International Conference on e-Business 2007 (INCEB 2007)
22 - 23 November 2007
P33

Data Mining and Knowledge Discovery for Chronic Fatigue Syndrome Decease

Maliwan Boonploy and Nuanwan Soonthornphisaj

Department of Computer Science
Faculty of Science, Kasetsart University
E-mail: g4864200@ku.ac.th, fscinws@ku.ac.th

Abstract—Chronic Fatigue Syndrome (CFS) patients are unable to perform daily activities because of their unending fatigue and problems with short-term memory. Chronic Fatigue Syndrome is a state of continual fatigue that exists which can not be explained by other means. The CFS is complicated and difficult to diagnose. The objective of this study is to explore if double decision tree together with rule mining algorithm could improved the accuracy of CFS classifications. The dataset of this study are collected from 191 patients. Blood test for each patients and their diagnosis are collected. We proposed a new method called double decision tree. The experimental results obtained from our study shows that using data from blood test together with patient diagnosis can achieve 83.5% of accuracy which is better than those of the decision tree forest (1.43%.) Furthermore, the result of our study showed that the classification by double decision tree applied solely to the data from blood test was more effective than classification by decision tree forest 25.99%.

Keyword: Data Mining, Knowledge Discovery, Chronic Fatigue Syndrome, Decision Tree

1. Intorduction

A research [1] done by Kenneth J. Reynolds and his team indicated approximately between 400,000 to 800,000 American people have Chronic Fatigue Syndrome (CFS). The causes of this syndrome are the body abnormality or unique mind and related complexity which create fatigue condition, rheumatism on joints and muscular. Relaxation with sleep won't make the patient feel any better, this in turn worsen the working and social abilities or reduce the daily activities [2]. The Chronic Fatigue Syndrome can happen to both genders at all ages and levels. It is very likely that women may suffer more than men at an approximate ratio of 6:1 [3]. At the moment, it is still very difficult to find the cause of the Chronic Fatigue Syndrome including an appropriate method to diagnosis and identify the Chronic Fatigue Syndrome.

From the above reasons, this research will try to develop an algorithm which classify the relationship of all the data which are related to the Chronic Fatigue Syndrome to create body of knowledge on the syndrome. By utilizing the Decision tree approach, the information regarding to the blood test and the diagnosis of 191 patients will be used to create the required body of knowledge to help the personals in the medical field to diagnose and prevent the Chronic Fatigue Syndrome in the future. The data in this research can be classified into 3 classes, they are as follow:

1.1) Non-Fatigued (NF) class – normal people.
1.2) Insufficient Number of Symptoms or Fatigue Severity (ISF) class – patients who are missing some characteristic of the Chronic Fatigue Syndrome.
1.3) Chronic Fatigue Syndrome (CFS) class – the Chronic Fatigue Syndrome patients.

2. Related works

2.1 Theory

Classification Approach: This theory emphasis on learning from System Training Data to build data model for each data type. There are 3 types of learning, they are: the Supervised Learning, the Unsupervised Learning and the Semi-Supervised Learning [13].

Type classification using the Decision tree: This approach relies on the non-uniformed Entropy and Gain to select superior attribute(s) in classifying important data [4]. Related formula in calculating the non-uniformed data and gain are in Eq. 1 and 2 respectively:

Formula in calculating the non-uniformed Entropy(S)

$$\text{Entropy(S)} = \frac{-P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N} \quad (1)$$

where:

S = a typical set of system learning data
P = number of sets of positive system learning data
N = number of sets of negative system learning data

Formula in calculating the value of Gain.

$$\text{Gain(S,A)} = \text{Entropy(S)} - \sum_{v \in \text{Values(A)}} \frac{|S_v|}{|S|}\text{Entropy}(S_v) \quad (2)$$

where:

A = attribute data A
$|S_v|$ = number of classes with attribute A which has data v
$|S|$ = total number of classes found in system learning data set S
Entropy(S) = the non-uniformed value of system learning data set before the classification with attribute A
Entropy(S_v) = the non-uniformed value of system learning data set after the classification with the attribute A

The decision tree will show the result in term of tree structure which in turn can be changed to easy and understandable rules. Each node of the tree structure gives the attribute value. Each branch gives the test condition and the leaf node gives the designated data type. Fig. 1 is a typical result from using the decision tree.

