quinlan [17]. C.45 extension that is added to the algorithm ID3 decision tree. This structure could be used for C4.5 classification and this reason it is called frequently for the statistical classifier in the C4.5 algorithm to build decision trees from the same training data ID3 principle of information entropy [18]. The C4.5 uses an accuracy of each list of attributes data for decision to split the data into sub-groups which will review the C4.5 normalized information gain (the difference in entropy). Results from the selected distribution list for the group data by feature with the highest normalized information gain is one of a decision.

C. K-Nearest neighbor

K-nearest neighbor technique is suitable for the classification problem which is algorithm in a group of supervised learning by setting data that close to the same group. This technique will imply that which class will replace the new conditions by examining the number of K if the terms of the decision is complicated. This approach can generate effective model and different from other techniques in that it does not use the training data to build the model but will use the data as a model. In the K-NN algorithm, we have to specify a positive integer to k which is defined as the number of cases to find the predicted new cases. K-NN algorithm such as 1-NN, 2-NN, 3-NN, …. K-NN where k instead of a positive integer such as 4-NN means this algorithm will find four cases are similar to new case (4 nearest cases) to predict new cases.

Classification of abnormal uses Euclidean metric to measure the dissimilarities between examples represented as vector inputs [19]. Euclidean distance is defined as the following formula.

$$d(x_i, y_j) = \sqrt{\sum_{r=1}^{n} w_r \left(a_r(x_i) - a_r(x_j)\right)^2} \qquad (1)$$

Define x is vector input (a_1, a_2, a_3, . . . , a_n), n is number of attributes of the vector input, a_r is the attribute, w_r is the weight of the attribute, r is from 1 to n

D. 10 fold Cross-validation

Cross-validation is a method to evaluate performance of classifier by divide data into two groups for use in training and testing are independent. This method is a standard procedure in the experiments of classification with small number of sample.

In the case of K - fold cross-validation, the data is divided into K set equally and calculates error values K round. Each round is calculated for a set of K data set and is selected for testing and the other K - 1 set is used as input for learning.

In this research used 10-fold cross-validation [20], Dataset is divided into 10 subsets (fold) for each subset data were the same. Dataset are divided into 10 subsets (Training data are 9 subsets and keep another for testing). The experiment is repeated 10 times, but changed the data set for training and testing.

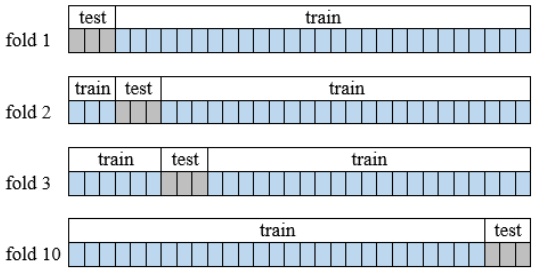


Fig. 3. 10-fold cross validation on a data

Experimental Result

In classification of abnormal divided data into three sets are right nasal cavity, left nasal cavity and both sides of nasal cavity by using algorithm C4.5 decision tree, Ripper Rule, K-Nearest

neighbor. Then compere performance of classifier by paired t-test which is used 10-fold cross-validation to run data. Each of the test, new data is generated based on cross-validation and tested 30 times. The result will be an average of all the tests by compare with a baseline classifier (here it is the C4.5 decision tree), the output uses the annotation v or * to indicate that a specific result is statistically better (v) or worse (*) than the baseline scheme at the significance level specified (currently 0.05

TABLE II. SUMMARY THE PERFORMANCE OF THE CLASSIFICATION FROM THE RIGHT NASAL CAVITY 1,452 RECODES (NUMBERS IN PARENTHESES ARE STANDARD DEVIATIONS.)

Algorithm	Accuracy	ROC	Sensitivity	Specificity
Ripper Rule	96.07 (1.92)	0.97 (0.02)*	0.96 (0.02)	0.96 (0.03)
C4.5 decision tree	99.48 (0.60)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
K-Nearest neighbor	93.62 (1.62)	0.94 (0.01)*	0.93 (0.03)	0.94 (0.03)

TABLE III. SUMMARY THE PERFORMANCE OF THE CLASSIFICATION FROM THE LEFT NASAL CAVITY 1,452 RECODES (NUMBERS IN PARENTHESES ARE STANDARD DEVIATIONS.)

