VOL. 10, NO. 16, SEPTEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

IMPROVING CLASSIFICATION PERFORMANCE OF K-NEAREST NEIGHBOUR BY HYBRID CLUSTERING AND FEATURE SELECTION FOR NON-COMMUNICABLE DISEASE PREDICTION

Daniel Hartono Sutanto and Mohd. Khanapi Abd. Ghani
Applied Research Group, Biomedical Computing and Engineering Technologies (BIOCORE), Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
E-Mail: p031220009@student.utem.edu.my

ABSTRACT

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 redundancy data, missing data, noisy class 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 k-nearest neighbour as classifier technique. The result shows that k-means + weight by SVM + k-nn improved the classification accuracy on most of all NCDs dataset (accuracy; AUC), likely Pima Indian Dataset (96.82; 0.982), Breast Cancer Diagnosis Dataset (97.36; 0.997), Breast Cancer Biopsy Dataset (96.85; 0.994), Colon Cancer (99.41; 1.000), ECG (97.80; 1.000), Liver Disorder (97.97; 0.998).

Keywords: k-nearest neighbour, prediction, non-communicable disease, k-means, weight by SVM.

INTRODUCTION

Non-communicable Diseases (NCDs) are leading mortality rate and cause of death in worldwide. NCDs also known as chronic diseases are a long-lasting condition that can be controlled, but not cured. Top three main types of NCDs are diabetes mellitus, cardiovascular diseases and cancers [1]. In data mining, a method that is used to extract the hidden knowledge from large amounts of data, is commonly used [2]. To enhance non-communicable disease prediction model, data mining is the prediction technique to diagnose disease [3]. For data mining task, classification is the most widely used method. Classification algorithms are supervised methods that uncover the hidden relationship between the target class and the independent variables [11]. Supervised learning algorithms allow labels to be assigned to the observations so that new data can be classified based on the training data [6]. Examples of classification tasks are image and pattern recognition, medical diagnosis [4], [5]. However, the prediction model using classification algorithm for non-communicable disease is needed to improve the quality of healthcare [6].

Related works

Feature selection has known as research and development in machine learning and data mining for decades [5]. Guyon has been reviewed advantages of feature selection such as enhance learning efficiency, increase predictive accuracy, and reduce complexity of learned results already proven in theory and practice [6]. Noisy data detected in diabetes dataset, and most of NCDs dataset has irrelevant attribute [7].

In previous research, Patil applied pre-processing technique to delete some attribute and used k-means only to handle the noisy class in PIMA dataset, the classification accuracy shown at 92.38 [8]. Gurbuz used adaptive NN to classify Pima, Breast Cancer, Liver, and the accuracy shown 97.39, 99.51, 84.63, respectively [9]. Anirundha applied K-means as clustering technique and Genetic Algorithm as wrapper feature selection to predict PIMA dataset, the accuracy shown that 97.86 [10]. There is chance to improve classification accuracy for NCDs dataset.

Based on literature review, it was found that the noisy dataset and irrelevant attribute haven't solved coincide yet. By other hand, the noisy dataset is handled by clustering technique, k-means. Moreover, irrelevant attribute is resolved by using a feature selection technique, attribute weighting by SVM. The classification section used k-nn classifier. Han and Kamber (2009) stated one of contribution to data mining is hybrid the methods [11]. The noisy class is reduced by clustering technique, k-means. Meanwhile, irrelevant attribute is solved by using feature selection technique, attribute weighting by SVM. The expectation of this research is hybrid three methods will improve classification accuracy. The propose model shown at section 2.

METHODOLOGY

A propose NCDs prediction model

This section draw the propose model for NCDs prediction based on k-nn classifier.

