A Study of Data Randomization on a Computer Based Feature Selection for Diagnosing Coronary Artery Disease

D. W. Prabowo, N. A. Setiawan, H. A. Nugroho
Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia

Abstract

The objective of this research is to investigate the randomization of data on a computer based feature selection for diagnosing coronary artery disease. The randomization on Cleveland dataset was conducted because the performance value is different for each experiment. Assuming the performance values have a Gaussian probability distribution is a solution to handle different performance value provided by the process of randomizing dataset. The final performance is taken from the mean value of all performance value. In this research, computer based feature selection (CFS), medical expert based feature selection (MFS) and combined both of MFS and CFS (MFS+CFS) are also conducted to improve the performance of the classification algorithm. Also, this research found a different characteristic on Cleveland dataset from previous work. This difference obviously can affect the feature selection result and the final performance. In summary, the randomization dataset and computing the final performance can generally represent the performance of the classification algorithm.

Keywords: CFS, Classification algorithm, Coronary artery disease, Cleveland dataset, Gaussian probability distribution, MFS, Randomization.

1 Introduction

Coronary Artery Disease (CAD), sometimes called as Coronary Heart Disease (CHD) is the most common heart diseases. CAD occurs when the blood flow to the heart muscle in the coronary arteries is blocked by atherosclerosis (fatty deposits) [1]. It has a very high mortality rate, e.g. in 2008 an estimated 7.3
million deaths in the world are caused by CAD [2]. The initial diagnosis usually uses medical history and physical examination, then further testing can be done. For further testing, coronary angiography provides the “gold standard” diagnosis of disease in the coronary arteries [3]. Coronary angiography test is preferred by cardiologists to diagnose the presence of CAD with high accuracy even though invasive, risky and expensive [4].

From the shortcomings of this test, it is necessary to develop a method which is capable of diagnosing CAD before coronary angiography test. The goal is to avoid invasive, risky and expensive diagnostic procedures to the patient. Therefore, this motivates the development of a computer based method to be able to diagnose the presence of CAD. The computer based method can provide diagnostic procedures to patients in a way that is non-invasive, safe and less expensive.

Various computer based methods have been developed to identify heart related diseases. The methods of neural network [5], fuzzy [6] and data mining [7] are proposed to diagnose CAD. The neural network based methods have advantages on nonlinear prediction, strong on parallel processing and the ability to tolerate faults, but they have a weakness in the need for a large training data, over-fitting, slow convergence and local optimum [8]. Fuzzy logic offers reasoning at a higher level by using linguistic information obtained from domain experts, but fuzzy systems lack the ability to learn and cannot adjust to the new environment [9]. Data mining which is a process of extracting hidden knowledge from the data offers other advantages. This method can reveal patterns and relationships among large amounts of data in a single dataset or not [10].

In medical diagnosis, data reduction is an important issue. Medical data often contain a large number of features that are irrelevant, redundant and relatively small number of cases that can affect the quality of disease diagnosis [11]. Therefore, the feature selection process can be used to select relevant features in medical data. Feature selection is proposed in many researches [11][12][13][14][15] to improve the accuracy in the diagnosis of CAD.

Nahar et al. [14] performed computer-based feature selection process. This process is called by the computer feature selection (CFS). CFS selects features randomly so there is the possibility to dispose of medical significant factors. To avoid the loss of medical significant factors, the feature selection process needs to be carried out by medical experts (termed as MFS). These significant factors are age, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting heart rate, maximum heart rate and exercise-induced angina. For CFS, Nahar et al. used CfsSubsetEval as attribute selection method (using BestFirst search strategy) provided by Weka.