Algorithm	Accuracy	ROC	Sensitivity	Specificity
Ripper Rule	95.64 (2.16)	0.96 (0.02)*	0.96 (0.02)	0.94 (0.04)
C4.5 decision tree	98.90 (0.91)	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)
K-Nearest neighbor	92.31 (1.83)	0.92 (0.02)*	0.93 (0.03)	0.91 (0.04)

TABLE IV. SUMMARY THE PERFORMANCE OF THE CLASSIFICATION FROM THE BOTH SIDES OF NASAL CAVITY 1,452 RECODES (NUMBERS IN PARENTHESES ARE STANDARD DEVIATIONS.)

Algorithm	Accuracy	ROC	Sensitivity	Specificity
Ripper Rule	96.07 (1.92)	0.97 (0.02)*	0.96 (0.02)	0.96 (0.03)
C4.5 decision tree	99.48 (0.60)	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)
K-Nearest neighbor	93.62 (1.62)	0.94 (0.01)*	0.93 (0.03)	0.94 (0.03)

Table II-III show that the best efficiency classification algorithm is C4.5 decision tree and when test data of both sides of nasal cavity from Table IV has ROC 99.49% (standard deviations 0.1), Sensitivity 0.99 and Specificity 0.99 with Tree of classification from Figure 2.

CONCLUSION

The results showed that C4.5 decision tree is an algorithm which is suitable for classification of abnormal nasal cavity detected by Acoustic Rhinometry. From Figure 2 when analyzed Tree found that abnormalities of nasal cavity is approximately 0.3 - 5 cm and nasal cross sectional area less than 0.55 cm2 which is close to medical research related to the average of Nasal Cross sectional area of Thailand [21].

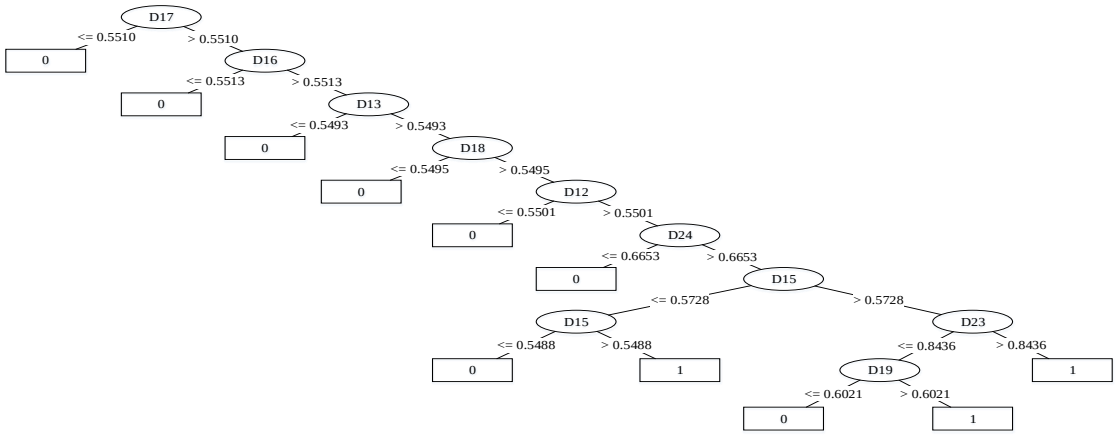


Fig. 4. hows the tree from the classification of the abnormal nasal cavity with the C4.5 decision tree.Define 0 is abnormal, 1 is normal

ACKNOWLEDGMENT

The authors acknowledge Asst.Prof. Visan mahasithiwat MD. from Otolaryngplogy head and neck surgery Department, HRH Princess Maha Chakri Sirindhorn medical center, Srinakhanarinwirot University for his assistance on capturing nasal cross sectional area data.

References

[1] CS. Kim, BK. Moon, DH. Jung, and YG. Min, "Correlation between nasal obstruction symptoms and objective parameters of acoustic rhinometry and rhinomanometry," Auris Nasus Larynx, vol. 25, pp. 45–48, 1998

[2] Kern EB, **"**Committee report on standardization of rhinomanometry," Rhinology, pp. 231-236, 1981

[3] Roithmann R, Cole P, and Chapnik J, "Acoustic rhinometry in the evaluation of nasal obstruction," Laryngoscope, vol. 105, pp. 275–281, 1995.

[4] Fisher EW, Scadding GK, and Lund VJ. "The role of acoustic rhinometry in studying the nasal cycle," Rhinology, vol. 31, pp. 57–61, 1993.