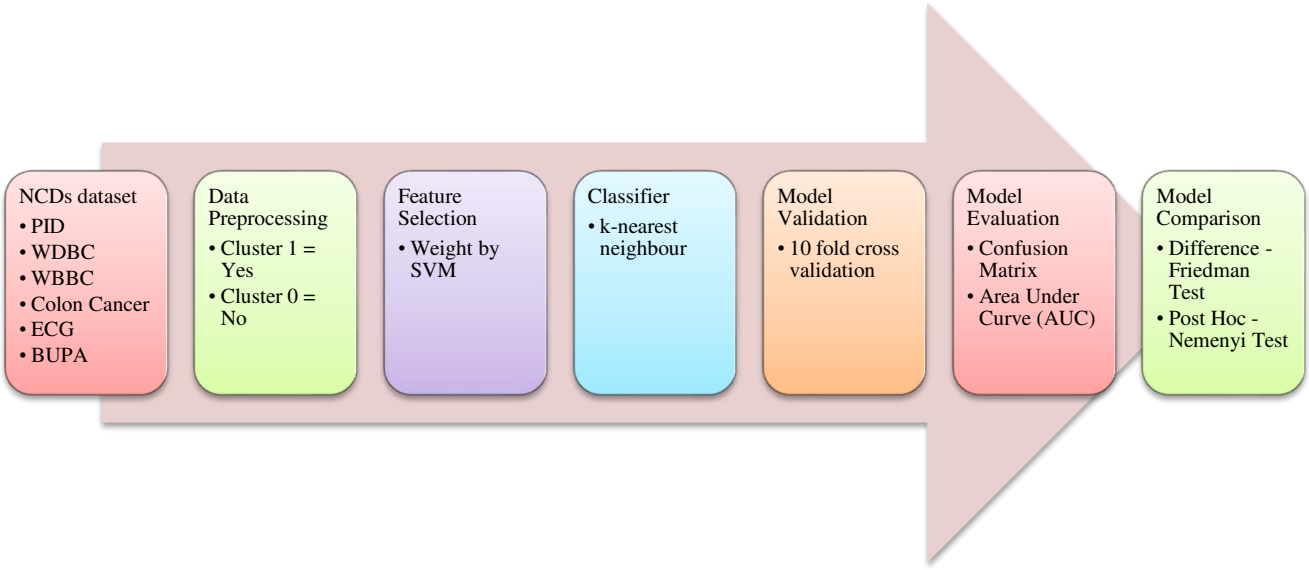

Figure-1. Non-communicable disease prediction model based on k-nearest neighbour.

NCDs dataset

NCDs datasets have been collected from internet repositories, mainly 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.

NCDs dataset	Abbr.	Contributor	Researcher	Instance	Attribute	Class	Task
Pima Indian Diabetes	PID	[12]	[13]-[16]	768	8	2	Classification
Wisconsin Biopsy Breast Cancer	WBBC	[17]	[18]-[20]	699	12	2	
BUPA Liver Disorder	BUPA	[21]	[22], [23]	345	6	2	
Echocardiogram	ECG	[24]	[13], [18], [20]	132	12	2	
Colon Cancer	CC	[25]	[26]	1858	17	2	
Wisconsin Diagnostic Breast Cancer	WDBC	[17]	[13], [20]	569	32	2	

Data preprocessing

Data processing is done 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 is normalized using z-score normalization.

In order to improve the accuracy of classification, the data preprocessing is needed to be completed. 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 form 768 instances).

Data cleaning

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.

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 z-score normalization. In any such situations, data type transformations are required. The cleaned data is 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.

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) [27]. 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,..., C_k be the k disjoint clusters of D. Then the error function is defined as

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x_1 \mu(C_i)) \quad (2)$$

where $\mu(C_i)$ is the centroid of cluster $C_i, d(x_1 \mu(C_i))$ 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 ith cluster. {Initialization Phase}

a) : (C1, C2,..., Ck) = initial partition of D. {Iteration Phase}

b) : repeat

c) : dij = distance between case i and cluster j;

d) : ni = arg min1 6 j 6 k dij;

e) : Assign case i to cluster ni;

f) : Recompute the cluster means of any changed clusters above;

g) : 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.

Attribute weighting by SVM

Feature selection plays a very significant role for the success of the system in fields like pattern recognition and data mining. Feature selection provides a smaller 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 good classification performance. This results in both reduced processing time and increased classification accuracy. Feature selection algorithms are grouped into randomized, exponential and sequential algorithms.

