Evaluation of Symptoms in Heart Disease Patients by using k - Nearest Neighbor Classification

Hnin Yu Maw; Khin Sandar, May Phyo Oo
Computer University(Pathein)
hninyu155@gmail.com , drkhinsandar@gmail.com, mayphyooo@gmail.com

Abstract

In many application domains, classification of complex measurements is essential in a diagnosis process. Correct classification of measurements may in fact be the most critical part of the diagnostic process. In this system, we intend to determine whether a patient has coronary artery disease (CAD) or not and if we have heart disease (CAD) what stage is it by using k - nearest neighbor classification. The k -nearest neighbor (k NN) is a sample and widely used technique which has found in several applications on classification problem. We can get classification accuracy by using k - nearest neighbor algorithm. Experiments were evaluated on some public datasets collected from the Cleveland Clinic Foundation in the UCI (University of California, Irvine) machine learning repository in order to test this system.

Keywords

Machine learning, k - nearest neighbor classifier, classifier accuracy, k nearest neighbor algorithm, Coronary Artery Disease.

1. Introduction

Coronary artery disease is the most common form of heart disease. It occurs when the arteries that supply blood to the heart muscles (coronary arteries) become hardened and narrowed. Every year, millions of deaths worldwide are attributed to CAD (coronary artery disease). Therefore, finding cost effective methods to predict and control CAD is one of the greatest challenges in public health. Research in this area usually involve using medical profile and family history information to predict the risk for CAD. Along with the rapid advancement of biomedical technologies, now we can evaluate with k nearest neighbor to predict the risk of CAD, and achieved reasonably good results. Patients with coronary artery disease (CAD) often show no symptoms before experiencing myocardial infarction (heart attack). Approximately 50% of men and 63% of women who die suddenly from CAD show no previous symptoms of the disease .Many different results, obtained from tests with substantial imperfections, must be integrated into a diagnostic conclusion about the probability of disease in a given patient. If we don’t know the symptoms, we could be at risk and not even it. In this paper, we will evaluate the classification using heart disease data sets from the machine learning database repository at the University of California, Irvine (UCI).

This paper is organized as follows. The next section reviews some of the related works. In the section 3, we present the process flow of the system and k – nearest neighbor classification. In section 4, we describe the data sets and their attribute information. In section 4 and 5, we review the classifier accuracy of the system and finally conclude the article.

2. Related work

The research in medical k nearest neighbor (kNN) is concentrated to Europe and US, as with in general. The medical domain of kNN is generally focused on producing systems for specific tasks, such as diagnosing a specific symptom. This system use the kNN classifier as a classification algorithm. It provides good generalization accuracy on many applications. Despite its simplicity, it has many advantages, e.g. it does not require any knowledge about statistical properties of data beforehand, it may competitive performance compare to many other methods. The nearest neighbor stated that any distance function can be used to determine how “close” one instance is to another, but only Euclidean distance was used in their experiments. A variety of other distance functions proposed, by far the most commonly used is the Euclidean Distance function. The predicted accuracy is mostly depending on the distance function[7].

3. Proposed System

Many classification and prediction methods are proposed by researchers in machine learning, expert systems, statistics and neurobiology. Some other approaches of classification, such as Case _ based reasoning, genetic algorithm, rough set and fuzzy logic techniques are introduced. Nearest Neighbor (NN) classifier is the most simple classifier found up till now. In NN classifier no special procedure is required to do training. All the available data (as maximum as possible) is stored to perform classification, where each test pattern is compared for similarity with all the available training data (pattern). The test pattern is assigned the class label of that training pattern, which is the closest to the test pattern . A major drawback of NN approach is its large total parameter requirement to perform classification task. When train patterns and test patterns are closely matched then accuracy obtained by NN approach is good. But when the test patterns do not match with train patterns, NN approach provides poor performance (in terms of accuracy). The classification accuracy of NN approach can be improved by making the decision of a test pattern for class labeling based on *k* nearest patterns. This method is known as *k*-Nearest Neighbor (kNN) technique.