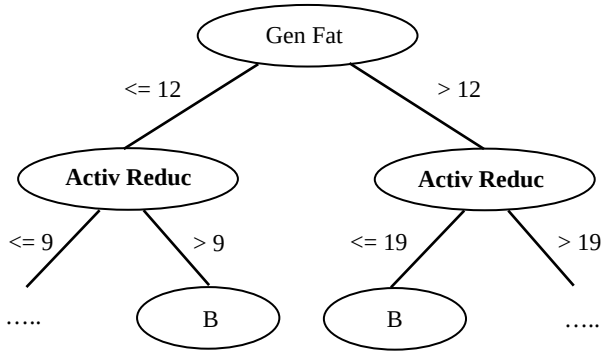

Fig. 1. Typical result from using the decision tree

The type classification using the decision tree not only helps in classifying the data in the medical field, it can also be used to predict a service model in the E-commerce [12], etc.

Type classification with forest: This is a collection of decision tree which can be used as predictor to create the required prediction for the forest. There is a format similar to the TreeBoost method which the number of trees keep on increasing. The increases may not in the same format and will not generate result until the final tree is created. The drawback of the forest classification type is the current format is very complex [14].

Type classification with forest has been used in the medical field to analyze the Single Nucleotide Polymorphisms (SNPs) [15].

2.2 Related researches

Today there are many researches which concentrated on the Serious Chronic Fatigue Syndrome. Most of them are studies on the impact of the Chronic Fatigue Syndrome on the patients in the economics area [1], the quantity of the blood circulation [5], the illness that follow the syndrome [6] and the issues that impact the cause of the Chronic Fatigue Syndrome [7]. The studies that involved the type classification of the Chronic Fatigue Syndrome are in the area which analyze the patient's Gene data [8], [9] to classify the data type of the Chronic Fatigue Syndrome. These classification have the accuracies of approximately 72.56%.

There are additional researches in finding new body of knowledge in the medical field. Most of the findings in the medical area use the decision tree and the Neural Network methods to analyze the required data for example, analyzing data involving the heart disease [10] and the Thalasimia disease [11], etc.

3. Research Methods

3.1 Typical data used in the research.

In this research, there are 2 types of patient, data they are:

1. the result of blood test data from the patients

2. the result of evaluation of the symptom data from the patient which consists of the illness evaluation value based on the doctor's diagnosis by dividing data into 3 main groups and 14 sub-groups as indicated in Table 1. There are 36 characteristics from blood test data as indicated in Table 2. and the patient's condition data consists of 83 characteristic data which can be grouped into 4 groups of General Health, Summary Scores from SF36, Summary Scores from Multidimensional Fatigue Inventory and CDC Symptom Inventory. The details of the data are in Table 3.

Table 1. Typical classification of the strong Chronic Fatigue Syndrome

Class names	number of patients
1. CFS (Chronic Fatigue Syndrome)	43
CFSMed	10
CFSPsy	1
CFSMedPsy	1
CFS-MDDm	12
2. ISF (Insufficient Number of Symptoms or Fatigue Severity)	61
ISFMed	9
ISFPsy	1
ISF-MDDm	8
ISF-MDDmMed	1
3. NF (Non-Fatigued)	59
NFMed	7
NFPsy	1
NF-MDDm	5

