A Non-symmetrical Weighted K-means with Rank-Based Artificial Bee Colony Algorithm for Medical Diagnosis

Jenn-Long Liu and Chung-Chih Li

***Abstract*—This study proposes a non-symmetrical weighted k-means (NSWKM) clustering algorithm to improve the accuracy of clustering result. The similarity distance of original k-means algorithm is modified by adding weights with a non-symmetrical form to the distance measurement. Namely, different weights for attributes are applied for clusters such that the contribution of attributes can be adjusted adaptively during the clustering process. In this work, the weights are given via an optimization process using a rank-based artificial bee colony (RABC) algorithm. Furthermore, the proposed NSWKM clustering algorithm combined with the RABC, termed NSWKM-RABC herein, is then applied to the medical diagnoses of five data sets of diseases, including breast cancer, cardiac disease, diabetes, liver disease and hepatitis, to evaluate the performance of the proposed algorithm.**

***Index Terms*—Data sets of diseases, medical diagnoses, non-symmetrical weighted k-means clustering algorithm, rank-based artificial bee colony algorithm.**

I. Introduction

Data mining technology is an effective tool to mining useful knowledge from the databases in hand, and it is a critical analysis step for achieving the knowledge-discovery in databases (KDD). Basically, the processes of KDD are getting data, choosing target data, preprocessing data, transforming data, discovering patterns/rules, and performing the optimal decision for action. According to the research work of [1] presented in 2008, the top ten data mining algorithms were C4.5, k-means, SVM (Support Vector Machine), Apriori, EM (Expectation-Maximization), PageRank (Google's Page Rank), AdaBoost (Adaptive Boosting), kNN (k-Nearest Neighbor), Naïve Bayes, and CART (Classification and Regression Trees). Detailed technologies and references of the ten algorithms were presented in [1]. Among the ten algorithms, k-means algorithm is an iterative method to divide a given data set to a user-specified number of clusters, k for example, by using the measurement of similarity distance. Each instance will be assigned to its belonging cluster with the smallest distance from the instance to assigned cluster centroid. Generally, the distance measurement is often using Euclidean or Manhattan distance with specifying the same weights, all equal to 1, for attributes to evaluate the distance. However, the contribution of attributes to clustering may be different due to the "curse of dimensionality" problem that significantly affects the classification accuracy [2]. Namely, the weighting values for attributes are always different with bias to clustering. In the literature [3]-[6], several works using weighting methods for similarity distance have been applied to pattern recognition problems. According, this work adopts a non-symmetrical weighted k-means (NSWKM) algorithm to address the deficiency of original k-means algorithm for enhancing the classification accuracy.

However, the contribution degree of weights to classification is hardly to known in prior with our limited knowledge. Recently, the evolutionary computation shows the promising development to obtain optimal solutions in a widely research domains. The recently developed artificial bee colony (ABC) algorithm is a swarm-based optimization with substantially powerful for finding optimal solutions by exploring new nectar sources and exploiting nectar sources discovered via artificial bees. Therefore, this work uses the ABC optimization method to achieve the optimums ofweights. The basic ABC algorithm, proposed by Karaboga *et al*. [7], [8], introduces three classes of artificial bees: employed, onlooker, and scout bees. Both the employed and onlooker bees perform the substantially exploitation and exploration processes by managing the capacity of nectar sources and discovering new candidates using local search around their current occupied food sources, whereas scout bees employ the exploration process by choosing new food sources randomly in the search space. Although the basic ABC can find out an optimum effectively, the solution search equation for the algorithm is poor at exploitation [9]. Therefore, this study modifies the basic ABC algorithm by including a rank-based selection mechanism in the onlooker bee phase and a modified abandoned rule in the scout bee phase to improve the searching ability of the algorithm. Accordingly, this work combines the NSWKM clustering algorithm and rank-based ABC (RABC) to the medical diagnosis of diseases. The combination of the proposed algorithm is named NSWKM-RABC in this work. Five data sets of diseases published on the web site of UCI repository [10] are evaluated by the proposed algorithm and also Naïve Bayes, C4.5, and k-means classifiers.

Manuscript received November 6, 2014; revised February 26, 2015. This work was supported in part by the Ministry of Technology and Science of Republic of China under Grant 103-2918-I-214-001.