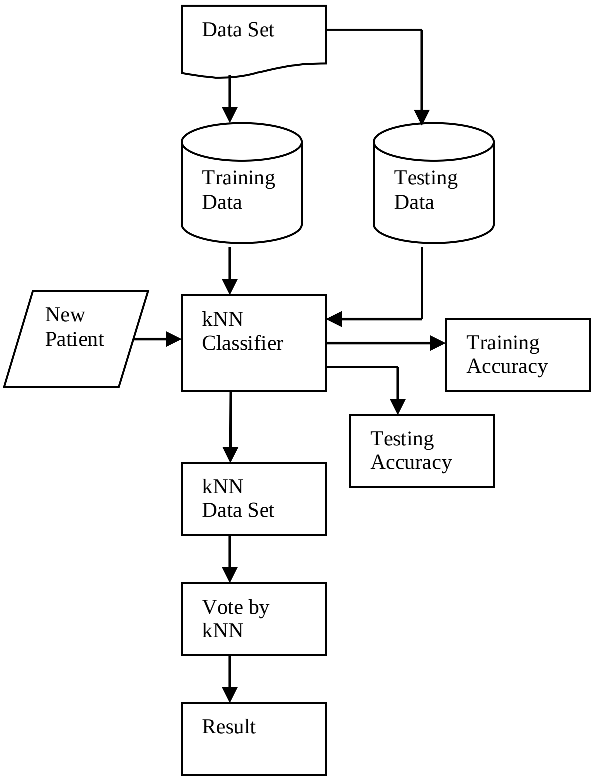


Figure 1. Process Flow Diagram

3.1 K - Nearest Neighbor Classification

The basic concept underlying the nearest neighbor classifier was first introduced by Fix and Hodges. In 1967, Cover and Hart formally defined the nearest neighbor rule and applied it to the pattern recognition problem. Since then, the nearest neighbor algorithm had been under extensive study. k - nearest neighbor was a generalization of metric distance minimization, instead of returning the parameters of the point that was most similar to the query point .

The popularity of kNN was due to its simplicity and the fact that the underlying class-conditional density need not be known. This was often the case with other methods. kNN is well known and has shown to give good performance on real-world data sets. It is competitive with other methods.

In machine learning community, kNN was often called instance based learning or memory based learning. A classifier predicts the class of a query pattern. kNN matches the query pattern against already classified examples. The k nearest examples forms a subset. The query pattern's class was predicted as the class that occurs most frequently in that subset. kNN unlike other common classifiers it does not build a classifier in advance .

There were four algorithmic parameters associated with the kNN rule: the value of k, the choice of distance measure, the distance weighting measure (weighted or no weighted) and the method of counting votes. Many distance metrics had been proposed in the literature but Euclidean distance is the most commonly used.

4. Data Sets Description

This data set contains information concerning heart disease diagnosis. The data was collected from the Cleveland Clinic Foundation and it is available from the UCI machine learning repository at www.ics.uci.edu/~mlearn/ML.Repository.html. The goal field refer to the presence of heart disease in the patient .Experiment with the Cleveland data set have concentrated on simply attempting to distinguish sick (values 1,2,3,4) from healthy (value 0).Number of instance: 303.Number of attributes: 14 (including class attributes).Attribute information and sample data set of this system is described in section 4.1 and 4.2.

4.1 Attribute information

Attributes; 8 symbolic, 6 numeric

1.Age in years (integer)

2.Sex (male, female)

3.Chest pain type

value 1: typical angina (angina)

value 2:atypical angina (abnong)

value 3:non_anginal pain (notang)

value 4:asypmtomatic (asympt)

4.Resting blood pressure (integer)

5.Serum colestoral in mg/dl (integer)

6. Fasting blood sugar < 120 (true or false)

7. Resting electrocardiographic results

value 0: normal (norm)

value 1: having ST_T wave abnormality (abn)

value 2: showing probable or definite left ventricular hypertrophy by Estes'criteria (hyper)

8. Maximum heart rate achieved (real)

9. Exercise induced angina (true, false)

10. Oldpeak: ST depression induced by exercise relative to rest

11. The slope of the peak exercise ST segment

value 1: upsloping (up)

value 2: flat (flat)

value 3: downsloping (down)

12. Number of major vessels (0_3) colored by flourosopy (real)

13. Thallium scintigraphy :

normal (norm)

fixed defect (fixed)

reversable defect (rever)

14. Class attribute: healthy (buff) (H)

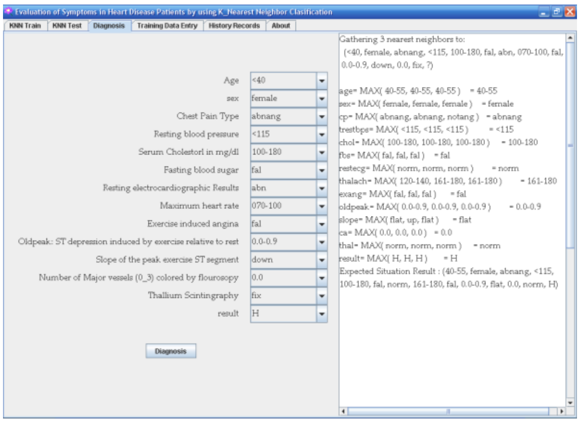
heart disease (sick) (S1,S2,S3,S4)

4.2 Sample Data Set

There are 303 sample data set in this system. We present six of them in this section.

