

Visualization of Intelligent System using Decision Tree and Fuzzy Clustering for Heart Disease Early Detection

Wiwik Anggraeni, Achmad Pramono, Retno Aulia Vinarti

Department of Information System, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember

Keywords:

decision tree
fuzzy clustering
heart disease
intelligent systems

ABSTRACT

Heart disease is one of serious ailments in most countries especially in Indonesia. Based on Indonesia Heart Foundation reports, there is almost 27 heart disease sufferers in each 100 persons. This rate is the second highest rate of death cause in Indonesia. In addition, the foundation reported that the one who suffered by this illness are youngsters. Early detection is certainly needed in order to reduce the death rate caused by heart failure. Intelligent Systems is one of detection methods that can help in decision making accuracy related to heart disease early detection.

In this research, intelligent system is built by combining decision tree algorithm with fuzzy clustering system. Due to accuracy testing of this combination performance, UCI dataset in Machine Learning Repository was used to categorize the range of heart failure seriousness started from low rate up to high rate of severity.

The AUC (Area under ROC) showed 87.8% as its results which means that decision tree classifier algorithm has a good classification performance by grouping the testing data 87 instance data correctly out of 100. Finally, the long-term objective of this research is to help doctors as a second opinion for early heart diagnoses.

*Copyright © 2013 Information Systems International Conference.
All rights reserved.*

Corresponding Author:

Retno Aulia Vinarti,
Departement of Information System, Faculty of Information Technology,
Institut Teknologi Sepuluh Nopember,
Jalan Raya Kampus ITS, Gedung Sistem Informasi, Sukolilo, Surabaya, Indonesia.
Email: zahra_17@is.its.ac.id

1. INTRODUCTION

Heart disease is non-contagious illness that caused the highest death caused rate in human. According to World Health Organization of United Nations in 2005, heart disease case is 29% or exactly 17.1 million patients died annually in world. Compared to Heart Disease Indonesia Foundation report [1], it shows slight difference which 26.8% with continue increases and age of patients is getting younger year by year. In fact, early detection can successfully do by doctors who have cardiac specialization. However, the cardiac specialists is not common in Indonesia or other developing countries, furthermore, it needs special facilities in hospitals to diagnose the early sign of heart diseases.

Therefore, through this research, the doctors' algorithm in decision making activities to diagnose heart patients was extracted and Intelligent Systems was built based on human experts' algorithms and experience. Even though this application was developed to diagnose patients, but the main decision maker is doctors. This application just helps doctor to make second judgement or as an additional tool.

A similar previous research has developed Intelligent Heart Disease Prediction System (IHDPS) using three algorithms: decision tree, Naive Bayes and Artificial Neural Network (ANN) [2]. The accuracy of each algorithm was 89%, 86.53% and 85% respectively [3]. Based on those results, it can be concluded that decision tree has the best performance in accuracy [4].

Another previous experimentation yielded the same result (99.2%) which is heart disease prediction using classifier algorithm with WEKA machine learning tool. It outnumbered Naive Bayes and clustering which produced accuracy rate 96.5% and 88.3% [5]. Besides the Combining decision tree and fuzzy c-means for predicting and classifying breast cancer has already well-implemented as its high accuracy (92.3%) [6].

Hence, in this work used a combination of those two methods sequentially. Decision tree algorithm suited to classify which one is patient's heart disease severity range. Before doing classification by Decision Tree, the Fuzzy Clustering (FCM) must be used to specify how many cluster created by patient's stage.

2. RESEARCH METHOD

This research was conducted through three main processes which are Data Processing, Rule Generation and Clustering as shown in Figure 1.