There is a difference in the characteristics of the Cleveland dataset between current research and Nahar et al. [14]. The difference lies in the total of positive class instances. This difference obviously can affect the final performance. But there is something important issue that Nahar et al. did not considered. This issue is the effect of randomizing process in data which can affect the performance of computer based diagnosis. In this research, the study of randomizing medical data (Cleveland dataset) is discussed.
2 Material and method

2.1 Cleveland dataset

In this research, feature selection process is used in Cleveland dataset. It used a maximum of 14 attributes of 76 attributes of the Cleveland dataset (with 303 total instances). Table 1 describes the attributes and their data types used in the dataset [14],[16].

| Attribute | Description                              | Value description                                      |
|-----------|------------------------------------------|-------------------------------------------------------|
| age       | Age                                      | Numeric                                               |
| sex       | Sex                                      | male; female                                          |
| cp        | Chest pain type                          | typical angina (angina); atypical angina (abnang); non-anginal pain (notang); asymptomatic (asympt) |
| trestbps  | Patient’s resting blood pressure in mm Hg at the time of admission to the hospital | Numeric                                               |
| chol      | Serum cholesterol in mg/dl               | Numeric                                               |
| fbs       | Boolean measure indicating whether fasting blood sugar is greater than 120 mg/dl | 1 if true; 0 if false                                 |
| restecg   | Electrocardiographic results during rest | normal (norm); abnormal (abn): having ST-T wave abnormality; ventricular hypertrophy (hyp) |
| thalach   | Maximum heart rate attained              | Numeric                                               |
| exang     | Boolean measure indicating whether exercise induced angina has occurred | 1 if yes; 0 if no                                      |
| oldpeak   | ST depression brought about by exercise relative to the rest | Numeric                                               |
| slope     | The slope of the ST segment for peak exercise | upsloping; flat; downsloping                          |
| ca        | A number of major vessels colored by fluoroscopy | Numeric                                               |
| thal      | The heart status                         | normal; fixed defect, reversible defect               |
| class     | The value is either healthy or heart disease | Sick type: 1, 2, 3, and 4                             |

This dataset is converted from a multi-class problem into a binary-class in order to obtain five datasets with characteristics in accordance with table 2. From table 2, it can be seen that the total of positive class is same with the total instance of Cleveland dataset (303 instances). Logically, this must happen because the datasets (H-0, Sick-1, Sick-2, Sick-3 and Sick-4) are same with the Cleveland dataset. The difference is only the class label considered as positive. From table 3, it can be seen that the total of positive class is 308 instances. The total of positive class is different from the total instance of Cleveland dataset.
Table 2: Characteristic of Cleveland dataset

| Dataset Name | Positive Class | Number of Positive Class | Number of Negative Class | Status indicated by positive class |
|---------------|----------------|--------------------------|--------------------------|-----------------------------------|
| H-0           | Health         | 165                      | 138                      | Health, Sick                      |
| Sick-1        | S1             | 54                       | 249                      | S1, Negative                      |
| Sick-2        | S2             | 36                       | 267                      | S2, Negative                      |
| Sick-3        | S3             | 35                       | 268                      | S3, Negative                      |
| Sick-4        | S4             | 13                       | 290                      | S4, Negative                      |

Table 3: Characteristic of Cleveland dataset (Nahar et al.)

| Dataset Name | Positive Class | Number of Positive Class | Number of Negative Class | Status indicated by positive class |
|---------------|----------------|--------------------------|--------------------------|-----------------------------------|
| H-0           | Health         | 165                      | 138                      | Health, Sick                      |
| Sick-1        | S1             | 56                       | 247                      | S1, Negative                      |
| Sick-2        | S2             | 37                       | 266                      | S2, Negative                      |
| Sick-3        | S3             | 36                       | 268                      | S3, Negative                      |
| Sick-4        | S4             | 14                       | 289                      | S4, Negative                      |

2.2 Motivated feature selection (MFS)

Motivated feature selection is the process of feature selection by medical experts. There are eight factors to be considered by the medical significance of MFS in the process of feature selection. These factors are age, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting heart rate, maximum heart rate and exercise induced angina [14].

2.3 Computer feature selection (CFS)

For CFS, Nahar et al. use CfsSubsetEval as attribute selection method (using BestFirst search strategy) provided by Weka. CFS selects features randomly so there is the possibility to dispose of medical significant factors [14].

2.4 Classifier algorithm

In this research, the six well known classifiers (Naïve Bayes, SMO, IBK, AdaBoostM1, J48 and PART) were used. This is the reason why Cleveland dataset has to convert to binary-class, because these algorithms are binary classifier.
2.4.1 Naïve Bayes
A Naïve Bayes is a probabilistic classification algorithm based on applying Bayes theorem assuming a strong independent. Eqn (1) is the probability of data record $X$ that has a label $C_j$.

$$
P(C_j | X) = \frac{P(X | C_j)^* P(C_j)}{P(X)}
$$

(1)

$C_j$ is class label with the largest conditional probability value determines the category of the data record [7].

