

IMPROVING CLASSIFICATION ACCURACY FOR NON-COMMUNICABLE DISEASE PREDICTION MODEL BASED ON SUPPORT VECTOR MACHINE

Mohd. Khanapi Abd. Ghani, Daniel Hartono Sutanto*

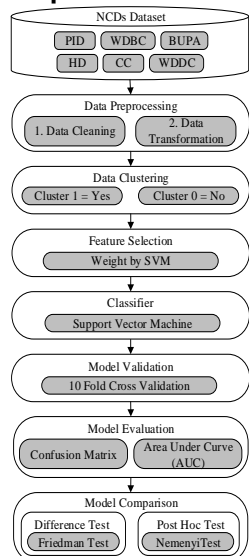
Biomedical Computing and Engineering Technologies (BIOCORE) Applied Research Group, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Article history

Received
15 May 2015
Received in revised form
1 July 2015
Accepted
11 August 2015

*Corresponding author
P031220009@student.utem.edu.my

Graphical abstract



Abstract

Over recent years, Non-communicable Disease (NCDs) is the high mortality rate in worldwide likely diabetes mellitus, cardiovascular diseases, liver and cancers. NCDs prediction model have problems such as redundant data, missing data, imbalance dataset and irrelevant attribute. This paper proposes a novel NCDs prediction model to improve accuracy. Our model comprises k-means as clustering technique, Weight by SVM as feature selection technique and Support Vector Machine as classifier technique. The result shows that k-means + weight SVM + SVM improved the classification accuracy on most of all NCDs dataset (accuracy; AUC), likely Pima Indian Dataset (99.52; 0.999), Breast Cancer Diagnosis Dataset (98.85; 1.000), Breast Cancer Biopsy Dataset (97.71; 0.998), Colon Cancer (99.41; 1.000), ECG (98.33; 1.000), Liver Disorder (99.13; 0.998). The significant different performed by k-means + weight by SVM + SVM. In the time to come, we are expecting to better accuracy rate with another classifier such as Neural Network.

Keywords: Prediction, non-communicable disease, data mining, feature selection, classification, k-means, weight by SVM, support vector machine

© 2015 Penerbit UTM Press. All rights reserved

1.0 INTRODUCTION

1.1 Overview

Over recent years, Non-communicable Diseases (NCDs) are leading causes of death and disability worldwide. NCDs also known as chronic diseases are a long-lasting condition that can be manipulated, but could not be healed immediately. Top three main types of NCDs are diabetes mellitus, cardiovascular diseases and cancers [1]. There are some aspect affects the quality of health care. Firstly, inequity of diagnosis of NCDs due to discrepancy numbers between patients and doctors [2],[3], [4].

1.2 Related Work

Guyon has been reviewed feature selection [5]. Feature selection has been an active and fruitful field of research and development for decades in machine learning and data mining [6]. It has proven in both theory and effective practice in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learning results. Noisy data detected in diabetes dataset, and most of NCDs dataset has irrelevant attribute.

Most of researchers used clustering, feature selection, or both of them for handling the NCDs problems. Patil used pre-processing technique to delete some instance and used K-means to handle

the noisy class, the classification accuracy shown at 92.38 to predict PIMA dataset [7]. Gurbuz used adaptive SVM to classify Pima, Breast Cancer, Liver, and the accuracy shown 97.39, 99.51, 84.63, respectively[8]. Anirundha applied k-means as clustering technique and Genetic Algorithm as wrapper feature selection to predict PIMA dataset, the accuracy shown that 97.86[9]. However, there is a chance to improve classification accuracy for NCDs dataset.

In this paper, we found the noisy problem and irrelevant attribute haven't handled yet while using SVM classifier in the same time. By other hand, the noisy problem is handled by clustering technique, k-means. Secondly, irrelevant attribute is resolved by using a feature selection technique, attribute weighting by SVM. The classification section used SVM classifier. Finally, the hybrid methods using k-means, weight by SVM and SVM classifier will be expected able to improve classification accuracy. The propose model shown in section 2.

2.0 METHODOLOGY

2.1 A Propose NCDs Prediction Model

This section draw the propose model for NCDs prediction based on SVM classifier. The proposed model is indicated in Figure 1.