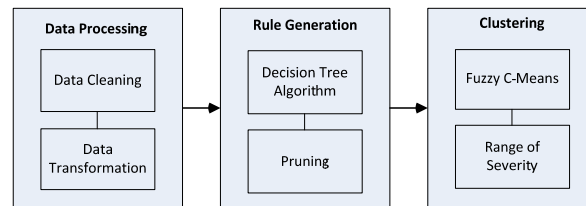


Figure 1. Research Methodology

a. Data Processing

In order to support this research, 14 attributes data from UCI Machine Learning Repository is used to supply heart disease data which are originally from Cleveland Clinic Foundation. First, pre-processing data was performed through two stages: data cleansing and data transformation. The cleansing process filtered attributes which has missing values. By this process, it is discovered that from 303 records contain 6 rows missing values. Therefore, those records were eliminated in order to sharpen the accuracy in the next stage. The second point on the pre-processing step is data transformation. This process converted decimal notation in integer data, for example, 48.0 was converted to 48. These cleansed and transformed data then was inputted to Weka machine learning to define several rules of classification.

Machine Learning requires specific extension format to process data which is arff format. Arff format consists of header and data itself. The arff format for these data was presented in figure 2. The header exhibits a relation name and at the following rows display attributes names and their values range. Variable definition is marked with @ATTRIBUTE and followed by its name and range of values. As aforementioned attributes, there are 13 attributes with 1 class attribute. The attributes are *age* (in years), *sex* which is 1 represent male and 0 represent female, *cp* which is acronym for chest pain have four types: typical angina, atypical angina, non-anginal pain and asymptomatic, *trestbps* represents resting blood pressure (in mmHg), *chol* which is serum cholestoral (in mg/dl), *fbs* is acronym for fasting blood sugars that will be 1 if fbs > 120 mg/dl and 0 if fbs < 120 mg/dl, *restecg* means resting electrocardiographic results which has three values: 0 for normal, 1 for having ST-T wave abnormality and 2 for showing probable or definite left ventricular hypertrophy, *thalach* is maximum heart rate achieved, *exang* represents exercise induced angina which 1 for yes and 0 for no, *oldpeak* which is ST depression induced by exercise relative to test, slope means peak exercise ST segment has three values: 1 for up-sloping, 2 for flat and 3 for down-sloping, *ca* is number of major vessels coloured by fluoroscopy varied from 0 to 3 and the last attribute is *thal* which 3 shows normal, 6 shows fixed defect and 7 shows reversible defect.

```

@RELATION stat_log

@ATTRIBUTE age NUMERIC
@ATTRIBUTE sex {0,1}
@ATTRIBUTE cp {1,2,3,4}
@ATTRIBUTE trestbps NUMERIC
@ATTRIBUTE chol NUMERIC
@ATTRIBUTE fbs {0,1}
@ATTRIBUTE restecg {0,1,2}
@ATTRIBUTE thalach NUMERIC
@ATTRIBUTE exang {0,1}
@ATTRIBUTE oldpeak REAL
@ATTRIBUTE slope {1,2,3}
@ATTRIBUTE ca {0,1,2,3}
@ATTRIBUTE thal {3,6,7}
@ATTRIBUTE class {0,1}

@DATA
29,1,2,130,204,0,2,202,0,0,1,0,3,0
34,1,1,118,182,0,2,174,0,0,1,0,3,0
34,0,2,118,210,0,0,192,0,0,7,1,0,3,0
  
```

Figure 2. Data Transformation to Arff file

b. Rule Generation

Before entering the main process of the classification, all records of data is divided into two parts based on 1:9 proportion. Those two parts are used separately in training and testing session. The aim for this divide is objectivity of judgment for accuracy because data in testing session are completely different to training session. With those proportion obtained 30 records data were used in testing session and 267 records were used to generate decision tree rule. Figure 3 displays accuracy level with J48 decision tree algorithm by Weka is 89.51%.