[5] Mostafa BE, "Detection of adenoidal hypertrophy using acoustic rhinomanometry," Eur Arch Otorhinolaryngol, vol. 254, suppl. 1, pp. S27–S29, 1997.

[6] Lenders H, and Pirsig W. Acoustic rhinometry: A diagnostic tool for patients with chronic rhonchopathies. Rhinol Suppl 14:101–105, 1992.

[7] Djupesland P, Kaastad E, and Franzen D, "Acoustic rhinomety in the evaluation of congenital choanal malformations," Int J Pediatr Otorhinolaryngol, vol. 41, pp. 319–337, 1997.

[8] Marais J, Murray JAM, Marshall I, "Minimal crosssectional area, nasal peak flow and patients' satisfaction in septoplasty and inferior turbinectomy," Rhinology vol. 32, pp. 145–147, 1994.

[9] Hilberg O., "Objective measurement of nasal airway dimensions using acoustic rhinometry: methodological and clinical aspects," Allergy, vol. 57, suppl. 70, pp. 5-39, 2002

[10] Hilberg O., "Jackson AC, Swift Dl, and Pedersen OF. Acoustic rhinometry: Evaluation of nasal cavity geometry by acoustic reflection," Journal of Applied Physiology, vol. 66, pp. 295–303, 1989.

[11] Clement PA, and Gordts F, "Standardization committee on objective assessment of the nasal airway, IRS and ERS," Rhinology, vol. 43, pp. 169–179, 2005.

[12] Huang ZL, Wang DY, Zhang PC, Dong F, and Yeoh KH, "Evaluation of Nasal Cavity by Acoustic Rhinometry in Chinese, Malay and Indian Ethnic Groups," Acta Otolaryngol, vol. 121, pp. 844–848, 2001

[13] Alireza M, Mohammad F, and Artemis E, "Assessment of Nasal Volume and Cross-Sectional Area by Acoustic Rhinometry in a Sample of Normal Adult Iranians," Archives of Iranian Medicine, vol. 11, pp. 555 – 558, 2008

[14] Abe H., Yokoi H., Ohsaki M., Yamaguchi T., "Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining", Seventh IEEE International Conference on Data Mining - Workshops, pp. 127 – 132, 2007

[15] W. Lee, K. W. Mok, and S. J. Stolfo. Mining sequential patterns: Techniques, visualization, and applications. submitted for publication, 1998.

[16] W. W. Cohen, "Fast effective rule induction," Machine Learning: the 12th International Conference, Lake Taho, CA, 1995.

[17] JR. Quinlan, C4.5: Programs for Machine Learning.Morgan Kaufmann Publishers, 1993.

[18] JR. Quinlan, "Improved use of continuous attributes in c4.5," Journal of Artificial Intelligence Research, vol. 4, pp. 77-90, 1996.

[19] K.Q. Weinberger and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," The Journal of Machine Learning Research, vol. 10, pp. 207-244, 2009.

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA ,pp. 1137–1143 ,1995

[21] P Tantilipikorn, P Jareoncharsri, S Voraprayoon,C Bunnag, Clement, and Peter A., "Acoustic rhinometry of Asian noses," American Journal of Rhinology, Vol. 22, pp. (4)620-617, 2008

Worachartcheewan A,Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P and Prachayasittikul V, "Identification of metabolic syndrome using decision tree analysis," Diabetes Research and Clinical Practice, vol 90, pp. e15-e18, 2010

BIOGRAPHY

NAME	Mr. Wasin Srisawat
DATE OF BIRTH	18 December 1985
PLACE OF BIRTH	Tak, Thailand
INSTITUTIONS ATTENDED	University of the Thai Chamber of Commerce, 2009-2011 Bachelor of Engineering (Computer Engineering) Mahidol University, 2011-2013 Master of Science (Technology of Information System Management**)**
HOME ADDRESS	1083/1, Taksin Road, Muang District Tak 63000 Tel 083-219-9863 Email: tong.wasin@gmail.com
PUBLICATION / PRESENTATION	**Srisawat W.**, Kiattisin S., Leelasantitham A. , Wongseree W., Diagnose Abnormal Nasal based on the C4.5 Modeling using Cross Section Area Curve from Acoustic Rhinometry., Communications and Information Technologies (ISCIT), Thailand, 4-6 Sept. 2013