J. L. Liu is with the Department of Information Management, I-Shou University, Kaohsiung, 84001 Taiwan (e-mail: jlliu@isu.edu.tw).

C. C. Li is with the School of Information Technology, Illinois State University, Normal, IL 61790 USA (e-mail: cli2@ilstu.edu).

II. Procedures of the Proposed Algorithm

A. Non-symmetrical Weighted K-means Clustering Algorithm

The k-means clustering algorithm usually computes the similarity distances for all instances first in order to partition

a given data set into k clusters. The generally used measurement of similarity distance is Minkowski distance which is obtained by computing the p-norm distance between two points x_i and y_i in n-dimensional space as below.

$$\left(\sum_{i=1}^{n} \left| x_i - y_i \right|^{p} \right)^{1/p} \quad (1)$$

In the case of p=2, we obtain Euclidean distance; in the case of p=1, we obtain Manhattan distance. This work adopts the Manhattan distance for the proximity measure. When there are n attributes for a data set, the Manhattan distance between a given instance x_i and centroid of j-th cluster x_c^j can be computed as below.

$$\sum_{i=1}^{n} \left| x_i - x_c^j \right|, \; j = 1, k \quad (2)$$

The processes of k-means algorithm are listed as follows:

1) To give a user-specified value of k.
2) To choose k initial centroids.
3) To calculate the similarity distance, and assign each instance to its belonging cluster based on the proximity measure.
4) To update the new centroid of each cluster.
5) To check step (4) to see if the centroids alter. If so, back to step (3). When the centroids of clusters remain at the same position, we achieve the convergence state of k-means algorithm.

The k-means clustering algorithm has been widely applied to a variety of applications. However, the classification accuracy of using the original k-means algorithm is always deteriorated when clusters are of differing sizes, densities, and non-globular shapes, and also when the data set contains outliers. Therefore, this study adopts a more flexible measure by incorporating weights ($\vec{w}$) to the similarity measure to adaptively adjust the contribution of attributes to clustering. Therefore, the similarity distance is modified as follows:

$$\sum_{i=1}^{n} w_i^j \times \left| x_i - x_c^j \right|, \; j = 1, k \quad (3)$$

The weights are devised with n dimensions for each cluster. Also, as shown in (3), the value of weights for each cluster can be different, i.e. non-symmetrical form, resulting in a non-symmetrical weighting for the similarity distance computing. The modified k-means algorithm was named non-symmetrical weighted k-means (NSWKM) in this work.

B. Basic Artificial Bee Colony Algorithm

The basic ABC algorithm proposed by Karaboga *et al*. [7], [8] is inspired from the foraging behavior of honey bees and designed to be an effective tool for solving optimization problems. Initially, a total of N food sources are generated according to a random search equation in the search space, and then the value of objective function (f_i) at food source i is evaluated using the current variables to measure the quality of nectar source. Hence, the fitness value for a certain food source i, $fitness_i$, can be obtained based on its value of objective function (f_i) as follows:

$$fitness_i = \begin{cases} 1/(1+f_i), \text{ if } f_i \geq 0 \\ 1+|f_i|, \text{ if } f_i < 0 \end{cases} \quad (4)$$

The basic ABC algorithm mainly composes three phases: employed bee phase, onlooker bee phase, and scout bee phase. The details of the three phases are listed as follows:

1) Employed bee phase

In this phase, N employed bees are sent to the food sources ($x_{i,j}$) first, and then the employed bees explore candidate positions in the neighborhood after they occupied food sources $x_{i,j}$. A candidate position ($v_{i,j}$) is obtained using the information of neighbor food source (r) and a random number $\phi_{i,j}$ as follows:

$$v_{i,j} = x_{i,j} + \phi_{i,j}(x_{i,j} - x_{r,j}) \quad (5)$$

where $\phi_{i,j} \in [-1.0, 1.0]$ and $j \in \{1,2,\ldots,n\}$. The index j is selected randomly from the interval of [1, n]. If the quality of nectar at the food source r is better than that at i, the bee choices r as the new food source. Namely, the greedy selection is applied in the employed bee phase.