Weight by SVM [28] 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 $x_1, \ldots, x_l$ where $x_i = (x_{i1,}, \ldots, x_{id})$, 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

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

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.

K-Nearest Neighbour

K-nearest neighbour (k-nn) is a well-known supervised learning algorithm for pattern recognition that first introduced by Fix and Hodges (1951), and is still one of the most popular nonparametric models for classification problems [29]. K-nearest neighbour assumes that observations which are close together are likely to have the same classification. The probability that a point x belongs to a class can be estimated by the proportion of training points in a specified neighbourhood of x that

belong to that class [29]. The point may either be classified by majority vote or by a similarity degree sum of the specified number (k) of nearest points. In majority voting, the number of points in the neighbourhood belonging to each class is counted, and the class to which the highest proportion of points belongs is the most likely classification of x. The similarity degree sum calculates a similarity score for each class based on the K-nearest points and classifies x into the class with the highest similarity score. Due to its lower sensitivity to outliers, majority voting is more commonly used than the similarity degree sum [30]. In this paper, majority voting is used for the data sets. In order to determine which points belong in the neighbourhood, the distances from x to all points in the training set must be calculated. Any distance function that specifies which of two points is closer to the sample point could be employed [29]. The most common distance metric used in K-nearest neighbour is the Euclidean distance [31]. The Euclidean distance between each test point ft and training set point fs, each with n attributes, is calculated using the equation:

$$d = [(f_{t1} - f_{s1})^2 + (f_{t2} - f_{s2})^2 + \cdots + (f_{tn} - f_{sn})^2]^{1/2} \quad (4)$$

In general the following steps are performed for the k-nearest neighbour model [32]:

a) Chosen of k value.
b) Distance calculation.
c) Distance sort in ascending order.
d) Finding k class values.
e) Finding dominant class

One challenge to use the k-nearest neighbour is to determine the optimal size of k, which acts as a smoothing parameter. A small k will not be sufficient to accurately estimate the population proportions around the test point [33]. A larger k will result in less variance in probability estimates but the risk of introducing more bias [31]. K should be large enough to minimize the probability of a non-Bayes decision, but small enough that the points included give an accurate estimate of the true class. Enas and Choi (1986) found that the optimal value of k depends upon the sample size and covariance structures in each population, as well as the proportions for each population in the total sample. For cases in which the differences in the covariance matrices and the difference between sample proportions were either small or both large, Enas and Choi (1986) found that the optimal k to be N3/8, where N is the number of samples in the training set. When there was a large difference between covariance matrices and a small difference between sample proportions, or vice versa, Enas and Choi (1986) determined N2/8 to be the optimal value of k.

In additional, when the boundaries between classes cannot be described as hyper-linear or hyper-conic, K-nearest neighbour performs better than the linear and quadratic discriminant functions. Enas and Choi (1986) found that the linear discriminant performs slightly better than k-nearest neighbour when population covariance matrices are equal, a condition that suggests a linear boundary. As the differences in the covariance matrices increases, k-nearest neighbour performs increasingly better than the linear discriminant function [33]. However, despite of all the advantages cited for the k-nearest neighbour models, they also have some disadvantages. K-nearest neighbour model cannot work well if large differences are present in the number of samples in each class. K-nearest neighbour provides poor information about the structure of the classes and of the relative importance of each variable in the classification. Furthermore, it does not allow a graphical representation of the results, and in the case of large number of samples, the computation can become excessively slow. In addition, K-nearest neighbour model much higher memory and processing requirements than other methods. All prototypes in the training set must be stored in memory and used to calculate the Euclidean distance from every test sample.

Model validation

This research uses a stratified 10-fold cross-validation for learning and testing data. This means that this research divides the training data into 10 equal parts and then perform the learning process 10 times. As shown in Table-3, each time, this research chose another part of dataset for testing and used the remaining nine parts for learning. After, this research calculated the average values and the deviation values from the ten different testing results. This research 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 [34].

Model evaluation

Classification accuracy is measured performance result likely confusion matrix, shown at (Table-2).

Table-2. Performance accuracy.