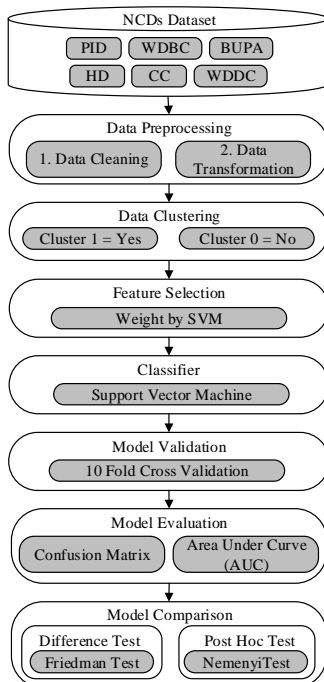


Figure 1 Non-Communicable Disease Prediction Model Based On Support Vector Machine

2.2 NCDs Dataset

NCDs datasets have been picked up from internet repositories, primarily from the UCI Machine Learning Repository. This research used 6 secondary datasets, that consist of diabetes, heart, and cancer datasets (Table 1).

Table 1 Dataset Detail

Researcher	Abbr.	Instance	Attribute	Class	Task
[10]–[13]	PID	768	8	2	Classification
[14]–[16]	WBBC	699	12	2	
[10], [14], [16]	ECG	132	12	2	
[17]–[18]	BUPA	345	6	2	
[19]	CC	1858	17	2	
[10], [16]	WDBC	569	32	2	

2.2.1 Pima Indian Diabetes

This data set includes a total of 768 instances depicted by 8 attributes and a predictive class. Out of 768 instances, 268 instances belong to class '1' which indicate that diabetic cases and 500 instances belong to class '0' means non diabetic cases i.e. they are healthy persons. Most of the cases contain missing values[20]. Number of missing values corresponding to each attribute in the data set is shown in Table 9.

2.2.2 Wisconsin Breast Cancer

The data set contains total 699 instances described by 9 attributes and a predictive class. The attribute values for all 9 attributes lie from 1 to 10. The class attribute has only two categories, namely benign and malignant. Class 'malignant' has 241 instances and class 'benign' consist of 458 instances. This data set consists of missing values[21]. The number of missing values with their attribute name is mentioned in Table 9.

2.2.3 Echocardiogram (ECG)

All the patients suffered heart attacks at some point in the past. The patients are however alive and some are not. The survival variables and still-alive variables, when taken together, show whether a patient survived following the heart attack [22]. The past researchers addressed problem to predict from the other variables whether or not the patient will survive at least one year. The difficult part is correctly predicting that the patient will not survive.

2.2.4 BUPA Liver Disorder

The BUPA Liver Disorders dataset consists of 345 male samples, each with 6 features concerning with the patients' biological markers, the amount of daily alcoholic beverage consumption, and the class attribute (presence of liver disorders). Those biological markers include the mean corpuscular erythrocyte

volume (MCV), carbohydrate-deficient transferrin (CDT), gamma-glutamyltransferase (GGT), total plasma homocysteine, and folate. The BUPA dataset contains 200 positive cases and 145 negative cases. In order to reduce the bias of scale, all the entries of all attributes in all datasets were linearly scaled into the interval of [0,1].

2.2.5 Colon Cancer

The data collection from one of the first successful trials of adjuvant chemotherapy for colon cancer [23]. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic chemotherapy agent. The output consists of two records per person, one record of recurring and one record for death.

2.2.6 Wisconsin Diagnostic Breast Cancer

Ten real-valued features are computed for each cell nucleus such as radius attribute, texture attributes, perimeter attributes, area attributes, compactness attributes, concavity attributes, concave points attribute, symmetry attributes, and fractal dimension attributes [21]. The mean, standard error, and "worst" or larger of these features were computed for each image, resulting in 30 features. The result is predicting diagnosis: B = benign or M = malignant and data sets are linearly separable using all 30 input features.

2.3 Data Preprocessing

Data processing is served as the raw dataset obtained may be noisy, irrelevant, incomplete and inconsistent. Initially, the dataset is preprocessed to remove noise points and missing values and then the data are normalized using z-score normalization.

In order to improve the accuracy of classification, the data preprocessing is needed to be done. A preliminary analysis of Pima Dataset indicates missing data. The number of missing values for the feature serum-insulin and triceps skin fold are very high (374 and 227, respectively from 768 instances).

2.3.1 Data Cleansing

The missing values of the data set that are considered for the experiment are denoted with the value zero. All the tuples that result in the value zero are removed. For Type-2 diabetes Pima Indians dataset, it is noticed that some attributes like plasma glucose have a value as zero. As no human can have that low count, it is removed so as not to affect the quality of the result.