2) Onlooker bee phase

Afterward the employed bee memorized the final position of the new food source and flew back to the hive to share her information with onlookers. As onlooker bees got the information shared from employed bees, each onlooker bee selects the one of food sources explored by employed bees based on a probability selection process. Namely, basic ABC algorithm applies the roulette wheel selection mechanism to select a food source in the onlooker bee phase. The probability (p_i) of selecting an explored food source by an onlooker bee is based on the following proportional probability formulation.

$$p_i = fitness_i / \sum_{j=1}^{N} fitness_j \quad (6)$$

After an onlooker bee selected a good food source as her current solution, she also performed a local search to search a possible new solution, and then chose the better food source with higher nectar as the new solution, the same process of greedy selection with an employed bee.

3) Scout bee phase

If the quality of a certain food source cannot be improved over "*Limit*" trials done by employed or onlooker bees, the bee occupying at the food source becomes a scout bee. The scout bee will abandon the current food source and explore a new one randomly.

C. Rank-Based Artificial Bee Colony Algorithm

In the presented rank-based ABC, there are two

modifications for the employed bee and scout bee phases. First, the probability (p_i) is modified based on the rank of fitness of a food source. The rank of fitness equaling to 1 means that the value of fitness at the food source is maximum, whereas the rank equaling to N means the value of fitness at the food source is minimum. Here the value of N is identical to the number of food source. The nonlinear rank-based selective pressure model is formulated as

$$p_i^{rank} = \frac{1}{1-(1-q)^N} \times q \times (1-q)^{rank-1} \quad (7)$$

with satisfying the constraint: $\sum_{i=1}^{N} p_i = 1$. The value of parameter q is a positive real with the range of $0 < q < 1$, and a larger value of q implies a stronger selective pressure for the value of probability. In the presented rank-based ABC algorithm, the value of q is devised as a dynamic function related to the number of iteration (*iter*) and maximum number of iteration ($iter_{max}$) as below.

$$q = q_{\min} + (q_{\max} - q_{\min}) \times (iter - 1) / (iter_{\max} - 1) \quad (8)$$

where q_{max} and q_{min} represent the minimum and maximum values of selective pressure, respectively. In this work, the parameters of q_{min} and q_{max} were 0.5 and 0.9, respectively.

Second, the abandoned mechanism of basic ABC is modified by abandoning bad food sources every iteration in the scout bee phase to enhance the exploration ability of the algorithm, whereas an abandoned criterion of "*Limit*" is used for published ABC algorithms. The role of scout bee is to perform a global search in the search space. In this work, the number of scouts was set to 5% of the size of employed bees according to the mean number of scouts which is about5%-10% of the colony from the observation of real bees [8]. Therefore, this work incorporated the two modifications into the basic ABC algorithm to achieve a balance for the exploitation and exploration. The proposed enhanced version of ABC was names RABC algorithm in this work.

D. Proposed NSWKM-RABC Procedure

This study uses RABC to find out the best weight (w_c^j), then provides the best weight as the required input of NSWKM algorithm for evaluating the classification accuracy. In this work, the objective function is set to minimize the classification error, the sum of the absolute value of difference between the predicted class C_{pred} and the actual class C_{actual}, presented as follows:

$$f(\vec{w}) = Min\left(\sum_{i=1}^{n} \left| \left(C_{pred}\right)_i - \left(C_{actual}\right)_i \right| \right) \quad (9)$$

The flowchart of the proposed NSWKM-RABC is plotted in Fig. 1. As shown in Fig. 1, the NSWKM clustering algorithm gives the results of predicted class C_{pred} to the employed bee, onlooker bee and scou

RABC according to the new position of updated food source $\vec{x}$. Then the best solution so far is memorized in the "memorized the best solutions so far" module after completing the three bee phases. Check the convergence criterion to see if the convergence state is arrived. If it is not the case, RABC outputs an updated weight vector $\vec{w}$ to NSWKM module to re-evaluate the predicted class C_{pred} required for the three bee phases of RABC. The iterative process will be repeated until the optimal classification is achieved.