Table 2. Typical data characteristic of blood test from patients

No.	Characteristics
1.	Intake_Classific
2.	Empiric
3.	White Blood Cell (WBC)
4.	Red Blood Cell (RBC)
5.	Hemoglobin (HGB)
6.	Hemotocrits (HCT)
7.	Mean Corpuscular Volume (MCV)
8.	Mean Cell Hemoglobin (MCH)
9.	Mean Cell Hemoglobin Concentration (MCHC)
10.	Red Cell Distribution Width (RDW)
11.	Platelets Count (PLT)
12.	% granulocytes
13.	% lymphocytes
14.	% mononuclear cells
15.	% eosinophils
16.	% basophils
17.	# granulocytes
18.	# lymphocytes
19.	# mononuclear cells
20.	# eosinophils
21.	# basophils
22.	sodium
23.	potassium
24.	chloride
25.	CO2
26.	anion gap
27.	glucose
28.	Blood Urea Nitrogen (BUN)
29.	creatinine
30.	total protein
31.	albumin
32.	calcium
33.	Total Bilirubin (bili total)
34.	AST/SGOT
35.	ALT/SGPT
36.	Alkaline Phosphates (alk phos)

Table 3. Typical characteristics on patient's condition data

No.	Characteristic
1.	General Health consists of 13 characteristics.
2.	Summary Scores from SF36 consists of 8 characteristics.
3.	Summary Scores for Multidimensional Fatigue Inventory consists of 5 characteristics.
4.	CDC Symptom Inventory consists of 57 characteristics.

3.2 Research procedure

At this step, the researcher prepares the result of the blood test and result of the patient diagnosis to find the superior attribute(s) from the above results. This can be done by dividing the data into 2 classes, the Normal Class (NF) and the Abnormal Class (CFS, ISF). When the superior attribute(s) has been identified for each data set, a combine of the superior attribute(s) is then begin. The result will then feed to the decision tree to obtain the Normal Class and the Abnormal Class. For the data set in the Abnormal Class, 2 set of data will be obtained, they are blood test data set for both the CFS and ISF and the patient diagnosis data set for both the CFS and ISF. Both groups of data set will be evaluated to identify the superior attribute(s), combine the superior attribute(s) of both groups together and feed them to the decision tree to obtain the final 2 new groups, they are CFS and ISF groups. Fig. 2 is a typical steps of analyzing all the data sets.

A details of the above procedure can be divided 4 steps (from Fig. 2.) as follow:

3.2.1) Data preparation step:

Data preparation is necessary to eliminate any patient data set which some values are missing or any patient data set in which the patient took some medication before the blood test. Some medication may have impact on the result of the blood test which will create a fault result. After the above data preparation, the total number of valid data set is 191 data sets. These data sets are then classified into 3 main classes, they are: NF class which consists of 65 data sets of the NF, NFPsy and NF-MDDm; CFS class which consists of 56 data sets of the CFS, CFSsy and CFS-MDDm; ISF class which consists of 70 data sets of the ISF, ISFsy and the ISF-MDDm respectively.