2.4.2 SMO
There are two components of the SMO algorithm. These components are an analytical method to solve two Lagrange multipliers and heuristic methods to determine the optimizing multiplier. This algorithm was introduced by John Platt in 1998 at Microsoft Research [17].

2.4.3 IBK
The algorithm found a group of $k$ objects in the training set that are closest to the test object and the label on the basis of the assignment of a certain class domination. It discusses the main issue in many datasets that may not be exactly matching one object with another object, as well as the fact that conflicting information about the class of an object can be obtained from the nearest objects [18].

2.4.4 AdaBoostM1
“Boosting” is a general method for improving the performance of any learning algorithm. The boosting can be used to significantly reduce the error of any “weak” learning algorithm that consistently generates classifiers which need only be a little bit better than random guessing [19].

2.4.5 J48
J48 is a classification algorithm that implements the C4.5 algorithm [10]. C4.5 algorithm is intended for supervised learning. C4.5 learns to mapping an attribute values to a class that can be applied to classify a new class (unseen instance) [18].

2.4.6 PART
PART algorithm builds a tree using C4.5's heuristics with the parameters specified by the user same with J48. The rules of the classification algorithm derived from the partial decision tree. Partial decision tree is a decision tree that contains branches of undefined sub-trees [10].

2.5 Research design
The CAD diagnosis process is provided by classifying Cleveland dataset using six well known algorithms. To provide a comparison among the classification algorithms, four performance metrics were used (accuracy, true positive rate, f-measure and training time).
The process of randomizing instances in the Cleveland dataset can affect the performance of computer-based diagnosis. Every randomizing process provides different performance values. Assume the performance values have a distribution $X$ and $\bar{X}$ is the mean of a random sample of size $n$ taken from a population with mean $\mu$ and finite variance $\sigma^2$, then the limiting form of the distribution of $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ as $n \to \infty$, is the standard Gaussian distribution $N(\mu, \sigma^2)$. Therefore, assuming the performance values have a Gaussian probability distribution is a solution to handle different performance values provided by the process of randomizing dataset. In this research, the minimum sample size is 100. The 100 times randomizing instance give 100 different performance values. Then the final performance is taken from the mean of those values. For each randomization instance, the performance of the algorithm can be obtained by using two-ways. First, applying the 10-fold cross-validation process on the dataset Cleveland. Second, applying the train-test split to the dataset and then used 10-fold cross-validation to choose the best parameters in the training process. In the process of train-test split, each dataset was subjected to a stratified sampling process to select two-thirds of the data for training and the rest for prediction. One of the tools provided by Weka (CVParameter) is used in the train-test split. Figure 1 and 2 describe these processes.
3 Experiment result and analysis

3.1 Computer Feature selection

A feature selection result that obtained from CFS can be seen in table 4. For each dataset, the features that are selected by CFS are different.

| Dataset | Feature Selection Result |
|---------|--------------------------|
| H-0     | chest pain, resting ECG, maximum heart rate, exercise induced angina, old peak, the number of vessels coloured, thal |
| Sick-1  | sex, chest pain, fasting blood sugar, resting ECG, exercise induced angina, thal |
| Sick-2  | chest pain, fasting blood sugar, maximum heart rate, exercise induced angina, oldpeak, number of vessels coloured, thal |
| Sick-3  | maximum heart rate, exercise induced angina, oldpeak, number of vessels coloured, thal |
| Sick-4  | resting ECG, oldpeak, number of vessels coloured |

It can be seen from table 4, CFS does not select the features that are considered as medical significant factor by MFS. Therefore, to avoid the medical significant factor is not selected, it is necessary to combine MFS and CFS.

3.2 Classification algorithm performance

Table 5 shows the final performance of the classification algorithm when performed on the five datasets. The bold values indicate the best algorithm for each dataset. Applying 10-fold cross validation and CVP 10-fold, SMO is the best algorithm (in terms accuracy) in dataset Sick-1, Sick-2, Sick-3 and Sick-4 whereas Naïve Bayes is the best algorithm in dataset H-0. In terms true positive rate and f-measure, Naïve Bayes is the best algorithm in dataset Sick-2, Sick-3 and Sick-4.