55-65,male,angina,141-150,181-240,true,hyp,141-160,fal,2.0-2.9,down,0.0,fix,H
>65,male,asympt,151-200,240-300,fal,hyp,100-120,true,1.0-1.9,flat,3.0,norm,S2
>65,male,asympt,115-125,100-180,fal,hyp,120-140,true,2.0-2.9,flat,2.0,rev,S1
<40,male,notang,126-130,240-300,fal,norm,181-202,fal,3.0-3.9,down,0.0,norm,H
40-55,female,abnang,126-130,181-240,fal,hyp,161-180,fal,1.0-1.9,up,0.0,norm,H
55-65,male,abnang,115-125,181-240,fal,norm,161-180,fal,0.0-0.9,up,0.0,norm,H

4.3 Implementation of the System


Evaluation of Symptoms in Heart Disease Patients by using K_Nearest Neighbor Clasification
KNN Train
KNN Test
Diagnosis
Training Data Entry
History Records
About
Diagnosis

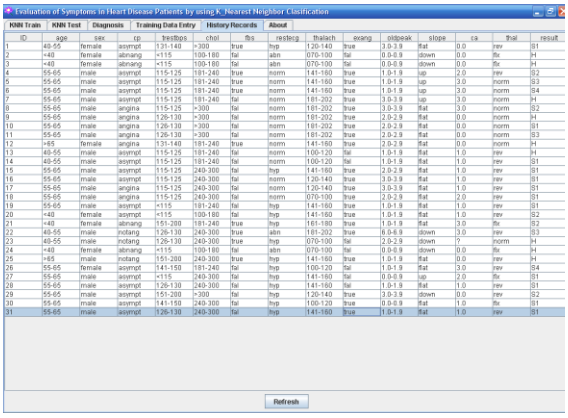
Age <40
sex female
Chest Pain Type abnang
Resting blood pressure <115
Serum Cholestorl in mg/dl 100-180
Fasting blood sugar fal
Resting electrocardiographic Results abn
Maximum heart rate 070-100
Exercise induced angina fal
Oldpeak: ST depression induced by exercise relative to rest 0.0-0.9
Slope of the peak exercise ST segment down
Number of Major vessels (0_3) colored by flourosopy 0.0
Thallium Scintingraphy fix
result H

```
Gathering 3 nearest neighbors to:
 (<40, female, abnang, <115, 100-180, fal, abn, 070-100, fal,
 0.0-0.9, down, 0.0, fix, ?)

age= MAX( 40-55, 40-55, 40-55 )    = 40-55
sex= MAX( female, female, female )    = female
cp= MAX( abnang, abnang, notang )    = abnang
trestbps= MAX( <115, <115, <115 )         = <115
chol= MAX( 100-180, 100-180, 100-180 )     = 100-180
fbs= MAX( fal, fal, fal )    = fal
restecg= MAX( norm, norm, norm )        = norm
thalach= MAX( 120-140, 161-180, 161-180 )        = 161-180
exang= MAX( fal, fal, fal )     = fal
oldpeak= MAX( 0.0-0.9, 0.0-0.9, 0.0-0.9 )       = 0.0-0.9
slope= MAX( flat, up, flat )     = flat
ca= MAX( 0.0, 0.0, 0.0 )   = 0.0
thal= MAX( norm, norm, norm )     = norm
result= MAX( H, H, H )      = H
Expected Situation Result : (40-55, female, abnang, <115,
100-180, fal, norm, 161-180, fal, 0.0-0.9, flat, 0.0, norm, H)
```