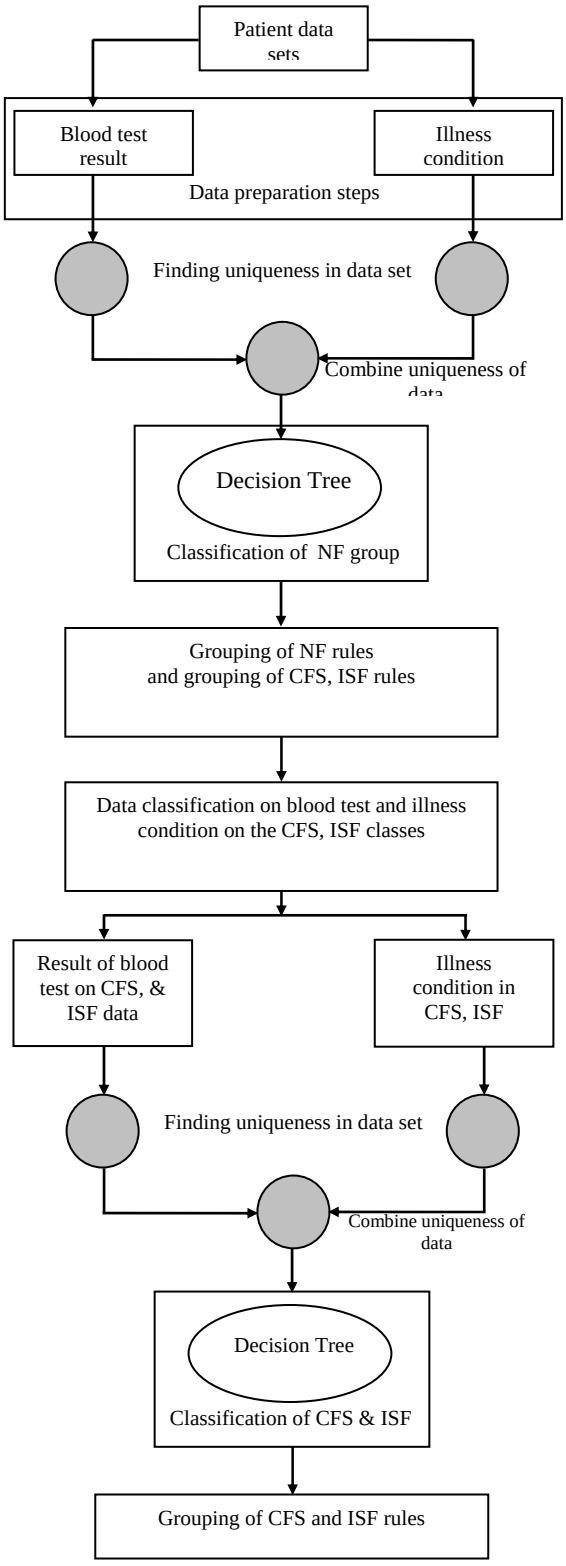

Fig. 2. Typical steps to analyze all data sets.

3.2.2) Finding the superior attribute(s) of data set step:

The researcher then proceed to identify the superior attribute(s) of the data set in the blood test and the patient diagnosis data set using the decision tree. The input data of Table 4. is the data set of the patient which consists of the Attribute.

Table 4. Typical step to identify the superior attribute(s) of the data set.

Typical step to identify the superior attribute(s) of the data set.
Input: D_{test}, D_{train}, $A_{train} = \{a_1, a_2, \ldots, a_n\}$ 1. *Model* = TrainDTree(D_{train} , A_{train}) 2. *acc_max* = TestDTree(Model , D_{test}) 3. For i = 1 to n 3.1 $A_{train} = A_{train} - \{a_i\}$ 3.2 *Model* = TrainDTree(D_{train} , A_{train}) 3.3 *Acc* = TestDTree(Model , D_{test}) 3.4 IF *acc_max* > *acc* Then $A_{train} = A_{train} \cup \{a_i\}$ 3.5 IF *acc_max* < *acc* Then *Acc_max* = *acc* 3.6 i = i +1 End For Output: A_{train}

Table 4. is a typical steps to identify the superior attribute(s) of the data set using the decision tree. From the test, the results from identifying the superior attribute(s) from the patient data set is divided into 2 groups as follow:

1. Step to classify the Non-Fatigued (NF): From the blood test data set, 25 superior attribute(s) out of 36 characteristic from each data set can be identified. For the patient diagnosis data sets, 2 superior attribute(s) out of 83 characteristic can also be identified.

2. Step to classify the Fatigued (CFS and ISF): From the blood test data sets, 10 superior attribute(s) out of 36 characteristic from each data set can be identified. For the patient siagnosis data, 4 superior attribute(s) out of 83 characteristic can also be identified.