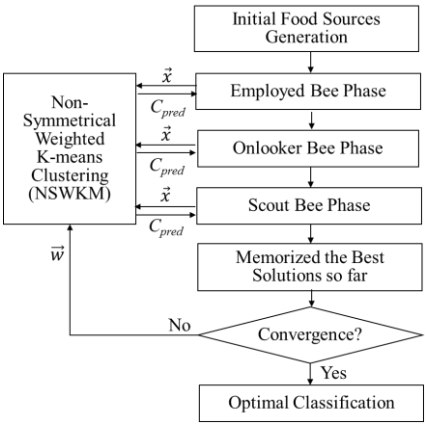

Fig. 1. Flow chart of the proposed NSWKM-RABC algorithm.

III. Ten-Fold Cross-Validation for a Data Set and Performance Parameters of Classification

This work uses 10-fold cross-validation (CV) method in partitioning training and test sets to give more reliable results. The process of 10-fold CV is to divide the data set into 10 segments: 9 training sets and 1 testing set. Accordingly, the initial data set are partitioned into 10 mutually exclusive folds, F_1, F_2, …, and F_{10}, each of approximately equal size. In iteration i, partition F_i is picked as the test set, and the remaining partitions are collectively used to train the model or pattern. Therefore, there are 10 iterations in performing training and testing processes for a run. This work also adopts stratified 10-fold CV, the same process as Weka software developed at the University of Waikato, New Zealand [11], instead of random one to achieve well "balanced" data sets.

In addition, in data mining the confusion matrix (Table I) is generally used to evaluate the computing accuracy of data bases to achieve the classified accuracy (CA), which is CA=(*tn*+*tp*)/(*tn*+*tp*+*fn*+*fp*). That is, the inaccuracy of classification is 1-CA. However, in clinical diagnosis, cost matrix still has to be counted in addition to the consideration of classification accuracy. This is because in diagnosing a medical case, the inaccuracy lies mainly in false positivity (*fp*) and false negativity (*fn*). The inaccuracy of false positivity/negativity means the ratio a patient is no/with disease but is diagnosed contrarily. Accordingly, necessary cost resulting from misdiagnosis is analyzed to form a cost matrix. This work assigns 1 unit of cost to be paid once a healthy person is misdiagnosed with a disease. On the contrary, five units of cost will have to be paid when a sick person is misdiagnosed without a disease. This shows the later has to pay 5 times as much as the former. Therefore, showing in Table I, the cost value, filled in the parenthesis, from diagnosing mistakenly is *fp*+5×*fn* . The higher the ratio

of *fn*, the higher cost will have to be paid, but the result will be far from being satisfactory.

TABLE I: CONFUSION AND COST MATRICES OF CLASSIFICATION

Actuality	Prediction	
	Negative	Positive
Negative	*tn* (0)	*fp* (1)
Positive	*fn* (5)	*tp* (0)

The classification result will be compared using confusion matrix and also performance parameters, which are presented as follows:

1) Accurac: $\frac{tn+tp}{tn+fn+fp+tp}$; Sensitivity (P): $\frac{tp}{fn+tp}$
2) Specificity (R): $\frac{tn}{tn+fp}$; F-measure: $\frac{2\times P\times R}{P+R}$
3) Cost: $fp+fn\times 5$

IV. COMPUTATIONAL RESULTS OF MEDICAL DIAGNOSIS

The proposed NSWKM-RABC algorithm was then applied to the medical diagnoses with five data sets of diseases to assess the performance of the algorithm. The data sets published on the web site of UCI repository [10] are: Breast Cancer Wisconsin, Heart-staglog, PIMA, ILPD, and Hepatitis. The data with missing attributes were removed, and the data in each attribute was normalized using z-score normalization, so that the mean and standard deviation of the data set are 0 and 1, respectively. This work also performed three classifiers listed in the top ten data mining algorithms for the medical diagnoses. The three classifiers are Naïve Bayes [12], [13], C4.5 [14], and k-means [15]. In addition, each case was performed 30 times of independent runs.