Parameter	Formula	
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	(5)
Sensitivity (TP Rate)	$\frac{TP}{TP + FN}$	(6)
Specificity (FP Rate)	$\frac{FP}{FP + TN}$	(7)
Positive Predictive Value (PPV)	$\frac{TP}{TP + FP}$	(8)
Negative Predictive Value (NPV)	$\frac{TN}{TN + FN}$	(9)

Table-3. 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	

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*. [35] and Li *et al*. [22] stated the use of the AUC to improve cross study comparability. The AUC has the 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 [36]. In the proposed framework, this research added the symbols for easier interpretation and understanding of AUC (Table-3).

Experimental setting

In this research, the experiment equipped with infrastructure consists Rapid Miner Toolkit and XLSTAT. Rapidminer is an open-source system composed of some data mining algorithms to analyze automatically a large data collection and extract useful knowledge [37]. It can be used for analysis and modeling of diabetes prediction as well [38]. 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 data analysis and statistics requirements [39]. The parameter of rapidminer 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-4. Rapidminer Setting.

Section	Method	Item	Detail
Clustering	k-means	K	2 class
		Max run	10
		Max optimization	100
		Measure type	n/a
		Divergence	n/a
Feature Selection	w-SVM	n/a	n/a
Classification	k-nn	k	10
		Measure type	Mixed Measure
		Mixed measure	Mixed Euclidean Distance

Experimental results

In this section, we acknowledged the result of prediction model from 6 NCDs dataset, the detail shown in Table-5.

Table-5. Prediction result using 6 NCDs dataset.

Ref	Method	PIMA		WDBC		WBBC		Colon		ECG		BUPA	
		Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
[18]	Sim	75.29	0.762										
	Sim+F1	75.84	0.703										
	Sim+F2	75.97	0.667										
[40]	FMM	69.28	0.661	95.26	0.961							95.26	0.961
	FMM-CART	71.35	0.683	95.71	0.973							95.71	0.973
	FMM-CART-RF	78.39	0.732	98.84	0.987							98.84	0.987
[9]	k-means+C45	92.38	0.824										
[20]	B-MLP					81.14				79.04			
[41]	GRD-XCS + SVM								0.87				
[16]	MFWX+NN	93.50	0.880										
[10]	Adaptive SVM	97.39	0.972	99.51	0.991							99.51	0.991
[11]	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+NN	97.47	0.865										
[19]	mk-means+SVM	96.71	0.900										
This study	k-nn	74.89	0.794	95.85	0.987	88.58	0.940	53.99	0.520	94.46	0.960	67.51	0.708
	k-means + k-nn	89.88	0.978	96.30	0.990	95.00	0.950	96.19	0.947	95.71	0.980	96.83	0.960
	k-means+ w-SVM+k-nn	96.82	0.982	97.36	0.997	96.85	0.994	99.41	1.000	97.80	1.000	97.97	0.998

The result showed that our proposed model improve the accuracy of prediction model. Most of NCDs dataset has accuracy more than 98% and AUC more than 0.98. The result also showed that the proposed model improves the accuracy of the prediction model. The NCDs prediction model focused on improving accuracy and comparing the previous research. The model has three steps; furthermore k-nn classifier has less tedious because the number of attributes and missing value has been eliminated.

Statistical analysis was performed to know whether the results are significant or not. This research took AUC as benchmark for every prediction model. The optimal prediction model on each dataset is black highlighted. In Figure-2, the highest Friedman score (R) is k-means + w-SVM + k-nn (proposed model: M4), followed by k-means + k-nn (PM3), existing optimal prediction model (M1), and k-nn classifier (M2).

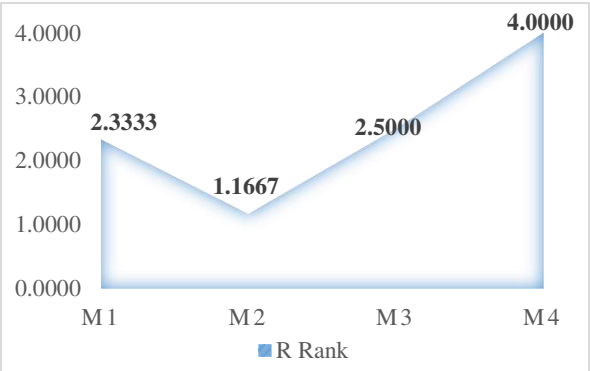

Figure-2. R Rank of prediction model.