3.3 Comparison of performance between CVP-10 fold, CFS and MFS

Table 6 shows the comparison of the final performance of the classification algorithm (in terms accuracy, true positive rate and f-measure) before and after the feature selection process is used. The bold values indicate the best algorithm for each dataset. Applying CFS process, the accuracy of CFS is better than MFS for all algorithm in dataset H-0. For dataset Sick-1, the accuracy of CFS is better than MFS for three cases (SMO, J48 and PART). For dataset Sick-2 and Sick-3, the accuracy of CFS is better than MFS for one case (PART). For dataset Sick-4, the accuracy of CFS is better than MFS for two cases (Naïve Bayes and PART). Table 6 also shows the performance result for CVP-10 fold (no feature selection). The highlighted values indicate the accuracy of MFS or CFS is better than CVP-10 fold. For H-0, the accuracy of CFS is better than CVP-10 fold in...
four cases (Naïve Bayes, SMO, AdaBoostM1 and J48). For Sick-1, the accuracy of CFS is better than CVP-10 fold in five cases (Naïve Bayes, IBK, AdaBoostM1, J48 and PART) and the accuracy of MFS is better than CVP-10 fold in four cases (Naïve Bayes, IBK, AdaBoostM1 and J48). For Sick-2, the accuracy of CFS and MFS is better than CVP-10 fold in three cases (Naïve Bayes, AdaBoostM1 and PART). For Sick-3, the accuracy of CFS is better than CVP-10 fold in three cases (Naïve Bayes, AdaBoostM1 and PART) and the accuracy of MFS is better than CVP-10 fold in three cases (Naïve Bayes, AdaBoostM1 and J48). For Sick-4, the accuracy of CFS and MFS is better than CVP-10 fold in three cases (Naïve Bayes, J48 and PART).