3.2.3) Classification steps for the NF group:

To classify the NF group, the data will be divided into 2 classes, they are: the Non-Fatigued class which consists of the NF, NFPsy and NF-DMMm data sets, the Fatigued class which consists of the CFS, CFSPsy, CFS-MDDm, ISF, ISFPsy and ISF-MDDm data sets. Next step is to identify the superior attribute(s) of data in the blood test and the patient diagnosis data sets. When the superior attribute(s) of the blood test group and the patient diagnosis group have been identified, a combine of

each superior attribute(s) from both groups is required. The result of the combination is then fed to the decision tree which utilized the K-Fold Cross Validation algorithm. After the decision tree process, 2 groups of the NF and the CFS, ISF rules can be obtained. At this stage, classification of the NF and the CFS, ISF can be done using the above rules as those of Table 5.

Table 5. Classification steps for the NF group

Classification steps for the NF group
Input: $Blood_{train}$, $Symptom_{train}$, $Blood_{test}$, $Symptom_{test}$ 1. $BloodData_{train}$ = FeatureAttribute($Blood_{train}$) 2. $SymptomData_{train}$ = FeatureAttribute($Symptom_{train}$) 3. $BloodSym_{train}$ = FusionData($BloodData_{train}$, $SymptomData_{train}$) 4. $BloodData_{test}$= FeatureAttribute($Blood_{test}$) 5. $SymptomData_{test}$ = FeatureAttribute($Symptom_{test}$) 6. $BloodSym_{test}$ = FusionData($BloodData_{test}$, $SymptomData_{test}$) 7. $Model_{train}$ = TrainDecisionTree ($BloodSym_{train}$) 8. *Rule* = TestDecisionTree ($Model_{train}$, $BloodSym_{test}$) Output : Rules Set

3.2.4) Classification steps for the CFS and ISF Group:

This step will classify the patients in the CFS class or those who miss some of the characteristic in the ISF class. This step can be divided into 2 smaller steps, they are: sub-step to select the NF data set and sub-step to classify those of the CFS and ISF group. The steps are shown in Table 6. and 7. respectively as follow:

Table 6. Classification sub-step selection for the NF

Sub-step selection for the NF
Input: Patient = $\{P_1, P_2, \ldots, P_n\}$ Attribute = $\{A_1, A_2, \ldots, A_n\}$ 1. For i = 1 to n 1.1 *valEmp* = checkClass(P_i) 1.2 IF *valEmp* = "NF" Then Patient = Patient - P_i 1.3 i = i +1 End For Output: Data

Table 7. Classification sub-step for the CFS and ISF

Classification sub-step for the CFS and ISF
Input: $Blood_{train}$, $Symptom_{train}$, $Blood_{test}$, $Symptom_{test}$ 1. $BloodCI_{train}$ = CutRecord($Blood_{train}$) 2. $SymCI_{train}$ = CutRecord($Symptom_{train}$) 3. $FeBloodCI_{train}$ = FeatureAttribute($BloodCI_{train}$) 4. $FeSymCI_{train}$ = FeatureAttribute($SymCI_{train}$) 5. $BloodSymCI_{train}$ = FusionData($FeBloodCI_{train}$, $FeSymCI_{train}$) 6. $BloodCI_{test}$ = CutRecord($Blood_{test}$) 7. $SymCI_{test}$ = CutRecord($Symptom_{test}$) 8. $FeBloodCI_{test}$ = FeatureAttribute($BloodCI_{test}$) 9. $FeSymCI_{test}$ = FeatureAttribute($SymCI_{test}$) 10. $BloodSymCI_{test}$ = FusionData($FeBloodCI_{test}$, $FeSymCI_{test}$) 11. $Model_{train}$ = TrainDecisionTree ($BloodSymCI_{train}$) 12. *Rule* = TestDecisionTree ($Model_{train}$, $BloodSymCI_{test}$) Output : *Rule*

3.3 Steps to measure efficiency

To measure the correctness of the efficiency in classifying the CFS data set, the researcher uses the following formula:

$$\text{Sensitivity} = \frac{TP}{P} \quad (3)$$

$$\text{Specificity} = \frac{TN}{N} \quad (4)$$

$$\text{Accuracy} = \text{Sensitivity}\left[\frac{P}{P+N}\right] + \text{Specificity}\left[\frac{N}{P+N}\right] \quad (5)$$

where:

True Positives (TP) = Total number of patients being classify from the data set as real patients and they are correct.