2.3.2 Data Transformation

Some algorithms are sensitive to the scale of data. If you have one attribute whose range spans millions (of dollars, for example) while another attribute is in a few tens, then the larger scale attribute will influence the outcome. In order to eliminate or minimize such bias,

we must normalize the data. Normalization implies transforming all the attributes to a common range. This is easily achieved by dividing each attribute by its largest value, for instance. This is called Range Normalization. Another way to normalize is to calculate the difference between each attribute value and the mean value of the attribute and dividing by the standard deviation of the attribute. This is called a z-score normalization. In any such situations, data type transformations are required. The cleaned data are now normalized by using z-score normalization as given by Equation 1. This is done so that during classification or clustering the attributes may be scaled to fall within the given range of values and to generalize their values.

$$v' = \frac{v - \mu}{\sigma} \quad (1)$$

Where v' is the normalized value, v is the experimental value, μ is the mean and σ is the standard deviation [2].

2.4 Data Clustering

The clustering technique is k-means clustering to remove the outliers. As the experimental datasets have two classes the number of clusters used in the proposed method is two ($k=2$). One of the most used clustering algorithms was first described by MacQueen (1967) [24]. It was designed to cluster numerical data in which each cluster has a center called the mean. Let D be a data set with n instances, and let C_1, C_2, \dots, C_k be the k disjoint clusters of D . Then the error function is defined as

$$E = \sum_{i=1}^k \sum_{x \in C_i} d(x, \mu(C_i)) \quad (2)$$

where $\mu(C_i)$ is the centroid of cluster C_i , $d(x, \mu(C_i))$ denotes the distance between x and $\mu(C_i)$, and a typical choice of which is the Euclidean distance. Where D represents the Data set, k is number of Clusters, d is the dimensions, and C_i is the i th cluster. {Initialization Phase}

1: (C_1, C_2, \dots, C_k) = initial partition of D . {Iteration Phase}

2: repeat

3: d_{ij} = distance between case i and cluster j ;

4: $n_i = \arg \min_j d_{ij}$;

5: Assign case i to cluster n_i ;

6: Recompute the cluster means of any changed clusters above;

7: until no further changes of cluster membership occur in a complete iteration.

The k-means algorithm can be divided into two phases: the initialization phase and the iteration phase. In the initialization phase, the algorithm randomly assigns the cases into k clusters. In the iteration phase, the algorithm computes the distance

between each case and each cluster and assigns the case to the nearest cluster.

2.5 Attribute Weighting by SVM

Feature selection plays a very significant role in the success of the system in fields like pattern recognition and data mining. Feature selection provides a small but more distinguishing subset compared to the starting data, selecting the distinguishing features from a set of features and eliminating the irrelevant ones. Our goal is to reduce the dimension of the data by finding a small set of important features that can give a good classification performance. This results in both reduced processing time and increased classification accuracy [25]. Feature selection algorithms are grouped into randomized, exponential and sequential algorithms.

Weight by SVM [24] has purpose for retaining the highest weighted features in the normal has been independently derived in a somewhat different context in [10]. The idea is to consider the feature important if it significantly influences the width of the margin of the resulting hyper-plane; this margin is inversely proportional to $\|w\|$, the length of w . Since $w = \sum_i a_i x_i$ for a linear SVM model, one can regard $\|w\|^2$ as a function of the training vectors x_1, \dots, x_i where $x_i = (x_{i1}, \dots, x_{id})$, and thus evaluate the influence of feature j on $\|w\|^2$ by looking at absolute values of partial derivatives of $\|w\|^2$ with respect to x_{ij} . (Of course this disregards the fact that if the training vectors change, the values of the multipliers a_i would also change. Nevertheless, the approach seems appealing.) For the linear kernel, it turns out that

$$\sum_i |\partial \|w\|^2 / \partial x_{ij}| = k|w_j| \tag{3}$$

where the sum is over support vectors and k is a constant independent of j . Thus the features with higher $|w_j|$ are more influential in determining the width of the margin. The same reasoning applies when a non-linear kernel is used because $\|w\|^2$ can still be expressed using only the training vectors x_i and the kernel function.