A. Breast Cancer Disease

This data set of breast cancer named "Breast Cancer Wisconsin" contains 699 instances described by 9 attributes, and the numbers of benign and malignant for breast cancer were 458 and 241, respectively. There were 683 instances being used for classification after we deleted 16 instances with missing data. Table II lists the means and standard deviations (SD) of the training and test sets after 30 independent runs. The means of the training and test sets were 98.00% 97.33%, respectively. Clearly, the mean of training set was higher than that of test set. In addition, the standard deviations for the training and test sets were 0.0297% and 0.2555%, respectively. The training set with a smaller value of SD than that of test set. Table III lists the confusion matrix of test set for assessing the performance of four algorithms to classification. The confusion matrix was obtained by combining the classification results of 10-fold test sets, the same process as Weka software [11] developed by Machine Learning Group at the University of Waikato. This works also performed Naïve Bayes, C4.5, and k-means classifiers included in the Weka software for the solution comparison. From Table III, the proposed NSWKM-RABC achieved the lowest false negativity with 3 only. The comparison of performance parameters was listed in Table IV. The proposed NSWKM-RABC provided the best solution in terms of accuracy, sensitivity, F-measure, and cost. The classification accuracy and cost achieved by the proposed NSWKM-RABC were 97.8% and 27, respectively. The computational cost for performing a 10-fold CV using NSWKM-RABC was 22 seconds on Windows 7 platform.

TABLE II: CLASSIFICATION SOLUTIONS OF BREAST CANCER

solution	training set	test set
mean	98.00%	97.33%
standard deviation	0.0297%	0.2555%

TABLE III: CONFUSION MATRIX FOR THE DATA SET OF BREAST CANCER

reality	algorithms	prediction	
		benign	malignant
benign	Naïve Bayes	425	19
	C4.5	428	16
	k-means	435	9
	NSWKM-RABC	432	12
malignant	Naïve Bayes	6	233
	C4.5	11	228
	k-means	18	221
	NSWKM-RABC	3	236

TABLE IV: PERFORMANCE PARAMETERS OF BREAST CANCER

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	96.34%	96.04%	96.05%	97.80%
sensitivity	97.49%	95.39%	92.47%	98.74%
specificity	95.72%	96.39%	97.97%	97.30%
F-measure	96.59%	95.88%	95.14%	98.02%
cost	49	71	99	27

TABLE V: CLASSIFICATION SOLUTIONS OF CARDIAC DISEASE

solution	training set	test set
mean	88.70%	82.73%
standard deviation	0.4720%	1.2264%

TABLE VI: CONFUSION MATRIX FOR THE DATA SET OF CARDIAC DISEASE

reality	algorithms	prediction	
		healthy	sick
healthy	Naïve Bayes	130	20
	C4.5	124	26
	k-means	122	28
	NSWKM-RABC	133	17
sick	Naïve Bayes	21	99
	C4.5	38	82
	k-means	27	93
	NSWKM-RABC	23	97

TABLE VII: PERFORMANCE PARAMETERS OF CARDIAC DISEASE

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	84.81%	76.29%	79.63%	85.19%
sensitivity	82.50%	68.33%	77.50%	80.83%
specificity	86.67%	82.67%	81.33%	88.67%
F-measure	84.53%	74.82%	79.37%	84.57%
cost	125	216	163	132

B. Cardiac Disease

The data set of cardiac disease named "Heart-staglog" has 13 variables and 1 class attribute. The total number of the data set is 270 with 120 healthy cases and 150 sick cases. Table V lists the means and standard deviations of the training and test sets after 30 independent runs. Clearly, the solutions of mean and SD for the training set were better than those of test set. The mean and SD of the training set were 88.70% and 0.4720% with higher classification accuracy and lower SD than the case of test set. Table VI lists the confusion matrix of test set and indicates that Naïve Bayes gave the lowest false negativity with 21 and presented algorithm gave

the lowest false positivity with 17. The comparison of the performance parameters was listed in Table VII. The proposed NSWKM-RABC provided the best solution in terms of accuracy, specificity, and F-measure, whereas Naïve Bayes gave the best solution in terms of sensitivity and cost. The computational cost for performing a 10-fold CV using NSWKM-RABC was 130 seconds on Windows 7 platform.

C. Gestational Diabetes

The data set of gestational diabetic disease named "Pima-Indians-Diabetes" (PIMA) contains 8 attributes and one class attribute. This data set has 768 instances including 500 patients with gestational diabetes and 268 patients without the disease. From Table VIII, the means of the training set and test set were 78.27% and 74.796%, respectively. Also, the SD of training set was smaller than that of test set. The confusion matrix of test set was listed in Table IX. Naïve Bayes gave the lowest false negativity with 104 and presented NSWKM-RABC algorithm gave the lowest false positivity with 72. Table X indicates the comparison of the performance parameters and displays Naïve Bayes gave the best solution in terms of accuracy, sensitivity, and F-measure. In this medical diagnosis, the proposed NSGKM-RABC and Naïve Bayes provided the same predicted accuracy of classification with 76.3%. The computational cost for performing a 10-fold CV using NSWKM-RABC was 43 seconds on Windows 7 platform.