Correctly Classified Instances	239	89.5131 %						
Incorrectly Classified Instances	28	10.4869 %						
Kappa statistic	0.7883							
Mean absolute error	0.1767							
Root mean squared error	0.2973							
Relative absolute error	35.5276 %							
Root relative squared error	59.6061 %							
Coverage of cases (0.95 level)	100	%						
Mean rel. region size (0.95 level)	91.573	%						
Total Number of Instances	267							
=== Detailed Accuracy By Class ===								
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
	0.93	0.145	0.881	0.93	0.905	0.925	0.9	0
	0.855	0.07	0.914	0.855	0.883	0.925	0.908	1
Weighted Avg.	0.895	0.11	0.896	0.895	0.895	0.925	0.904	
=== Confusion Matrix ===								
a	b	<-- classified as						
133	10	a = 0						
18	106	b = 1						

Figure 3. Capture of Decision Tree Accuracy

From this rule generating activity there was an option to enhance accuracy result by tree pruning. The tree pruning can fit the number of attributes to avoid rule under-fitting or over-fitting. Finally these attributes, as shown by Table 1, are valued to have the best accuracy in result. Based on Figure 3, it can be summarized that 10 from 143 suffered patients misclassified as non-heart disease patients. Moreover, 18 from 124 non-heart disease patients misclassified as heart disease patients. Thus, the developed intelligent system incorrectly classified 28 instance data.

The next step is rule implementation generated by Weka to intelligent systems. This rule is defined by decision tree logic which is consisted precedent and antecedent. Precedent is condition sequences at attributes; meanwhile, antecedent is class attribute from dataset. At Figure 4, antecedent is stated with if statement and precedent is presented as else statement. Implemented decision tree algorithm in Figure 4 is binary tree which only has two branches in each internal node.

Table 1 Selected Attributes through pruning activity

No	Attribute Names
1	age
2	sex
3	cp
4	trestbps
5	chol
6	exang
7	fbs
8	slope
9	thal
10	ca

c. Clustering

Fuzzy C-means has several steps that is needed to be accomplished to decide severity through membership degree between 0 and 1. FCM algorithm was implemented to application based on flows shown by Figure 5.

```

J48 pruned tree
-----
thal = 3
|
|   cp = 4
|   |
|   |   ca = 0
|   |   |
|   |   |   trestbps <= 145.0: 0 (23.0/2.0)
|   |   |   trestbps > 145.0: 1 (6.0/2.0)
|   |   |
|   |   |   ca != 0
|   |   |   |
|   |   |   |   sex = 0
|   |   |   |   |
|   |   |   |   |   slope = 1: 0 (2.0)
|   |   |   |   |   slope != 1: 1 (4.0/1.0)
|   |   |   |   |
|   |   |   |   |   sex != 0: 1 (12.0)
|   |   |   |
|   |   |   |   cp != 4: 0 (100.0/12.0)
|   |   |
|   |   |   thal != 3
|   |   |   |
|   |   |   |   ca = 0
|   |   |   |   |
|   |   |   |   |   exang = 0
|   |   |   |   |   |
|   |   |   |   |   |   fbs = 0
|   |   |   |   |   |   |
|   |   |   |   |   |   |   thal = 6: 0 (4.0)
|   |   |   |   |   |   |   thal != 6
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   age <= 52.0: 1 (10.0/2.0)
|   |   |   |   |   |   |   |   age > 52.0: 0 (12.0/3.0)
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   fbs != 0: 0 (4.0)
|   |   |   |   |   |   |
|   |   |   |   |   |   |   exang != 0
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   slope = 1
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   chol <= 248.0: 0 (3.0)
|   |   |   |   |   |   |   |   |   chol > 248.0: 1 (3.0)
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   slope != 1
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   trestbps <= 115.0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   trestbps > 115.0: 1 (17.0)
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   ca != 0: 1 (64.0/5.0)
|   |   |   |   |   |
|   |   |   |   |   |   thal != 3
|   |   |   |   |   |   |
|   |   |   |   |   |   |   ca = 0
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   exang = 0
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   fbs = 0
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   thal = 6: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   thal != 6
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   age <= 52.0: 1 (10.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   age > 52.0: 0 (12.0/3.0)
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   fbs != 0: 0 (4.0)
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   exang != 0
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   slope = 1
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   chol <= 248.0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   chol > 248.0: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   slope != 1
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   trestbps <= 115.0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   trestbps > 115.0: 1 (17.0)
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   ca != 0: 1 (64.0/5.0)
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   thal != 3
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   ca = 0
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   exang = 0
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   fbs = 0
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   thal = 6: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   thal != 6
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   age <= 52.0: 1 (10.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   age > 52.0: 0 (12.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   fbs != 0: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   exang != 0
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   slope = 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chol <= 248.0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chol > 248.0: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   slope != 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps <= 115.0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps > 115.0: 1 (17.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   ca != 0: 1 (64.0/5.0)
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   thal != 3
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   ca = 0
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   exang = 0
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   fbs = 0
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   thal = 6: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   thal != 6
|   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age <= 52.0: 1 (10.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   age > 52.0: 0 (12.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   fbs != 0: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   exang != 0
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   slope = 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chol <= 248.0: 0 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   chol > 248.0: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   slope != 1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps <= 115.0: 0 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   trestbps > 115.0: 1 (17.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   ca != 0: 1 (64.0/5.0)
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   thal != 3
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   ca = 0
|   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   exang = 0
|   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   fbs = 0
|   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   thal = 6: 0 (4.0)
|   |   |  
```