2.6 Support Vector Machine Classifier

The well-known classification algorithm is Support vector machines (SVM). SVM is a new pattern recognition tool theoretically founded on Vapnik's statistical learning theory [26]. Support vector machines, originally designed for binary classification, employ supervised learning to find the optimal separating hyper plane between the two groups of data. Having found such a plane, support vector machines can then predict the classification of an unlabeled example by asking on which side of the separating plane the example lies. Support vector machine acts as a linear classifier in a high dimensional feature space originated by a projection of the original input space, the resulting classifier is in

general non-linear in the input space and it achieves good generalization performances by maximizing the margin between the two classes. In the following, this research give a short outline of construction of support vector machine. Consider a set of training examples as follows:

$$\{(x_i y_i) \mid x_i \in R^n, y_i \in \{+1, -1\}; i = 1, 2, \dots, m, \tag{4}$$

where the x_i are real n -dimensional pattern vectors and the y_i are dichotomous labels. Support vector machine maps the pattern vectors $x \in R^n$ into a possibly higher dimensional feature space ($z = \phi(x)$) and construct an optimal hyperplane $w \cdot z + b = 0$ in feature space to separate examples from the two classes. For support vector machine with L1 soft-margin formulation, this is done by solving the primal optimization problem as follows:

$$\text{Min} \frac{1}{2} \|w\| + C \sum_{i=1}^m \epsilon_i \text{ s.t. } y_i (w \cdot z_i + b) \geq 1 - \epsilon_i x \tag{5}$$

$$\epsilon_i \geq 0, i = 1, 2, \dots, m$$

where C is a regularization parameter used to decide a tradeoff between the training error and the margin, and $\epsilon_i (i = 1, 2, \dots, m)$ are slack variables. The above problem is computationally solved using the solution of its dual form:

$$\text{Max}_\alpha \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{j=1}^m \alpha_i \alpha_j y_i y_j k(x_i, y_i) \tag{6}$$

$$\text{s.t. } \sum_{i=1}^m \alpha_i y_i = 0; 0 \leq \alpha_i \leq C, i = 1, 2, \dots, m,$$

where $k(x_i, y_i) = \phi(x_i) \cdot \phi(x_j)$ is the kernel function that implicitly defines a mapping ϕ . The resulting decision function is:

$$f(x) = \text{sgn} \left\{ \sum_{i=1}^m \alpha_i y_i k(x_i, x) + b \right\} \tag{7}$$

All kernel functions have to fulfill Mercer theorem, however, the most commonly used kernel functions are polynomial kernel and radial basis function kernel, respectively.

$$k(x_i, x_j) = (a(x_i, x_j) + b)^d, \tag{8}$$

$$k(x_i, x_j) = \exp(-g \|x_i, x_j\|^2), \tag{9}$$

Support vector machines differ from discriminant analysis in two significant ways. First, the feature space of a classification problem is not assumed to be linearly separable. Rather, a nonlinear mapping function (also called a kernel function) is used to represent the data in higher dimensions where the boundary between classes is assumed to be linear [27]. Second, the boundary is represented by support vector machines instead of a single boundary. Support vectors run through the sample patterns which are the most difficult to classify, thus the sample

patterns that are closest to the actual boundary [27]. Over fitting is prevented by specifying a maximum margin that separates the hyper plane from the classes. Samples which violate this margin are penalized. The size of the penalty is a parameter often referred to as C [28], [29].

2.7 Model Validation

This research use a stratified 10-fold cross-validation for learning and testing data. This means that it divides the training data into 10 equal parts and then perform the learning process 10 times. It takes another part of dataset for testing and used the remaining nine parts for learning. Then, it calculated the average values and the deviation values from the ten different testing results. It employs the stratified 10-fold cross validation, because this method has become the standard and state-of-the-art validation method in practical terms. Some tests have also shown that the use of stratification improves results slightly [30].






2.8 Model Evaluation

The performance of classification accuracy is measured by confusion matrix, shown in Eq.10.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

This research applies Area Under Curve (AUC) as an accuracy indicator in our experiments to evaluate the performance of classification algorithm. AUC is area under ROC curve. In some research, Lessmann *et al.* [31] and Li *et al.* [17] stated the use of the AUC to improve cross study comparability. The AUC has benefit to improve convergence across empirical experiments significantly, because it separates predictive performance from operating conditions, and represents a general measure of predictive. A rough guide for classifying the accuracy of a diagnostic test using AUC is the traditional system, presented by Belle [32]. In the proposed framework, this research added the symbols for easier interpretation AUC (Table 2).

Table 2 AUC Evaluation

AUC	Classification	Symbol
0.90 - 1.00	excellent	
0.80 - 0.90	good	
0.70 - 0.80	fair	
0.60 - 0.70	poor	
< 0.60	failure	

2.9 Model Comparison

In comparison test, there are three families of statistical tests that can be used for benchmarking two or more classifiers over multiple datasets:

1. Parametric tests (the paired t-test and ANOVA), non-parametric tests (the Wilcoxon and the Friedman test)
2. The non-parametric test that assumes no commensurability of the results (sign test).