TABLE VIII: CLASSIFICATION SOLUTIONS OF DIABETES

solution	training set	test set
mean	78.2700%	74.7960%
standard deviation	0.3031%	0.84010%

TABLE IX: CONFUSION MATRIX FOR THE DATA SET OF DIABETES

reality	algorithms	prediction	
		healthy	sick
healthy	Naïve Bayes	422	78
	C4.5	407	93
	k-means	380	120
	NSWKM-RABC	428	72
sick	Naïve Bayes	104	164
	C4.5	108	160
	k-means	135	133
	NSWKM-RABC	110	158

TABLE X: PERFORMANCE PARAMETERS OF DIABETES

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	76.30%	73.83%	66.79%	76.30%
sensitivity	61.19%	59.70%	49.62%	58.96%
specificity	84.40%	81.40%	76.00%	85.60%
F-measure	75.12%	68.88%	60.04%	69.82%
cost	598	633	795	622

D. Liver Disease

The data set of liver disease named "Indian Liver Patient Data Set" (ILPD) contains 10 variables, and it was collected from North East of Andhra Pradesh, India. The class attribute was divided as liver patient or not. This data set contains 414 patients with liver disease and 165 patients without liver disease when removing the records with missing data. From Table XI, the means of the training set and test set were 74.5343% and 70.7138%, respectively. Table XII lists the confusion matrix of test set and indicates that the proposed NSWKM-RABC gave the lowest false negativity with 28, and Naïve Bayes gave the lowest false positivity with 8. The comparison of performance parameters was listed in Table XIII. Clearly, the proposed NSWKM-RABC provided the best solution in terms of accuracy and cost. The computational cost for performing a 10-fold CV using NSWKM-RABC was 652 seconds on Windows 7 platform.

TABLE XI: CLASSIFICATION SOLUTIONS OF LIVER DISEASE

solution \ data set	training set	test set
mean	74.5354%	70.7138%
standard deviation	0.3432%	0.9685%

TABLE XII: CONFUSION MATRIX FOR THE DATA SET OF LIVER DISEASE

reality	algorithms	prediction	
		healthy	sick
healthy	Naïve Bayes	157	8
	C4.5	35	130
	k-means	49	116
	NSWKM-RABC	33	132
sick	Naïve Bayes	248	166
	C4.5	55	359
	k-means	91	323
	NSWKM-RABC	28	386

TABLE XIII: PERFORMANCE PARAMETERS OF LIVER DISEASE

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	55.78%	68.05%	64.25%	72.37%
sensitivity	95.15%	86.71%	91.79%	93.24%
specificity	40.09%	21.21%	29.70%	20.00%
F-measure	56.41%	34.08%	44.88%	32.94%
cost	1248	705	576	272

TABLE XIV: CLASSIFICATION SOLUTIONS OF HEPATITIS

solution	training set	test set
mean	98.9947%	87.5000%
standard deviation	0.9597%	2.7576%

TABLE XV: CONFUSION MATRIX FOR THE DATA SET OF HEPATITIS

reality	algorithms	prediction	
		live	die
live	Naïve Bayes	59	8
	C4.5	64	3
	k-means	40	27
	NSWKM-RABC	65	2
die	Naïve Bayes	5	8
	C4.5	8	5
	k-means	2	11
	NSWKM-RABC	4	9

TABLE XVI: PERFORMANCE PARAMETERS OF HEPATITIS

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	83.75%	86.25%	63.75%	92.50%
sensitivity	61.54%	38.46%	84.62%	69.23%
specificity	90.77%	95.52%	59.70%	97.01%
F-measure	73.35%	54.84%	70.00%	80.80%
cost	33	43	37	22

E. Hepatitis

This data set of liver disease named "Hepatitis" comprises 19 attributes and 1 class attribute with 155 patient instances. This case is a small size of data set and has many missing data. After we deleted the instances with missing data, there are only 80 patient instances which 67 patients are lived and 13 patients are died. As shown in Table XIV, the means of the training set and test set were 98.994% and 87.50%,

respectively. The solution accuracy for the training set was very high. Table XV lists the confusion matrix of test set and indicates that k-means gave the lowest false negativity with 2 and presented algorithm gave the lowest false positivity with 2. Furthermore, the comparison of coefficients of performance parameters was listed in Table XVI. The proposed NSWKM-RABC provided the best solution in terms of accuracy, specificity, F-measure, and cost. The computational cost for performing a 10-fold CV using NSWKM-RABC was 47 seconds on Windows 7 platform.