True Negatives (TN) = Total number of patients being classify from data set as non-real patients and they are correct.

Positives (P) = Total number of real patients

Negatives (N) = Total number of non-real patients

This research utilized the principle of the K-Fold Cross Validation algorithm to manipulate the input data by assigning the value of K as 10. The 9 portion of data sets will be allocated as Training Data and the last portion will be allocated as Testing Data data sets. Then all the data sets will be regrouped and the above process will be repeated 10 times to obtain the final result.

4. Result of Research

In this research, an efficiency analysis will be performed in classifying the blood test and the patient diagnosis data sets. Additional tests were run for comparison purposes, the first test utilizing the double decision tree on the blood test data set only

and the blood test data sets along with the patient diagnosis data sets. The second test utilizing the tree forest Trees on blood test data sets only and the blood test data sets along with the patient diagnosis data sets. The following are summary for the above comparison tests:

Table 8. Test result utilizing the double decision tree using only the blood test data sets.

BloodTest	NF	ISF	CFS
Sensitivity	69.99	58.57	54.29
Specificity	61.43	61.54	63.85
Accuracy	60.41	60.45	60.59

From Table 8, the classification of data sets in the CFS class results in the best efficiency when viewing the accuracy value of 60.59%. The ISF class follows with the accuracy value of 60.45%.

Table 9. Test result utilizing the double decision tree on blood test data sets along with the patient diagnosis data sets.

BloodTest + Symptom	NF	ISF	CFS
Sensitivity	100	75.712	77.14
Specificity	76.43	87.69	86.92
Accuracy	83.6	83.30	83.34

From Table 9, the classification of data sets in the NF class results in the best efficiency when viewing the accuracy and sensitivity in the NF class which has higher value than other classes. The CFS class follows with the accuracy value of 83.34%.

Table 10. Test result utilizing the decision tree forest using only the blood test data sets.

BloodTest	NF	ISF	CFS
Sensitivity	36.92	37.14	28.57
Specificity	33.33	33.06	37.04
Accuracy	34.43	34.55	34.5

From Table 10, the classification of data sets in the ISF class results in the best efficiency when viewing the accuracy in the ISF class of 34.55%. The CFS class follows with the accuracy value of 34.5%.

Table 11. Test result utilizing the decision tree forest on the blood test data sets along with the patient diagnosis data sets.

BloodTest + Symptom	NF	ISF	CFS
Sensitivity	100	75.71	69.64
Specificity	73.02	85.96	87.19
Accuracy	81.94	82.2	81.79

From Table 11, the classification of data sets in the ISF class results in the best efficiency when viewing the accuracy value of 82.2%. The NF class follows with the accuracy value of 81.94%.

Table 12. Comparison on efficiency on all data sets.

Method	Overall efficiency for all 3 groups		
	Sensitivity	Specificity	Accuracy
double decision tree using only the blood test data sets.	60.95	62.27	60.48
double decision tree on the blood test data sets along with the patient diagnosis data sets.	84.28	83.68	83.41
decision tree forest using only the blood test data sets.	34.21	34.38	34.49
decision tree forest using the blood test data sets and the patient diagnosis data sets.	81.78	82.06	81.98

From Table 12, the double decision tree provides better efficiency than those of the decision tree forest on both the only blood test data sets and the blood test data sets along with the patient diagnosis data sets. The double decision tree results in the highest accuracy of 83.41% on the blood test along with the patient diagnosis data sets while the decision tree forest provides only 81.98% on the same data sets. The result is the same for the double decision tree in case of using only the blood test data sets which provides an accuracy of 60.48% when compare to the 34.49% for the decision tree forest.