Demsar suggests the Friedman test for multiple benchmark classifiers, which relies on less restrictive assumptions [69]. Based on this recommendation, the Friedman test is applied to compare the AUCs in different classifiers. The Friedman test is calculated on the average ranked (R) performances of the classification algorithms on each dataset.

Let r_j^i be the rank of the j -th of C algorithms on the i -th of D datasets. The Friedman test has aim to compare the average ranks of algorithm $R_j = \frac{1}{D} \sum_{i=1}^D r_j^i$. Under the null-hypothesis, which states that all the algorithms are equivalent and so their ranks R_j should be fair. The statistic of Friedman is calculated as follows, and distributed according to χ_F^2 with $C - 1$ degrees of freedom, when variable D and C are big enough.

$$\chi_F^2 = \frac{12D}{C(C+1)} \left[\sum_j R_j^2 - \frac{C(C+1)^2}{4} \right] \tag{24}$$

If the null-hypothesis is rejected, it can be proceeded with a post-hoc test. When all classifiers are compared to each other, the Nemenyi test should be applied. Two classifiers have significantly different performance if the corresponding average ranks differ by at least the critical difference, shown by

$$CD = q_\alpha \sqrt{\frac{C(C+1)}{D}} \tag{25}$$

where critical values q_α are based on the studentized range statistic.

2.10 Experimental Setting

In this research, the experiment equipped with infrastructure consists RapidMiner Toolkit and XLSTAT. Rapidminer is an open-source system consisting of a number of data mining algorithms to automatically analyze a large data collection and extract useful knowledge[33], it can be used for analysis and modeling on diabetes prediction as well [34]. The XLSTAT statistical analysis add-in offers a wide variety of functions to enhance the analytical capabilities of Excel, making it the ideal tool for your everyday data analysis and statistics requirements[35]. The parameter should be adjusted to achieve the optimal performance and optimal accuracy for prediction model, rapidminer setting showed in Table 3. The hardware used CPU: HP Z420 Workstation, Processor: Intel® Xeon® CPU E5-1603 @ 2.80 GHz, RAM: 8,00 GB, and OS: Windows 7 Professional 64-bit Service Pack 1.

Table 3 Rapidminer Setting

Section	Method	Item	Detail
Clustering	k-means	K	2 class
		Max run	10
		Max optimization	100
		Measure type	
		Divergence	
Feature Selection	wSVM		
Classification	SVM	Type	C-SVC
		Kernel	Linier
		C	0.0
		Cache	80
		Epsilon	0.5

3.0 RESULTS AND DISCUSSION

In result section, we acknowledged the result of prediction model from 6 NCDs dataset, the detail shown at table 4-9.

Table 4 Result on Pima Indian Dataset

Contributor	Method	Accuracy	AUC
Lukka [14]	Sim	75.29	0.762
	Sim+F1	75.84	0.703
	Sim+F2	75.97	0.667
Seera [36]	FMM	69.28	0.661
	FMM-CART	71.35	0.683
	FMM-CART-RF	78.39	0.732
Patil [7]	k-means+C45	92.38	0.824
Zhu [13]	MFWX+k-NN	93.50	0.880
Gurbuz [8]	Adaptive SVM	97.39	0.972
Aninrudha [9]	k-means+GAFS+SVM	97.86	0.947
	k-means+GAFS+NB	97.86	0.947
	k-means+GA+DT	94.75	0.935
	k-means+GAFS+k-NN	97.47	0.865
Yilmaz [15]	mk-means+SVM	96.71	0.900
This study	SVM	76.69	0.823
	k-means + SVM	95.89	0.920
	k-means+w-SVM+SVM	99.52	0.999

Table 5 Result on Breast Cancer Diagnosis Dataset

Contributor	Method	Accuracy	AUC
Seera [36]	FMM	95.26	0.961
	FMM-CART	95.71	0.973
	FMM-CART-RF	98.84	0.987
Gurbuz [8]	Adaptive SVM	99.51	0.991
This study	SVM	97.54	0.992
	k-means + SVM	98.00	0.995
	k-means+w-SVM+SVM	98.85	1.000

Table 6 Result on Breast Cancer Biopsy Dataset

Contributor	Method	Accuracy	AUC
Belciug [16]	B-MLP	81.14	-
This study	SVM	95.34	0.958
	k-means + SVM	96.67	0.960
	k-means+w-SVM+SVM	97.71	0.998