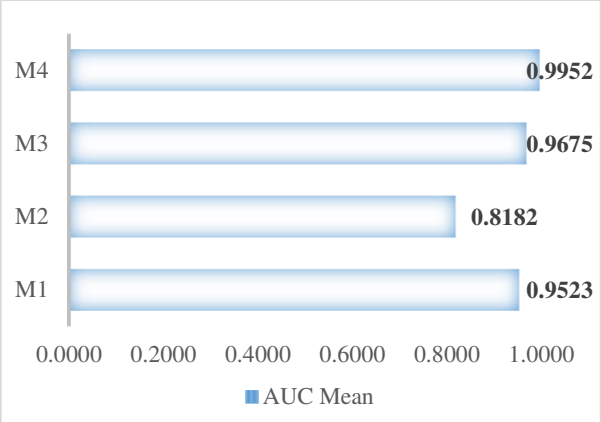

Figure-3. AUC Mean of prediction model.

In terms of R value (Figure-2) and AUC mean (M) (Figure-3), M4 (proposed model) also has the highest value, followed by M3, PM1, and PM2. The Friedman's test need to adjust the parameters, the details shown in Table-6.

Table-6. Friedman's test.

Q (Observed value)	**14.6000**
Q (Critical value)	7.8147
DF	3
p-value (Two-tailed)	0.0022
Alpha	0.05

Table-7. Pairwise differences.

	M1	**M2**	**M3**	**M4**
M1	0	1.1667	-0.1667	-1.6667
M2	-1.1667	0	-1.3333	-2.8333
M3	0.1667	1.3333	0	-1.5000
M4	1.6667	2.8333	1.5000	0

In this research, the statistical significance level (α) to was set 0.05. It means that there is a statistically significant difference, if P-value < 0.05. For detecting particular classifiers differ significantly, a Nemenyi post hoc test was applied. Nemenyi post hoc has the ability to calculate 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: 1.9149.

Table-8. P-values of Nemenyi post hoc test.

	M1	**M2**	**M3**	**M4**
M1	1	0.3986	0.9961	0.1136
M2	0.3986	1	0.2786	0.0008
M3	0.9961	0.2786	1	0.1833
M4	0.1136	0.0008	0.1833	1

Table-9. Significant differences.

	M1	**M2**	**M3**	**M4**
M1	No	No	No	No
M2	No	No	No	Yes
M3	No	No	No	No
M4	No	Yes	No	No

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.

From P-value analysis (Table-8), there is a significant difference between M4 (proposed model) and M2 (k-nn classifier), the rest is no significant different due to the AUC result almost 1.0. Significant difference of Nemenyi post hoc test shown in Table-9. The P-value result of Nemenyi post hoc test are featured in Table-8. P-value < 0.05 results is highlighted with black print, therefore there is a statistically significant difference between prediction model, in a column and a row. As shown in Table-9, PM4 outperforms other models in most NCDs datasets.

CONCLUSIONS

NCDs prediction model has been known to predict chronic disease. Meanwhile, there are some problem in NCDs dataset such as noisy data and irrelevant attributes, correspondingly. This paper proposes a novel NCDs prediction model to improve accuracy such as hybrid k-means as clustering technique, Weight SVM as feature selection technique and k-nearest neighbour as classifier technique. The result shows that k-means + weight by SVM + k-nn improved the classification accuracy on most of all NCDs dataset (accuracy; AUC), likely Pima Indian Dataset (96.82; 0.982), Breast Cancer Diagnosis Dataset (97.36; 0.997), Breast Cancer Biopsy Dataset (96.85; 0.994), Colon Cancer (99.41; 1.000), ECG (97.80; 1.000), Liver Disorder (97.97; 0.998). In future, we are able to improve accuracy rate with other classifiers such as Support Vector Machine and Neural Network.