5. Conclusion

The concept of using the double decision tree to classify the ISF type provides a better efficiency than those of the decision tree forest because the former method gives better accuracy than those of the decision tree forest. The double decision tree has divided the classification into 2 steps utilizing the data set selection on the patient data sets, this in turn reduces the complexity of the problem in this

research from the Multi-class Problem into an easier Two-class Problem. From the test result, the double decision tree can classify the input data sets with better efficiency. The values of class Sensitivity of 77.14% and the Accuracy of 83.34% on the patient in the CFS class from the double decision tree are higher than those of the decision tree forest at 7.5% and 1.55% respectively.

6. References

[1] Kenneth J Reynolds, Suzanne D Vernon, Ellen Bouchery and William C Reeves "The Economic Impact of Chronic Fatigue Syndrome" *Cost Effectiveness and Resource Allocation*, 2004.

[2] Rosane Nisenbaum, James F Jones, Elizabeth R Unger, Michele Reyes and William C Reeves, "A Population-Based Study of the Clinical Course of Chronic Fatigue Syndrome" *Health and Quality of Life*, 2003.

[3] L.D. Devanur, J.R. Kerr, "Review Chronic Fatigue Syndrome" *Journal of Clinical Virology*, vol. 37, pp. 139-150, August 2006.

[4] J.R. Quinlan, "Induction of Decision Trees" *Machine Learning*, vol. 1, pp. 81-106, 1986.

[5] David H. P. Streeten, David S. Bell, "Circulating Blood Volume in Chronic Fatigue Syndrome" *Journal of Chronic Fatigue Syndrome*, vol. 4, pp. 3-11, 1998.

[6] James F. Jones, Rosane Nisenbaum, Laura Solomon, Michele Reyes, and William C. Reeves, "Chronic Fatigue Syndrome and Other Fatiguing Illnesses in Adolescents: A Population-based Study" *Journal of Adolescent Health*, vol. 35, pp. 34-40, 2004.

[7] Laura Solomon and William C. Reeves, "Factors Influencing the Diagnosis of Chronic Fatigue Syndrome" *ARCH INTERN MED*, vol. 164, pp. 2241-2245, 2004.

[8] Toni Whistler, Elizabeth R Unger, Rosane Nisenbaum and Suzanne D Vernon, "Integration of gene expression, clinical, and epidemiologic data to characterize Chronic Fatigue Syndrome" *Journal of Translational Medicine*, 2003.

[9] Toni Whistler, Elizabeth R Unger, Rosane Nisenbaum and Suzanne D Vernon, "Exercise responsive genes measured in peripheral blood of women with Chronic Fatigue Syndrome and matched control subjects" *BMC Physiology*, 2005.

[10] Andrew Kusiak, Christopher A. Caldarone, Michael D. Kelleher, Fred S. Lamb, Thomas J. Persoon and Alex Burns, "Hypoplastic left heart syndrome: knowledge discovery with a data mining approach" *Computers in Biology and Medicine*, vol. 36, pp. 21-40, 2006.

[11] Waranyu Wongseree, Nachol Chaiyaratana, Kanjana Vichittumaros, Pranee Winichagoon and Suthat Fucharoen, "Thalassaemia classification by neural networks and genetic programming" *Information Sciences*, vol. 117, pp. 771-786, 2007.

[12] Sungjoo Lee, Seunghoon Lee and Yongtae Park, "A prediction model for success of services in e-commerce using decision tree: E-customer's attitude towards online service" *Expert Systems with Applications*, vol. 33, pp. 572-581, 2007.

[13] Mitchell T.M., "Machine Learning." *McGraw-Hill*, New York

[14] Breiman Leo, "Decision Tree Forests" *Machine Learning*, 5-32, 2001

[15] Qian Xie, Luke D Ratnasinghe, Huixiao Hong, Roger Perkins, Ze-Zhong Tang, Nan Hu, Philip R Taylor and Weida Tong, "Decision Forest Analysis of 61 Single Nucleotide Polymorphisms in a Case-Control Study of Esophageal Cancer; a novel method", *BMC Bioinformatics* , 2005