Table 7 Result on Colon Cancer

Contributor	Method	Accuracy	AUC
Abedini [37]	GRD-XCS + SVM	-	0.87
This study	SVM	54.46	0.597
	k-means + SVM	97.85	0.960
	k-means+w-SVM+SVM	99.41	1.000

Table 8 Result on ECG

Contributor	Method	Accuracy	AUC
Belciug [16]	B-MLP	79.04	-
This study	SVM	95.71	1.000
	k-means + SVM	96.67	1.000
	k-means+w-SVM+SVM	98.33	1.000

Table 9 Result on Liver Disorder

Contributor	Method	Accuracy	AUC
Seera [36]	FMM	67.25	0.671
	FMM-CART	92.61	0.917
	FMM-CART-RF	95.01	0.955
Gurbuz [8]	Adaptive SVM	84.63	0.900
This study	SVM	68.65	0.713
	k-means + SVM	96.85	0.900
	k-means+w-SVM+SVM	99.13	0.998

In discussion section, we conclude that most of NCDs dataset have accuracy more than 98% and AUC more than 0.99. Regarding model evaluation based on result of AUC, our proposed model improved AUC in Table 5. Meanwhile, in Table 6 showed accuracy below 98%, it caused by more than 30 features as attributes. The one of contribution to data mining is hybrid techniques [38], meanwhile the contribution of this work is hybrid between clustering, feature selection and classification technique using SVM for NCDs prediction model. The result showed that our proposed model improves the accuracy of the prediction model. The NCDs prediction model focused on high accuracy. The model has three steps, furthermore SVM classifier has less tedious because the number of attributes and missing value have been eliminated.

Statistical analysis should be performed to know whether the results are significant or not. The optimal prediction model on each dataset is black highlighted. The highest Friedman score (R) is kmeans + w-SVM + SVM (PM4), followed by kmeans + SVM (PM3), SVM (PM2), and other models (PM1). In statistical significance testing, the P-value is the probability of achieving a test statistic at least as extreme as the one that was actually observed, hence assuming that the null hypothesis is true. Usually, the research is used "rejects the null hypothesis" when the P- value is less than the predetermined significance level (α), showing the observed result would be highly unlikely under the null hypothesis.

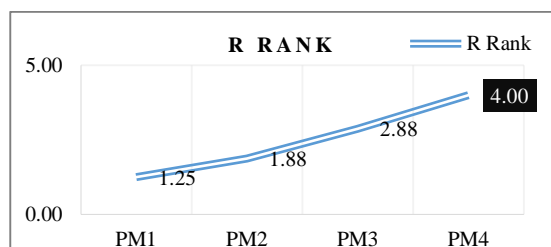


Figure 2 Result of R Rank of Prediction Model

In this research, it set the statistical significance level (α) to be 0.05. It means that there is a statistically significant difference, if P-value < 0.05. From the result of experiment, P-value is 0.0001, this is lower than the significance level α=0.05, hence one should reject the null hypothesis, and there is a significant difference, statistically. For detecting particular classifiers differ significantly, it can be used a Nemenyi post hoc test. Nemenyi post hoc has ability to calculates all pairwise benchmarks between different prediction model and find which performance differences of models exceed the critical difference. The results of the pairwise benchmarks of prediction model are shown in Table 10 with critical difference: 2.3452.

Table 10 Pairwise of Nemenyi Post Hoc Test

	PM1	PM2	PM3	PM4
PM1	0	0.6250	-1.0000	-2.1250
PM2	-0.6250	0	-1.6250	-2.7500
PM3	1.0000	1.6250	0	-1.1250
PM4	2.1250	2.7500	1.1250	0

P-value results of Nemenyi post hoc test are shown in Table 11. P-value < 0.05 results is highlighted with black print, furthermore there is a statistically significant difference between two classification algorithms, in a column and a row.

Table 11 P-value of Nemenyi Post Hoc Test

	PM1	PM2	PM3	PM4
PM1	1	0.9030	0.6923	0.0918
PM2	0.9030	1	0.2829	0.0138
PM3	0.6923	0.2829	1	0.6062
PM4	0.0918	0.0138	0.6062	1

As shown in Table 12, PM4 outperforms other models in most NCDs datasets. In terms of R value (Figure 2) and AUC mean (M) (Figure 3), PM4 also has the highest value, followed by PM1, PM2 and PM2.

From P-value analysis (Table 12), there is a significant difference PM3 and PM4 compare to 6 datasets. Significant difference table resulted by Nemenyi post hoc test shown in Table 12.