ACKNOWLEDGMENT

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] J. Han, M. Kamber, and J. Pei. 2001. Data Mining Concepts and Techniques. 40(6).

[3] J. T. L. Wang, M. J. Zaki, H. T. T. Toivonen and D. Shasha. 2005. Data Mining in Bioinformatics.

[4] 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.

[5] 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.

[6] V. Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos. 2011. Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Syst. Appl. 38(5): 5947-5957.

[7] I. Guyon. 2003. An Introduction to Variable and Feature Selection 1 Introduction. J. Mach. Learn. Res. 3: 1157-1182.

[8] V. Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos. 2013. A review of feature selection methods on synthetic data. Knowl. Inf. Syst. 34(3): 483-519.

[9] 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.

[10] E. Gürbüz and E. Kılıç. 2014. A new adaptive support vector machine for diagnosis of diseases. Expert Syst. 31(5): 389-397.

[11] 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.

[12] V. Sigillito. 1990. Pima Indians Diabetes Database. UCI Machine Learning Repository, National Institute of Diabetes and Digestive and Kidney Diseases.

[13] 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.

[14] M. A. Chikh, M. Saidi, and N. Settouti. 2012. Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor. J. Med. Syst. 36(5): 2721-9.

[15] 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.

[16] 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.

[17] 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.

[18] P. Luukka. 2011. Feature selection using fuzzy entropy measures with similarity classifier. Expert Syst. Appl. 38(4): 4600-4607.

[19] 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.

[20] 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: 329-337.

[21] R. S. Forsyth. 1990. Liver Disorders Data Set. BUPA Medical Research Ltd., Nottingham. p. 1990.

[22] 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.

[23] Y.-J. Fan and W. A. Chaovalitwongse. 2010. Optimizing feature selection to improve medical diagnosis. Ann. Oper. Res. 174(1): 169-183.

[24] S. Salzberg and Evlin Kinney. 1988. Echocardiogram Data Set. UCI Machine Learning Repository, the Reed Institute, Miami.

[25] 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.

[26] 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.

[27] 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. pp. 281-297.

[28] Y.-W. Chang and C.-J. Lin. 2008. Feature ranking using linear svm. Causation Predict. Chall pp. 53-64.

[29] E. Fix and J. L. H. Jr. 1951. Discriminatory analysis-nonparametric discrimination: consistency properties. CALIFORNIA UNIV BERKELEY.

[30] W. A. Chaovalitwongse, Y. Fan, S. Member, and R. C. Sachdeo. 2007. On the Time Series K -Nearest Neighbor Classification of Abnormal Brain Activity. Seizure. 37(6): 1005-1016.

[31] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene. 2002. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. J. Risk Insur. 69(3): 373-421.

[32] T. Yildiz, S. Yildirim, Y. Doç and D. T. Altilar. 2008. Spam filtering with parallelized KNN algorithm. Akademik Bilisim. pp. 627-632.

[33] G. G. Enas and S. C. Choi. 1986. Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. Comput. Math. With Appl. 12(2): 235-244.

[34] ian H. Witten, E. Frank and M. A. Hall. 2006. Data Mining : Practical Machine Learning Tools and Techniques. 3rd Edition.

[35] 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.

[36] 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.

[37] M. Hofmann and R. Klinkenberg. 2013. Rapid Miner: Data mining use cases and business analytics applications. CRC Press.

[38] J. Han, J. C. Rodriguez and M. Beheshti. 2008. Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner. 2008 Second Int. Conf. Futur. Gener. Commun. Netw. pp. 96-99.

[39] T. Fahmy and A. Aubry. 1998. XLstat. In Société Addinsoft SARL. p. 40.

[40] M. Seera and C. P. Lim. 2014. A hybrid intelligent system for medical data classification. Expert Syst. Appl. 41(5): 2239-2249.

[41] M. Abedini and M. Kirley. 2013. An enhanced XCS rule discovery module using feature ranking. Int. J. Mach. Learn. Cybern. 4(3): 173-187.