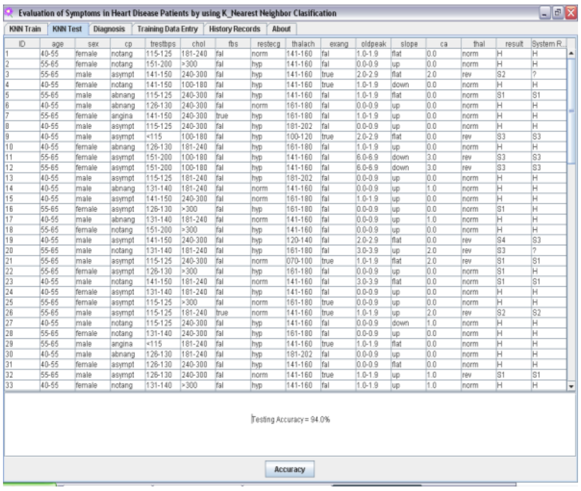
Figure 2. Diagnosis and Testing Result generated from the k-Nearest Neighbor analysis

Figure2 show diagnosis section and testing result generated from the kNN analysis for the system. When we fill the information needed by the system, we can see the result of the patient in Figure 3. This evaluation for heart disease classification is obtained by using 13 attributes with different attribute values and 5 classes for various kind of result.


Evaluation of Symptoms in Heart Disease Patients by using K_Nearest Neighbor Clasification
KNN Train
KNN Test
Diagnosis
Training Data Entry
History Records
About
Refresh

ID	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	result
1	40-55	female	asympt	131-140	>300	true	hyp	120-140	true	3.0-3.9	flat	0.0	rev	S1
2	<40	female	abnang	<115	100-180	fal	abn	070-100	fal	0.0-0.9	down	0.0	fix	H
3	<40	female	abnang	<115	100-180	fal	abn	070-100	fal	0.0-0.9	down	0.0	fix	H
4	55-65	male	asympt	115-125	181-240	true	norm	141-160	true	1.0-1.9	up	2.0	rev	S2
5	55-65	male	asympt	115-125	181-240	true	norm	141-160	true	1.0-1.9	up	3.0	norm	S3
6	55-65	male	asympt	115-125	181-240	true	norm	141-160	true	1.0-1.9	up	3.0	norm	S4
7	55-65	male	asympt	115-125	181-240	fal	norm	181-202	true	3.0-3.9	up	3.0	norm	H
8	55-65	male	angina	115-125	>300	fal	norm	181-202	true	3.0-3.9	flat	3.0	norm	S2
9	55-65	male	angina	126-130	>300	fal	norm	181-202	true	2.0-2.9	flat	0.0	norm	H
10	55-65	male	angina	126-130	>300	fal	norm	181-202	true	2.0-2.9	flat	0.0	norm	S1
11	55-65	male	angina	126-130	>300	fal	norm	181-202	true	2.0-2.9	flat	0.0	norm	S3
12	>65	female	angina	131-140	181-240	true	norm	141-160	true	2.0-2.9	flat	0.0	norm	H
13	40-55	male	asympt	115-125	181-240	fal	norm	100-120	fal	1.0-1.9	flat	1.0	rev	H
14	40-55	male	asympt	115-125	181-240	fal	norm	100-120	fal	1.0-1.9	flat	1.0	rev	S1
15	55-65	male	asympt	115-125	240-300	fal	hyp	141-160	true	2.0-2.9	flat	1.0	rev	S1
16	55-65	male	asympt	115-125	240-300	fal	norm	120-140	true	3.0-3.9	flat	1.0	rev	S1
17	55-65	male	angina	115-125	240-300	fal	norm	120-140	true	3.0-3.9	flat	1.0	rev	S1
18	55-65	male	angina	115-125	240-300	fal	norm	070-100	true	2.0-2.9	flat	2.0	rev	S1
19	55-65	male	asympt	<115	181-240	fal	hyp	141-160	true	1.0-1.9	flat	1.0	rev	S1
20	<40	female	asympt	<115	100-180	fal	hyp	141-160	true	1.0-1.9	flat	1.0	rev	S2
21	<40	female	abnang	151-200	181-240	true	hyp	161-180	true	1.0-1.9	flat	3.0	fix	S2
22	40-55	male	notang	126-130	240-300	true	abn	181-202	true	6.0-6.9	down	3.0	rev	S3
23	40-55	male	notang	126-130	240-300	true	hyp	070-100	fal	2.0-2.9	down	?	norm	H
24	<40	female	abnang	<115	100-180	fal	abn	070-100	fal	0.0-0.9	down	0.0	fix	H
25	>65	male	notang	151-200	240-300	true	hyp	141-160	true	1.0-1.9	flat	0.0	rev	H
26	55-65	female	asympt	141-150	181-240	fal	hyp	100-120	fal	1.0-1.9	flat	3.0	rev	S4
27	55-65	male	asympt	<115	240-300	fal	hyp	141-160	fal	0.0-0.9	up	2.0	fix	S1
28	55-65	male	asympt	126-130	240-300	fal	hyp	141-160	fal	1.0-1.9	flat	1.0	rev	S1
29	55-65	male	asympt	151-200	>300	fal	hyp	120-140	true	3.0-3.9	down	0.0	rev	S2
30	55-65	male	asympt	141-150	240-300	fal	hyp	100-120	true	0.0-0.9	flat	1.0	fix	S1
31	55-65	male	asympt	126-130	240-300	fal	hyp	141-160	true	1.0-1.9	flat	1.0	rev	S1