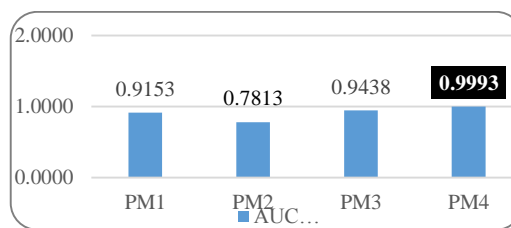


Figure 3 Result of AUC Mean (M) of Prediction Models

Table 12 Significant Differences of Nemenyi Post Hoc Test

	PM1	PM2	PM3	PM4
PM1	N	N	N	N
PM2	N	N	N	Y
PM3	N	N	N	N
PM4	N	Y	N	N

4.0 CONCLUSION

Non-communicable Disease (NCDs) is the high mortality rate in worldwide likely diabetes mellitus, cardiovascular diseases, liver and cancers. This paper proposes a novel NCDs prediction model to improve accuracy. Our model comprises k-means as clustering technique, Weight by SVM as feature selection technique and Support Vector Machine as classifier technique.

The result shows that k-means + weight SVM + SVM improved the classification accuracy on most of all NCDs dataset (accuracy; AUC), likely Pima Indian Dataset (99.52; 0.999), Breast Cancer Diagnosis Dataset (98.85; 1.000), Breast Cancer Biopsy Dataset (97.71; 0.998), Colon Cancer (99.41; 1.000), ECG (98.33; 1.000), Liver Disorder (99.13; 0.998). The significant different performed by k-means + weight by SVM + SVM. In the time to come, we are expecting to better accuracy rate with another classifier such as Neural Network.

Acknowledgement

This work was supported in part by a grant from LPDP Minister of Finance of Indonesia No. Kep56/LPDP/2014.

References

- [1] WHO. 2010. Global Status Report on Noncommunicable Diseases.
- [2] M. K. A. Ghani, R. K. Bali, R. N. G. Naguib, I. M. Marshall, and N. S. Wickramasinghe. 2010. Critical Analysis of the Usage of Patient Demographic and Clinical Records During Doctor-Patient Consultations: A Malaysian Perspective. *Int. J. Healthc. Technol. Manag.* 11(1/2): 113.
- [3] D. H. Sutanto, N. S. Herman, and M. K. A. Ghani. 2014. Trend of Case Based Reasoning in Diagnosing Chronic Disease: A Review. *Adv. Sci. Lett.* 20(10): 1740-1744.
- [4] M. K. A. Ghani, R. K. Bali, R. N. G. Naguib, I. M. Marshall, and N. S. Wickramasinghe. 2008. Electronic Health Records Approaches and Challenges: A Comparison Between