Table 5: Performance for 10-fold and CVP 10-fold

| Dataset | Algorithm  | Naïve Bayes | SMO | IBK | AdaBoostM1 | J48 | PART |
|---------|------------|-------------|-----|-----|------------|-----|------|
|         | 10-fold    | Accuracy (%)| CVP 10-fold | 10-fold | CVP 10-fold | 10-fold | CVP 10-fold | 10-fold | CVP 10-fold |
| H-0     | Naïve Bayes| 84.042      | 84.249 | 0.868 | 0.873 | 0.855 | 0.858 | 0.001 | 0.000 |
|         | SMO        | 83.065      | 83.183 | 0.839 | 0.866 | 0.842 | 0.848 | 0.083 | 1.039 |
|         | IBK        | 83.131      | 82.368 | 0.872 | 0.860 | 0.849 | 0.841 | 0.000 | 0.169 |
|         | AdaBoostM1 | 83.298      | 80.541 | 0.861 | 0.837 | 0.848 | 0.824 | 0.021 | 0.022 |
|         | J48        | 78.892      | 75.817 | 0.844 | 0.806 | 0.812 | 0.783 | 0.002 | 0.187 |
|         | PART       | 80.560      | 79.437 | 0.860 | 0.842 | 0.827 | 0.816 | 0.002 | 0.189 |
| Sick-1  | Naïve Bayes| 77.549      | 78.316 | 0.110 | 0.115 | 0.139 | 0.151 | 0.001 | 0.001 |
|         | SMO        | 82.194      | 82.175 | 0.000 | 0.000 | 0.000 | 0.000 | 0.026 | 1.102 |
|         | IBK        | 80.909      | 81.186 | 0.016 | 0.010 | 0.021 | 0.015 | 0.000 | 0.232 |
|         | AdaBoostM1 | 82.194      | 77.631 | 0.000 | 0.160 | 0.000 | 0.199 | 0.144 | 0.080 |
|         | J48        | 81.411      | 81.409 | 0.014 | 0.014 | 0.020 | 0.020 | 0.002 | 0.158 |
|         | PART       | 81.459      | 81.382 | 0.009 | 0.010 | 0.011 | 0.012 | 0.002 | 0.240 |
| Sick-2  | Naïve Bayes| 78.615      | 80.090 | 0.426 | 0.327 | 0.313 | 0.275 | 0.001 | 0.001 |
|         | SMO        | 88.036      | 88.148 | 0.000 | 0.000 | 0.001 | 0.000 | 0.027 | 1.275 |
|         | IBK        | 87.755      | 87.730 | 0.002 | 0.009 | 0.003 | 0.016 | 0.000 | 0.217 |
|         | AdaBoostM1 | 84.827      | 83.006 | 0.062 | 0.086 | 0.072 | 0.088 | 0.145 | 0.077 |
|         | J48        | 87.515      | 87.876 | 0.011 | 0.011 | 0.013 | 0.017 | 0.001 | 0.128 |
|         | PART       | 86.347      | 86.926 | 0.026 | 0.021 | 0.030 | 0.024 | 0.002 | 0.168 |
| Sick-3  | Naïve Bayes| 82.305      | 82.578 | 0.513 | 0.478 | 0.394 | 0.386 | 0.000 | 0.000 |
|         | SMO        | 88.132      | 88.422 | 0.025 | 0.000 | 0.035 | 0.000 | 0.025 | 1.153 |
|         | IBK        | 87.563      | 88.023 | 0.000 | 0.005 | 0.000 | 0.008 | 0.000 | 0.213 |
|         | AdaBoostM1 | 85.488      | 83.423 | 0.211 | 0.153 | 0.226 | 0.164 | 0.161 | 0.070 |
|         | J48        | 87.860      | 87.878 | 0.010 | 0.013 | 0.012 | 0.017 | 0.001 | 0.119 |
|         | PART       | 86.471      | 87.306 | 0.038 | 0.034 | 0.040 | 0.040 | 0.002 | 0.137 |
| Sick-4  | Naïve Bayes| 93.329      | 93.797 | 0.074 | 0.075 | 0.076 | 0.099 | 0.001 | 0.001 |
|         | SMO        | 95.731      | 95.703 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 | 0.860 |
|         | IBK        | 95.725      | 95.451 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.171 |
|         | AdaBoostM1 | 95.682      | 95.703 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.103 |
|         | J48        | 95.731      | 95.654 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.110 |
|         | PART       | 95.154      | 95.364 | 0.009 | 0.007 | 0.007 | 0.006 | 0.001 | 0.140 |
Table 6: Performance for CVP-10 fold, CFS and MFS