Figure 4 Part of classifier modelling

1. Initialization

This step requires five variables definition which are X matrices, number of clusters, weight, maximum iteration and stopping rule.

- X matrices contain data that will be clustered.
- Number of clusters should be defined in early stage in order to decide which cluster amount that suited the data most. Usually numbers of clusters are equal or more than two clusters and need a measurement (Sum Squared Error) to justify its correctness.
- Weight is used to define and revise clusters' centroids.
- There are two stopping rules criteria, first, maximum iteration that is used when the clustering has been trapping in local optimum. Second, minimum improvement threshold to determine looping activities when the revised matrices were not significantly performed. The minimum improvement is $1-e5$.

2. Matrices U

At first, partition matrices U are filled by random values. Mostly, this step can reduce the iteration number as the random values in matrices that accidentally close to optimum value will shorten the iteration numbers.

$$U = \begin{bmatrix} \mu_{1,1}(x_1) & \mu_{1,2}(x_2) & \cdots & \mu_{1,n}(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{m,1}(x_1) & \mu_{m,2}(x_2) & \cdots & \mu_{m,n}(x_n) \end{bmatrix}$$

3. Matrices V

Different from matrices U , matrices V contain centroid of each cluster. The centroid formula is shown below.

$$V_{ij} = \frac{\sum_{k=1}^n (M_{ik})^W \cdot x_{kj}}{\sum_{k=1}^n (M_{ik})^W}$$

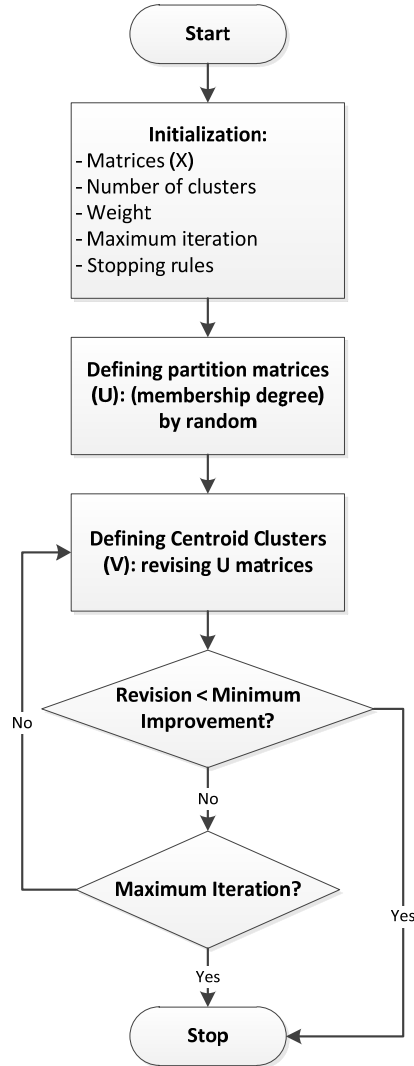


Figure 5. Clustering flowchart

4. Revision step

Membership degree in each cluster

$$u_{ik} = \left[\sum_{j=1}^c \left(\frac{d_{ij}}{d_{jk}} \right)^{\frac{2}{2/(m-1)}} \right]^{-1}$$

with the formula below

$$d_{ik} = d(x_{ik} - v_k) = \left[\sum_{j=1}^m (x_{ij} - v_{kj})^2 \right]^{1/2}$$