TABLE XVII: RANK VALUES OF PERFORMANCE PARAMETERS

(a) breast cancer

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	2	4	3	**1**
sensitivity	2	3	4	**1**
specificity	4	3	**1**	2
F-measure	2	3	4	**1**
cost	2	3	4	**1**
sum of rank	12	16	16	**6**

(b) cardiac disease

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	2	4	3	**1**
sensitivity	**1**	4	3	2
specificity	2	3	4	**1**
F-measure	2	4	3	**1**
cost	1	4	3	**2**
sum of rank	8	19	16	**7**

(c) gestational diabetes

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	**1**	3	4	**1**
sensitivity	**1**	2	4	3
specificity	2	3	4	**1**
F-measure	**1**	3	4	**2**
cost	**1**	3	4	**2**
sum of rank	6	14	20	**9**

(d) liver disease

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	4	2	3	**1**
sensitivity	**1**	4	3	**2**
specificity	**1**	3	2	4
F-measure	**1**	3	2	4
cost	4	3	2	**1**
sum of rank	11	15	12	**12**

(e) hepatitis

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	3	2	4	**1**
sensitivity	3	4	**1**	2
specificity	3	2	4	**1**
F-measure	2	4	3	**1**
cost	2	4	3	**1**
sum of rank	13	14	15	**6**

TABLE XVIII: COMPARISON OF ALGORITHMIC PERFORMANCE

parameter	Naïve Bayes	C4.5	k-means	NSWKM-RABC
accuracy	12	15	17	5
sensitivity	8	17	15	10
specificity	12	14	15	9
F-measure	8	17	16	9
cost	10	17	16	7
sum of rank	50	80	79	40
avg. of rank	2.00	3.20	3.16	1.60

V. DISCUSSION OF THE CLASSIFICATION RESULTS

From the above five medical diagnoses, the proposed NSWKM-RABC exhibited a good performance for the classifications of five data sets of diseases. We ranked the solutions of the performance parameters for obtaining more information relating to the four algorithms for the five disease classifications. Table XVII lists the rank values of the solutions based on the solutions of performance parameters from Table IV-Table XVI. The lower rank value represents the better solution we obtained. For example, when the rank value equals 1, the achieved solution obtained using a classifier is the best. Furthermore, we combined the values of rank recorded in Table XVII, the comparison of the five diseases classifications using the four classifiers was listed in Table XVIII. The averages of rank for Naïve Bayes, C4.5, k-means, and the proposed NSWKM-RABC were 2.00, 3.20, 3.16, and 1.60, respectively. Clearly, the proposed NSWKM-RABC outperforms the other three algorithms.

VI. CONCLUSION

This study proposed a classifier, termed NSWKM-RABC algorithm, by combining the non-symmetrical weighted k-means clustering algorithm and the rank-based artificial bee colony to classify the data sets of medical diseases. The similarity distance used in k-means was weighted by the optimal non-symmetrical weights which were obtained via RABC computation. Five data sets of diseases including breast cancer, cardiac disease, diabetes, liver disease, and hepatitis were adopted. Also three classifiers, including Naïve Bayes, C4.5 and k-means, were performed for the comparison of the classification solutions. From the results of medical diagnoses, the proposed NSWKM-RABC provided the most reliable solutions for the classifications. The average rank of the solution accuracy obtained using the proposed NSWKM-RABC was minimum. We also observed that the solutions obtained using Naïve Bayes were better than those of using C4.5 and k-means. Accordingly, the proposed algorithm is an effective algorithm for diagnosing the data sets of medical disease.

REFERENCES

[1] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, December 2008.