| Dataset  | Algorithm       | Accuracy (%) | TP       | F-measure |
|----------|-----------------|--------------|----------|-----------|
|          |                 | MFS          | CFS      | CVP 10-fold | MFS | CFS  | CVP 10-fold | MFS | CFS  | CVP 10-fold |
| H-0      | Naïve Bayes     | 75.106       | 85.075   | 84.249     | 0.801 | 0.906 | 0.873     | 0.777 | 0.869 | 0.858     |
|          | SMO             | 75.514       | 83.201   | 83.183     | 0.768 | 0.878 | 0.866     | 0.773 | 0.850 | 0.848     |
|          | IBK             | 75.659       | 81.680   | 82.368     | 0.836 | 0.889 | 0.860     | 0.789 | 0.841 | 0.841     |
|          | AdaBoostM1      | 75.009       | 83.443   | 80.541     | 0.799 | 0.878 | 0.837     | 0.776 | 0.852 | 0.824     |
|          | J48             | 71.313       | 77.321   | 75.817     | 0.781 | 0.827 | 0.806     | 0.747 | 0.798 | 0.783     |
|          | PART            | 73.486       | 78.864   | 79.437     | 0.789 | 0.852 | 0.842     | 0.763 | 0.814 | 0.816     |
| Sick-1   | Naïve Bayes     | 81.906       | 79.083   | 78.316     | 0.997 | 0.093 | 0.115     | 0.901 | 0.122 | 0.151     |
|          | SMO             | 82.168       | 82.175   | 82.175     | 1.000 | 0.000 | 0.000     | 0.902 | 0.000 | 0.000     |
|          | IBK             | 81.584       | 81.428   | 81.186     | 0.992 | 0.006 | 0.010     | 0.898 | 0.010 | 0.015     |
|          | AdaBoostM1      | 81.566       | 79.169   | 77.631     | 0.992 | 0.094 | 0.160     | 0.898 | 0.123 | 0.199     |
|          | J48             | 81.605       | 82.175   | 81.409     | 0.992 | 0.000 | 0.014     | 0.899 | 0.000 | 0.020     |
|          | PART            | 81.140       | 81.612   | 81.382     | 0.983 | 0.003 | 0.010     | 0.895 | 0.005 | 0.012     |
| Sick-2   | Naïve Bayes     | 86.001       | 81.628   | 80.090     | 0.962 | 0.396 | 0.327     | 0.923 | 0.338 | 0.275     |
|          | SMO             | 88.112       | 88.070   | 88.148     | 1.000 | 0.002 | 0.000     | 0.937 | 0.004 | 0.000     |
|          | IBK             | 87.588       | 87.449   | 87.730     | 0.994 | 0.029 | 0.009     | 0.934 | 0.043 | 0.016     |
|          | AdaBoostM1      | 84.977       | 83.348   | 83.006     | 0.952 | 0.183 | 0.086     | 0.918 | 0.194 | 0.088     |
|          | J48             | 87.773       | 87.614   | 87.876     | 0.996 | 0.013 | 0.011     | 0.935 | 0.017 | 0.017     |
|          | PART            | 87.243       | 87.352   | 86.926     | 0.988 | 0.028 | 0.021     | 0.932 | 0.036 | 0.024     |
| Sick-3   | Naïve Bayes     | 85.015       | 85.626   | 82.578     | 0.944 | 0.351 | 0.478     | 0.922 | 0.358 | 0.386     |
|          | SMO             | 88.284       | 87.984   | 88.422     | 0.997 | 0.029 | 0.000     | 0.938 | 0.042 | 0.000     |
|          | IBK             | 87.896       | 87.283   | 88.023     | 0.992 | 0.039 | 0.005     | 0.935 | 0.056 | 0.008     |
|          | AdaBoostM1      | 86.654       | 86.383   | 83.423     | 0.969 | 0.214 | 0.153     | 0.928 | 0.251 | 0.164     |
|          | J48             | 87.994       | 87.830   | 87.878     | 0.993 | 0.025 | 0.013     | 0.936 | 0.029 | 0.017     |
|          | PART            | 87.197       | 87.782   | 87.306     | 0.976 | 0.038 | 0.034     | 0.931 | 0.045 | 0.040     |
| Sick-4   | Naïve Bayes     | 95.313       | 95.567   | 93.797     | 0.996 | 0.148 | 0.075     | 0.976 | 0.209 | 0.099     |
|          | SMO             | 95.701       | 95.596   | 95.703     | 1.000 | 0.004 | 0.000     | 0.978 | 0.005 | 0.000     |
|          | IBK             | 95.313       | 95.139   | 95.451     | 0.996 | 0.003 | 0.000     | 0.976 | 0.002 | 0.000     |
|          | AdaBoostM1      | 95.119       | 94.247   | 95.703     | 0.994 | 0.053 | 0.000     | 0.975 | 0.059 | 0.000     |
|          | J48             | 95.711       | 95.664   | 95.654     | 1.000 | 0.000 | 0.000     | 0.978 | 0.000 | 0.000     |
|          | PART            | 95.508       | 95.567   | 95.364     | 0.997 | 0.002 | 0.007     | 0.977 | 0.003 | 0.006     |

3.4 Combination of MFS and CFS process

Table 7 shows the comparison of the final performance when MFS and CFS are combined. The bold values indicate the best algorithm for each dataset. The highlighted values indicate accuracy of MFS+CFS is better than MFS. For H-0, the accuracy of MFS+CFS is better than MFS for all algorithms. For Sick-1, the accuracy of MFS+CFS is better than MFS for two cases (J48 and PART). For Sick 2, the accuracy of MFS+CFS is better than MFS for two cases (IBK and PART). For Sick-3, the accuracy of MFS+CFS is better than MFS for one case (PART). For Sick-4, there is no accuracy of MFS+CFS better than MFS.
Table 7: Performance for MFS and MFS+CFS