Figure3. History Records


Evaluation of Symptoms in Heart Disease Patients by using K_Nearest Neighbor Clasification
KNN Train
KNN Test
Diagnosis
Training Data Entry
History Records
About
Accuracy

ID	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	result	System R...
1	40-55	female	notang	115-125	181-240	fal	norm	141-160	fal	1.0-1.9	flat	0.0	norm	H	H
2	55-65	female	notang	151-200	>300	fal	hyp	141-160	fal	0.0-0.9	up	0.0	norm	H	H
3	55-65	male	asympt	141-150	240-300	fal	hyp	141-160	true	2.0-2.9	flat	2.0	rev	S2	?
4	40-55	female	notang	141-150	100-180	fal	hyp	141-160	true	1.0-1.9	down	0.0	norm	H	H
5	55-65	male	abnang	115-125	240-300	fal	hyp	141-160	fal	1.0-1.9	flat	0.0	norm	S1	S1
6	40-55	male	abnang	126-130	240-300	fal	norm	161-180	fal	0.0-0.9	up	0.0	norm	H	H
7	55-65	female	angina	141-150	240-300	true	hyp	161-180	fal	1.0-1.9	up	0.0	norm	H	H
8	40-55	male	asympt	115-125	240-300	fal	hyp	181-202	fal	0.0-0.9	up	0.0	norm	H	H
9	40-55	male	asympt	<115	100-180	fal	hyp	100-120	true	2.0-2.9	flat	0.0	rev	S3	S3
10	40-55	female	abnang	126-130	181-240	fal	hyp	161-180	fal	1.0-1.9	up	0.0	norm	H	H
11	55-65	female	asympt	151-200	100-180	fal	hyp	141-160	fal	6.0-6.9	down	3.0	rev	S3	S3
12	55-65	female	asympt	151-200	100-180	fal	hyp	141-160	fal	6.0-6.9	down	3.0	rev	S3	S3
13	40-55	male	asympt	115-125	181-240	fal	hyp	181-202	fal	0.0-0.9	up	0.0	norm	H	H
14	40-55	male	abnang	131-140	181-240	fal	norm	141-160	fal	0.0-0.9	up	1.0	norm	H	H
15	40-55	male	asympt	141-150	240-300	fal	norm	161-180	fal	1.0-1.9	up	0.0	norm	H	H
16	55-65	female	asympt	126-130	>300	fal	hyp	161-180	fal	0.0-0.9	up	0.0	norm	S1	H
17	40-55	male	abnang	131-140	181-240	fal	norm	141-160	fal	0.0-0.9	up	1.0	norm	H	H
18	55-65	female	notang	151-200	>300	fal	hyp	141-160	fal	0.0-0.9	up	0.0	norm	H	H
19	40-55	male	asympt	141-150	240-300	fal	hyp	120-140	fal	2.0-2.9	flat	0.0	rev	S4	S3
20	55-65	male	notang	131-140	181-240	fal	hyp	161-180	fal	3.0-3.9	up	2.0	rev	S3	?
21	55-65	male	asympt	115-125	240-300	fal	norm	070-100	true	1.0-1.9	flat	2.0	rev	S1	S1
22	55-65	female	asympt	126-130	>300	fal	hyp	161-180	fal	0.0-0.9	up	0.0	norm	S1	H
23	40-55	male	notang	141-150	181-240	fal	norm	141-160	fal	3.0-3.9	flat	0.0	norm	S1	S1
24	40-55	female	asympt	131-140	181-240	fal	hyp	141-160	fal	0.0-0.9	up	0.0	norm	H	H
25	55-65	female	asympt	115-125	>300	fal	norm	161-180	true	0.0-0.9	up	0.0	norm	H	H
26	55-65	male	asympt	115-125	181-240	true	norm	141-160	true	1.0-1.9	up	2.0	rev	S2	S2
27	40-55	male	notang	115-125	240-300	fal	hyp	141-160	fal	0.0-0.9	down	1.0	norm	H	H
28	55-65	female	notang	131-140	240-300	fal	hyp	161-180	fal	0.0-0.9	up	0.0	norm	H	H
29	55-65	male	angina	<115	181-240	fal	hyp	141-160	true	1.0-1.9	flat	0.0	norm	H	H
30	40-55	male	abnang	126-130	181-240	fal	hyp	181-202	fal	0.0-0.9	up	0.0	norm	H	H
31	40-55	female	asympt	126-130	240-300	fal	hyp	141-160	fal	0.0-0.9	flat	0.0	norm	H	H
32	55-65	male	asympt	126-130	240-300	fal	norm	141-160	true	1.0-1.9	up	1.0	rev	S1	S1
33	40-55	female	notang	131-140	>300	fal	hyp	141-160	fal	1.0-1.9	up	1.0	norm	H	H