5. Stopping rules checking

If revised matrices show margin lower than minimum improvement then looping activity must be stopped. Another stopping rule is maximum iteration number which is performed if the margin still higher than minimum improvement but it exceeds allowed maximum iteration.

3. RESULTS AND ANALYSIS

Besides performing classification and clustering methods, this research also enclosed by visualization named Intelligent Heart Disease Prediction Systems (IHDPS). The interface can illustrate heart disease severity stage as shown by Figure 6. That figure also shows clogging in sufferers' blood vessels. Table 2 displays percentage range of blood vessels' clogging which is the main cause of heart disease.

3.1. Classification Validation Testing

Classification validation was performed by counting precision, recall and accuracy using AUC (Area Under ROC) [7]. The confusion matrix which is result of validation data testing is showed by Table 3.

Table 2. Range of blood vessels' clogging heading to heart organ

No	Cluster	Cluster Values	Clogging Percentage
1	Cluster 1 (50-100 %)	< 20 %	90 – 100 %
2	Cluster 1 (50-100 %)	20 – 40 %	80 – 90 %
3	Cluster 1 (50-100 %)	40 – 60 %	70 – 80 %
4	Cluster 1 (50-100 %)	60 – 80 %	60 – 70 %
5	Cluster 1 (50-100 %)	80 – 100 %	50 – 60 %
6	Cluster 2 (0 - 50 %)	< 20 %	40 – 50 %
7	Cluster 2 (0 - 50 %)	20 – 40 %	30 – 40 %
8	Cluster 2 (0 - 50 %)	40 – 60 %	20 – 30 %
9	Cluster 2 (0 - 50 %)	60 – 80 %	10 – 20 %
10	Cluster 2 (0 - 50 %)	80 – 100 %	< 10 %

Table 3. Confusion Matrix Testing

Actual Class	Predicted Class		
		class = yes	class = no
	class = yes	11	2
	class = no	1	16

Based on confusion matrix in table 3, precision and recall from validation testing output are showed below.

$$\text{Precision} = \frac{TP}{TP + FP} \times 100 = \frac{11}{11 + 1} \times 100 = 91,67\%$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100 = \frac{11}{11 + 2} \times 100 = 84,61\%$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 = \frac{11 + 16}{11 + 16 + 1 + 2} \times 100 = 90\%$$

For next validation testing in a classification case, most research used AUC. AUC is area below Receiver Operation Characteristic (ROC) curve. ROC consists of an absis which shows false positive rate an ordinate that measure true positive rate. A good classifier will have ROC Curve increase from left-top side in a graph. That shows ROC has a high true positive rate and low false positive rate. Figure 6 displays ROC performed by decision tree method by this research. Table 4 displays standard measurement for AUC [3]. It divides 5 ranges of classifier performance based on AUC range. IHDPS has 0.878 of AUC then it is classified as good classification which has a good performance to classify.

Table 4. AUC performance classification

No	Range	Goodness
1	0.90 - 1.00	<i>excellent classification</i>
2	0.80 - 0.90	<i>good classification</i>
3	0.70 - 0.80	<i>fair classification</i>
4	0.60 - 0.70	<i>poor classification</i>
5	0.50 - 0.60	<i>failure</i>