| Dataset | Algorithm       | Accuracy (%) | TP   | F-measure |
|---------|-----------------|--------------|------|-----------|
|         | MFS             | MFS+CFS      | MFS  | MFS+CFS   |
| H-0     | Naïve Bayes     | 75.106       | 83.697 | 0.801 | 0.880 | 0.777 | 0.854 |
|         | SMO             | 75.514       | 82.746 | 0.768 | 0.870 | 0.773 | 0.846 |
|         | IBK             | 75.659       | 81.640 | **0.836** | 0.875 | **0.789** | 0.838 |
|         | AdaBoostM1      | 75.009       | 82.445 | 0.799 | 0.862 | 0.776 | 0.842 |
|         | J48             | 71.313       | 76.428 | 0.781 | 0.815 | 0.747 | 0.789 |
|         | PART            | 73.486       | 85.239 | 0.789 | 0.845 | 0.763 | 0.808 |
| Sick-1  | Naïve Bayes     | 81.906       | 75.848 | 0.997 | 0.895 | 0.901 | 0.858 |
|         | SMO             | **82.168**   | **82.158** | **1.000** | **1.000** | **0.902** | **0.902** |
|         | IBK             | 81.584       | 81.392 | 0.992 | 0.989 | 0.898 | 0.897 |
|         | AdaBoostM1      | 81.566       | 78.837 | 0.992 | 0.939 | 0.898 | 0.879 |
|         | J48             | 81.605       | 81.844 | 0.992 | 0.995 | 0.899 | 0.900 |
|         | PART            | 81.140       | 81.195 | 0.983 | 0.985 | 0.895 | 0.896 |
| Sick-2  | Naïve Bayes     | 86.001       | 80.514 | 0.962 | 0.866 | 0.923 | 0.886 |
|         | SMO             | 88.112       | 87.889 | **1.000** | **0.996** | **0.937** | **0.935** |
|         | IBK             | 87.588       | 87.607 | 0.994 | 0.994 | 0.934 | 0.934 |
|         | AdaBoostM1      | 84.977       | 83.463 | 0.952 | 0.923 | 0.918 | 0.907 |
|         | J48             | 87.773       | 87.667 | 0.996 | 0.994 | 0.935 | 0.934 |
|         | PART            | 87.247       | 87.307 | 0.988 | 0.987 | 0.932 | 0.932 |
| Sick-3  | Naïve Bayes     | 85.915       | 83.177 | 0.944 | 0.885 | 0.922 | 0.903 |
|         | SMO             | 88.284       | 88.168 | **0.997** | **0.995** | **0.938** | **0.937** |
|         | IBK             | 87.896       | 87.701 | 0.992 | 0.990 | 0.935 | 0.934 |
|         | AdaBoostM1      | 86.654       | 84.711 | 0.969 | 0.931 | 0.928 | 0.915 |
|         | J48             | 87.994       | 87.964 | 0.993 | 0.992 | 0.936 | 0.936 |
|         | PART            | 87.197       | 87.681 | 0.976 | 0.987 | 0.931 | 0.934 |
| Sick-4  | Naïve Bayes     | 95.313       | 93.410 | 0.996 | 0.970 | 0.976 | 0.966 |
|         | SMO             | 95.701       | **95.663** | **1.000** | **0.999** | **0.978** | **0.978** |
|         | IBK             | 95.313       | 95.167 | 0.996 | 0.994 | 0.976 | 0.975 |
|         | AdaBoostM1      | 95.119       | 94.303 | 0.994 | 0.985 | 0.973 | 0.971 |
|         | J48             | **95.711**   | **95.556** | **1.000** | **0.998** | **0.978** | **0.977** |
|         | PART            | 95.508       | 95.215 | 0.997 | 0.995 | 0.977 | 0.975 |