Testing Accuracy = 94.0%

Figure 4. The result of testing accuracy

5. Classifier Accuracy

Classification algorithms can then be compared according to their accuracy. Accuracy is measured using a test set of objects for which the class label are known. Accuracy is decided as the number of correct class predictions, divided by the total number of test samples. For medical applications, two other measures are more frequently used than the classification accuracy, sensitivity and specificity .Sensitivity measures the fraction of positive cases that are classified as positive. Specificity measure the fraction of negative cases classified as negative.

.

Sensitivity = t_pos/pos,
Specificity= t_neg/neg,
Precision = t-pos / (t-pos+f-pos)
accuracy = sensitivity (pos / (pos+neg)) +specificity (neg / (pos+neg))

t_pos = the number of true positives
pos = the number of positive
t_neg = the number of true negatives
neg = the number of negative
f_pos = the number of false positive

There are 303 record in this system.202 data sets have been trained and 101 records are use as testing data. There are several methods of accuracy determination to assess the performance of k – nearest neighbor classification. We use one of them named hold-out method. When data sets of this system are trained, 94% of testing accuracy has been achieved according to figure 4.

6. Conclusion

A system has been developed to evaluate symptoms in heart disease patient using k – nearest neighbor classification. The Nearest neighbor algorithm is based on distance function, the attributes values should be such that a distance could be computed. Because the distance between distances is based on all attributes, less relevant attribute can delay processing time. In our future by define an appropriate distance function especially when samples are represented as complex symbolic expressions. This system can make the most possible decision instead of a doctor.

References

[1] E. E. Khing (MCSc) Dec, 2008,"Spatial Data Mining by using Nearest Neighbor Algorithm". University of Computer Study,Yangon.

[2] J. Han and M. Kamber, "Data Mining Concepts and Techniques", ACADEMIC.2001.

[3] K.T..Lynn ,"Case_Based Reasoning for Evaluation of Symptom in Bronchogenic Carcinoma Patients" University of Computer Studies, Yangon, Myanmar.

[4] N. Lavric, E. Keravnou, B. Zupan, "Intelligent Data Analaysis in Medicine".

[5] G. D. Magoulas and A. Prentza, "Machine Learning in Medical Applications".

[6] National Institute for Clinical Excellence. Chronic heart failure: Management of chronic heart failure in adults in primary and secondary care. July 2003. Clinical guideline 5. Available at http://www.nice.org.uk (accessed on 14 January 2008).

[7] S. Phyu University of Computer Studies, Yangon.A Hybrid Approach to Genetic Algorithm and k Nearest Neighbor Classifier on Information based Distance Metric" A dissertation submitted to the University of Computer Studies, Yangon in fulfillment of the requirement for the degree of Doctor of Philosophy May, 2004.

[8] T. P. Trappengerg and A. D Back, "A Classification Scheme for Applications with Ambiguous Data", IEEE-INNS-ENNS International Joint Conference on Neural Networks, Vol.6, 2000.

[9] W. M. Oo, "Biomedical Data Analysis for Diabetes using Hybrid Learning Method with Genetic Algorithms and Decision Tree" University of Computer Studies, Yangon, Myanmar.

[10] T. M. Cover, T. M., and P. E. Hart, "Nearest Neighbor Pattern Classification".Institute of Electrical and Electronics Engineers Transactions on Information Theory, Vol. 13, no.1, pp. 21-27, 1967.

[11] Han and Kamber, Jiawei Han, Micheline Kamber , Data Mining Concepts and Techniques, 2^{nd} Ed.Original ISBN : 978-1-55860-901-3, 2006, Elservier Inc