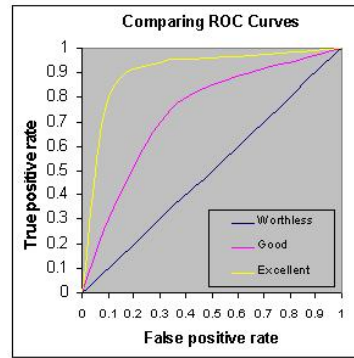


Figure 6. ROC classifier performances

3.2. Clustering Validation Testing

Clustering validation testing was done by comparing manual computing objective function (J) to Matlab FCM function. This comparison had been taken in consideration of SSE. SSE result taken from Matlab is 235716.93 then SSE counting performed manually which shows in Table 5. Iteration will be stopping if minimum improvement less than $1\text{-e}5$. It means that centroid clusters have been converged. Table 5 displays that in 25th iteration minimum improvement not higher than minimum improvement then it stopped at objective function 235716,935675. It is similar to Matlab result, thus, through clustering validation testing, it can be concluded that this method has been validated.

Table 5. Objective function manually counting

Iteration	Objective Function (J)	Min Improvement (1-e-5)
1	454761,4338	-
2	340594,4781	114166,9557
3	306860,1236	33734,3545
4	256355,7959	50504,328
5	238388,7909	17967,005
6	236137,2313	2251,5596
7	235843,3041	293,9272
8	235764,8719	78,4322
9	235735,8193	29,0526
10	235724,4132	11,4061
11	235719,8989	4,5143
12	235718,1101	1,7888
13	235717,4012	0,7089
14	235717,1202	0,281
15	235717,0088	0,1114
16	235716,9647	0,0441
17	235716,9472	0,0175
18	235716,9402	0,007
19	235716,9375	0,0027
20	235716,9364	0,0011
21	235716,9360	0,0004
22	235716,9358	0,0002
23	235716,93571	0,00009
24	235716,935686	0,00002
25	235716,935675	0,0000015

4. CONCLUSION




According to two validations testing in this research, it can be summarized that Decision Tree algorithm has a high AUC rate (0.878) which is categorized as good classifier. Another conclusion can be taken through Clustering Validation method it can be seen that objective function (SSE) has reached the optimum value at 25th iteration with SSE value is 235716.93567. For further development, it will be needed more training data in order to improve accuracy of classifier model. Besides improving training data, it needs to be

more tested to other ailments such as Breast Cancer or Diabetes which have a high death rate below Heart Disease.

REFERENCES

- [1] Antara. 2011. Surya. Retrieved February 15, 2012, Available: <http://www.surya.co.id/2011/09/15/268-persen-kasus-penyakit-jantung-di-indonesia>
- [2] D,W, Aha, 2007. Retrieved Februari 22, 2012, from UCI Machine Learning repository. Available : <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>
- [3] F. Gorunescu. 2011. Data Mining: Concepts, Models and Techniques. Springer.
- [4] J. Soni, U. Ansari, D. Shama, & S. Soni. 2011. Predictive Data Mining for Medical Diagnosis: An Overview. International Journal of Computer Applications, 17.
- [5] M, A., E, A., & N.CH.S.N, I. 2010. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology, 1-7.
- [6] S., Shanthi & V. M. Bhaskaran. 2011. Intuitionistic Fuzzy C-Means and Decision Tree Approach for. European Journal of Scientific Research, 345-351.
- [7] T. G. Tape. (n.d.). The Area Under an ROC Curve. Retrieved July 11, 2012, Available: Interpreting Diagnostic Tests: <http://gim.unmc.edu/dxtests/roc3.htm>

BIBLIOGRAPHY OF AUTHORS

	<p>Achmad Pramono Graduated from Information System Department in 2013 Member of Decision Support Systems Research Centre</p>
	<p>Wiwik Anggraeni Lecturer of Information Systems Department, Information Technology Faculty Institut Teknologi Sepuluh Nopember Member of Decision Support Systems Research Centre Expertise area in Forecasting and Statistics</p>
	<p>Retno Aulia Vinarti Lecturer of Information Systems Department, Information Technology Faculty Institut Teknologi Sepuluh Nopember Member of Decision Support Systems Research Centre Expertise area in Forecasting and Data Mining</p>