4 Conclusions and future works

The difference in the characteristics of the Cleveland dataset between current research and previous work, it obviously affects the feature selection result and the final performance. The total of positive class must same with the total instance of Cleveland dataset (303 instances). Gaussian probability distribution is a solution to handle the different performance value provided by the process of randomizing dataset. By dataset randomization then computes the final performance, it can generally represent the performance of the classification algorithm. Table 5 shows that CVP 10-fold improve the accuracy than 10-fold cross validation at dataset H-0 (Naïve Bayes and SMO), Sick-1 (Naïve Bayes and IBK), Sick-2 (Naïve Bayes, SMO, J48 and PART), Sick-3 (Naïve Bayes,
SMO, IBK, J48 and PART) and Sick-4 (Naïve Bayes, AdaBoostM1 and PART). From the analysis of final performance result, it can be seen that the feature selection process (CFS and MFS) improve the accuracy in some case than only apply CVP 10-fold (without feature selection). Then, to improve the ability of computer based feature selection, the method of combined MFS and CFS can be proposed. From table 7, the method of combined MFS and CFS improve the accuracy in some case for dataset H-0, Sick-1, Sick-2 and Sick-3 than only apply MFS process. For CFS, this research only use one attribute selection method (CfsSubsetEval) so this is not generally represent the CFS process. In the future works, the modification of the CFS method with other attribute selection is recommended to improve the performance of diagnosing coronary artery disease. Also, the modification of CFS can combined with MFS to ensure the medical expert about the diagnosis result.

5 Acknowledgements

The research work was supported by Intelligent System Research Group at Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada.

References

[1] Randall, O. S., Segerson, N. M. & Romaine, D. S., The Encyclopedia of the Heart and Heart Disease, 2nd ed. Facts on File, 2010.
[2] WHO, Global Atlas on Cardiovascular Disease Prevention and Control, 1st ed. World Health Organization, 2012.
[3] Phibbs, B., The Human Heart: A Basic Guide to Heart Disease, Second. Lippincott Williams & Wilkins, 2007.
[4] Setiawan, N. A., Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System, Universiti Teknologi Petronas, 2009.
[5] Khemphila, A. & Boonjing, V., Heart Disease Classification Using Neural Network and Feature Selection, 2011 21st International Conference on Systems Engineering (ICSENG), pp. 406–409, 2011.
[6] Pal, D., Mandal, K. M., Pal, S., Sarkar, D. & Chakraborty, C., Fuzzy expert system approach for coronary artery disease screening using clinical parameters, Knowl.-Based Syst., vol. 36, pp. 162–174, Dec. 2012.
[7] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B. & Sani, Z. A., A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed, 2013.
[8] Capparuccia, R., De Leone, R., and Marchitto, E., Integrating support vector machines and neural networks, Neural Netw., vol. 20, no. 5, pp. 590–597, Jul. 2007.
[9] Negnevitsky, M., Artificial Intelligence: A Guide to Intelligent Systems, 2nd ed. Addison-Wesley, 2004.
[10] Witten, I. H. & Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*, 2nd ed. Morgan Kaufmann, 2005.

[11] Chu, N., Ma, L., Li, J., Liu, P. & Zhou, Y., Rough set based feature selection for improved differentiation of traditional Chinese medical data, *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, vol. 6, pp. 2667–2672, 2010.

[12] Babaoglu, İ., Findik, O. & Ulker, E., A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine, *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3177–3183, Apr. 2010.

[13] Shilaskar, S. & Ghatol, A., Feature selection for medical diagnosis: Evaluation for cardiovascular diseases, *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[14] Nahar, J., Imam, T., Tickle, K. S., and Chen, Y.-P. P., Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, *Expert Syst. Appl.*, vol. 40, no. 1, pp. 96–104, Jan. 2013.

[15] Guan, D., Yuan, W., Jin, Z., and Lee, S., Undiagnosed samples aided rough set feature selection for medical data, *2012 2nd IEEE International Conference on Parallel Distributed and Grid Computing (PDGC)*, pp. 639–644, 2012.

[16] UCI, Heart disease dataset, Online. http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleve.mod.

[17] Platt, J. C., Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, *Advances In Kernel Methods - Support Vector Learning*, 1998.

[18] Wu, X. & Kumar, V., *The top ten algorithms in data mining*. Boca Raton: CRC Press, 2009.

[19] Freund, Y. & Schapire, R. E., *Experiments with a New Boosting Algorithm*. 1996.

[20] Walpole, R. E., Myers, R. H. & Ye, K. E., *Probability & statistics for engineers & scientists*. Boston: Prentice Hall, 2012.