[2] S. Ahan, H. Kodaz, S. Günеş, and K. Polat, "A new classifier based on attribute weighted artificial immune system," *Lecture Notes in Computer Science*, vol. 3280, pp. 11-20, Oct. 2004.

[3] S. Özşen and S. Günеş, "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems," *Expert Systems with Applications*, vol. 36, issue 1, pp. 386-392, Jan. 2009.

[4] P. Gancarskia, A. Blanschea, and A.Waniab, "Comparison between two coevolutionary feature weighting algorithms in clustering," *Pattern Recognition*, vol. 41, issue 3, pp. 983-994, March 2008.

[5] M. Nazari, J. Shanbehzadeh, and A. Sarrafzadeh, "Fuzzy c-means based on automated variable feature weighting," in *Proc. the International MultiConference of Engineers and Computer Scientists 2013*, Hong Kong , vol. 1, March 13-15, 2013.

[6] A. Blansché, P. Gançarski, and J. J. Korczak, "MACLAW: A modular approach for clustering with local attribute weighting," *Pattern Recognition Letters*, vol. 27, issue 11, pp.1299-1306, Aug. 2006.

[7] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, vol. 39, issue 3, pp. 459-471, Nov. 2007.

[8] D. Karaboga and B. Basturk, "On the Performance of Artificial Bee Colony (ABC) Algorithm," *Applied Soft Computation*, vol. 8, no. 1, pp. 687-697, Jan. 2008.
[9] G. Zhu and S. Kwong, "Gbest-guided artificial bee colony algorithm for numerical function optimization," *Applied Mathematics and Computation*, vol. 217, issue 7, pp. 3166-3173, Dec. 2010.
[10] C. L. Blake and C. J. Merz, *UCI Repository of Machine Learning Databases*, Irvine, CA: University of California, Department of Information and Computer Science, 1998.
[11] R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "WEKA-experiences with a java open-source project," *Journal of Machine Learning Research*, vol. 11, pp. 2533-2541, Sept. 2010.
[12] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, issue 2-3, pp. 103-130, 1997.
[13] E. Fix and J. L. Jr. Hodges "Discriminatory analysis, nonparametric discrimination: consistency properties," *International Statistical Review*, vol. 57, no. 3, pp. 238-247, Dec. 1989.
[14] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Approach*, vol. 4, pp. 77-90, 1996.
[15] D. L. Olson and Y. Shi, *Introduction to Business Data Mining*, Boston, MA: McGraw-Hill /Irwin Education, 2006.

Jenn-Long Liu received his MS and PhD degrees in aeronautics and astronautics from National Cheng Kung University, Taiwan in 1987 and 1991, respectively. He is a professor at the Department of Information Management of I-Shou University, Kaohsiung, Taiwan. He also had been a visiting scholar at University of California, Riverside (UCR) and Illinois State University (ISU) in 2010 and 2014, respectively. His current research interests include artificial intelligence, evolutionary computation, data mining, and wireless sensor network. Dr. Liu is a life member of the Taiwanese Association for Artificial Intelligence (TAAI).

Chung-Chih Li was born and in Taiwan and received his bachelor of engineering degree from Tamkang University, Taipei, Taiwan in 1986, master of science degree from Lamar University, Beaumont, Texas in 1991, and Ph.D. degree from Syracuse University, Syracuse, New York in 2001. All degrees Dr. Li received were in computer science. His doctoral dissertation was in the area of theoretical computer.

He worked as a programmer before moving into academia. Currently, he is a professor at the School of Information Technology, Illinois State University. Before that, he had been a visiting professor at Colgate University, assistant professor at Lamar University Texas from 2002 to 2006, assistant professor from 2006 to 2009) and associate professor from 2009 to 2013 at Illinois State University. His primary research is in the area of theoretical computer science, higher typed computability and complexity in particular. He also published several papers in machine learning, algorithms, and cryptography, as he has extended his research interests in these areas.

Dr. Li has been a number of serval professional societies including ACM SIGCSE, IEEE CCC (Conference on Computational Complexity), IEEE CIMCA (Intl. Conf. on Computational Intelligence for Modeling Control & Automation), CIE (Computability in Europe), TOC (International Conference on Theoretical Computer Science). Also, Dr. Li is an active member of NATPA (North America Taiwanese Professors' Association).