- Malaysia and four East Asian countries. *Int. J. Electron. Healthc.* 4(1): 78.
- [5] I. Guyon. 2003. An Introduction to Variable and Feature Selection 1 Introduction. *J. Mach. Learn. Res.* 3: 1157-1182.
- [6] V. Bolón-Canedo, N. Sánchez-Marroño, and A. Alonso-Betanzos. 2013. A Review of Feature Selection Methods on Synthetic Data. *Knowl. Inf. Syst.* 34(3): 483-519.
- [7] B. M. Patil, R. C. Joshi, and D. Toshniwal. 2010. Hybrid Prediction Model for Type-2 Diabetic Patients. *Expert Syst. Appl.* 37(12): 8102-8108.
- [8] E. Gürbüz and E. Kılıç. 2014. A New Adaptive Support Vector Machine for Diagnosis of Diseases. *Expert Syst.* 31(5): 389-397.
- [9] R. C. Anirudha, R. Kannan, and N. Patil. 2015. Genetic Algorithm Based Wrapper Feature Selection on Hybrid Prediction Model for Analysis of High Dimensional Data.
- [10] L.-Y. Chuang, C.-H. Yang, K.-C. Wu, and C.-H. Yang. 2011. A Hybrid Feature Selection Method for DNA Microarray Data. *Comput. Biol. Med.* 41(4): 228-37.
- [11] M. A. Chikh, M. Saidi, and N. Settoui. 2012. Diagnosis of diabetes Diseases Using An Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-Nearest Neighbor. *J. Med. Syst.* 36(5): 2721-9.
- [12] F. Beloufa and M. a Chikh. 2013. Design of Fuzzy Classifier for Diabetes Disease Using Modified Artificial Bee Colony Algorithm. *Comput. Methods Programs Biomed.* 112(1): 92-103.
- [13] J. Zhu, Q. Xie, and K. Zheng. 2015. An Improved Early Detection Method of Type-2 Diabetes Mellitus Using Multiple Classifier System. *Inf. Sci. (Ny)*. 292: 1-14.
- [14] P. Luukka. 2011. Feature Selection Using Fuzzy Entropy Measures with Similarity Classifier. *Expert Syst. Appl.* 38(4): 4600-4607.
- [15] N. Yilmaz, O. Inan, and M. S. Uzer. 2014. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. *J. Med. Syst.* 38(5): 48.
- [16] S. Belciug and F. Gorunescu. 2014. Error-correction Learning for Artificial Neural Networks Using the Bayesian Paradigm. Application to Automated Medical Diagnosis. *J. Biomed. Inform.* 52: 329-37.
- [17] D.-C. Li, C.-W. Liu, and S. C. Hu. 2011. A Fuzzy-based Data Transformation for Feature Extraction to Increase Classification Performance with Small Medical Data Sets. *Artif. Intell. Med.* 52(1): 45-52.
- [18] Y. J. Fan and W. A. Chaovalitwongse. 2010. Optimizing Feature Selection to Improve Medical Diagnosis. *Ann. Oper. Res.* 174: 169-183.
- [19] P. Ganesh Kumar, T. Aruldoss Albert Victoire, P. Renukadevi, and D. Devaraj. 2012. Design of Fuzzy Expert System for Microarray Data Classification Using a Novel Genetic Swarm Algorithm. *Expert Syst. Appl.* 39(2): 1811-1821.
- [20] V. Sigillito. 1990. Pima Indians Diabetes Database. UCI Machine Learning Repository, National Institute of Diabetes and Digestive and Kidney Diseases.
- [21] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. 1992. Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository, University of Wisconsin Hospitals Madison, Wisconsin, USA.
- [22] S. Salzberg and Evin Kinney. 1988. Echocardiogram Data Set. UCI Machine Learning Repository, The Reed Institute, Miami.
- [23] J. a. Laurie, C. G. Moertel, T. R. Fleming, H. S. Wieand, J. E. Leigh, J. Rubin, G. W. McCormack, J. B. Gerstner, J. E. Krook, J. Malliard, D. I. Twito, R. F. Morton, L. K. Tschetter, and J. F. Barlow. 1989. Surgical Adjuvant Therapy of Large-Bowel Carcinoma: An Evaluation of Levamisole and Their Combination of Levamisole and Fluorouracil. *J. Clin. Oncol.* 7(10): 1447-1456.
- [24] J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 281-297.
- [25] G. Chandrashekar and F. Sahin. 2014. A Survey on Feature Selection Methods. *Comput. Electr. Eng.* 40(1): 16-28.
- [26] V. Vapnik, S. E. Golowich, and A. Smola. 1998. Support Vector Method for Function Approximation. *Regression Estimation, and Signal Processing*. 281-287.
- [27] R. Duda O., P. Hart E., and D. Stork G. 2000. *Pattern Classification*.
- [28] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. 2000. Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *Proc. Natl. Acad. Sci. U. S. A.* 97(1): 262-267.
- [29] N. Cristianini and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines*.
- [30] Ian H. Witten, E. Frank, and M. A. Hall. 2006. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition.
- [31] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch. 2008. Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings." *IEEE Trans. Softw. Eng.* 34(4): 485-496.
- [32] V. Van Belle and P. Lisboa. 2014. White Box Radial Basis Function Classifiers with Component Selection for Clinical Prediction Models. *Artif. Intell. Med.* 60(1): 53-64.
- [33] M. Hofmann and R. Klinkenberg. 2013. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press,
- [34] J. Han, J. C. Rodriguez, and M. Beheshti. 2008. Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner. *2008 Second Int. Conf. Futur. Gener. Commun. Netw.* 96-99.
- [35] T. Fahmy and A. Aubry. 1998. XLstat. In *Société Addinsoft SARL*. 40.
- [36] M. Seera and C. P. Lim. 2014. A Hybrid Intelligent System for Medical Data Classification. *Expert Syst. Appl.* 41(5): 2239-2249.
- [37] M. Abedini and M. Kirley. 2013. An Enhanced XCS Rule Discovery Module Using Feature Ranking. *Int. J. Mach. Learn. Cybern.* 4(3): 173-187.
- [38] J. H. and M. Kamber. 2006. *Data Mining Concepts and Techniques*.