J Med Syst (2011) 35:283–289
DOI 10.1007/s10916-009-9364-8

ORIGINAL PAPER

An Application of Artificial Immune Recognition System for Prediction of Diabetes Following Gestational Diabetes

Hung-Chun Lin · Chao-Ton Su · Pa-Chun Wang

Received: 15 June 2009 / Accepted: 10 August 2009 / Published online: 25 August 2009

Abstract Diabetes mellitus (DM) is a disease prevalent in population and is not easily perceived in its initial stage but may sway a patient very seriously in later stage. In accordance with the estimation of World Health Organization (WHO), there will be 370 million diabetics which are 5.4% of the global people in 2030, so it becomes more and more important to predict whether a pregnant woman has or is likely to acquire diabetes. This study is conducted with the use of the machine learning—Artificial Immune Recognition System (AIRS)—to assist doctors in predicting pregnant women who have premonition of type 2 diabetes. AIRS is proposed by Andrew Watkins in 2001 and it makes use of the metaphor of the vertebrate immune system to recognize antigens, select clone, and memorize cells. Additionally, AIRS includes a mechanism, limited resource, to restrain the number of memory cells from increasing uncontrollably. It has also showed positive results on problems in which it was applied. The objective of this study is to investigate the feasibility in using AIRS to predict gestational diabetes mellitus (GDM) subsequent DM. The dataset of diabetes has imbalanced data, but the overall classification recall could still reach 62.8%, which is better than the traditional method, logistic regression, and the technique which is thought as one of the powerful classification approaches, support vector machines (SVM).

Keywords Artificial Immune Recognition System (AIRS) · Type 2 diabetes · Vertebrate immune system · Imbalanced data

H.-C. Lin · C.-T. Su
Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan

P.-C. Wang
Department of Otolaryngology, Cathay General Hospital, Fu Jen Catholic University School of Medicine, Taipei, Taiwan

C.-T. Su (✉)
Room 820, Engineering Building I, 101, Sec. 2, Kuang Fu Road, Hsinchu, Taiwan 30013, Republic of China
e-mail: ctsu@mx.nthu.edu.tw

Introduction

Diabetes mellitus (DM) is a disease prevalent in population. The vital statistics data provided by the American Diabetes Association (ADA) reveal that, 23.6 million children and adults, or 8.0% of the population, suffer from diabetes in United States. Based on the World Health Organization (WHO) estimation, diabetes patients will reach a number of 370 million (5.4% of the global population) by year 2030. Resulted from insulin deficiency or insulin resistance, diabetes mellitus is a disease complex that characterized by hyperglycemia clinically. The ADA classified diabetes into 4 types: type 1 diabetes (insulin-dependent), type 2 diabetes (non-insulin dependent), gestational diabetes mellitus (GDM), and other specific types.

GDM is a hyperglycemic status during pregnancy. Some women would develop insulin resistance to cause drastic glucose metabolic change in later trimesters of pregnancy. GDM is considered as pre-diabetic when a pregnant woman can not recover from her temporary hyperglycemic status during pregnancy. Unfortunately, causes of DM following GDM remain obscure [1–4].

Some models have been developed to predict GDM subsequent DM. For example, Cheung and Helmink [5] concluded that need for bedtime insulin during pregnancy is predictive of subsequent DM after labor. Lee et al. [6]

revealed that use of insulin, Asian women, and elevated 1-hour glucose test can predict later development of type 2 DM. Nam et al. [7] used the homeostasis model assessment (HOMA) and quantitative insulin-sensitivity check index (QUICKI) models to investigate important indices for subsequent DM. Carr et al. [8] performed a retrospective study and aimed to predict DM following GDM. An observational cohort study of women with GDM was conducted and screened for postpartum DM., and efforts to improve puerperal DM screening in the high-risk group are warranted [9].

Artificial Immune Recognition System (AIRS) is a resource limited supervised learning algorithm [10]. The algorithm uses resources competition, clone selection, maturation, mutation and memory cells generation for the development of prediction modeling. There have been many researches applying AIRS for predicting on many diseases, such as obstructive sleep apnea (OSA) syndrome, lung cancer and so on [11–14], and there are good performances with applying AIRS. However, there is no research utilizing AIRS for predicting DM so far. Thus, apart from traditional statistical technique—in which pre-assumptions are often insufficient, the aim of this study is to evaluate the feasibility in using AIRS to predict DM development following GDM.

Materials and methods

Diabetes classification problem

In this study, the dataset we used is obtained from the medical cases of pregnant women in a particular medical center in Taipei, Taiwan. This center usually carries out glucose change test (GCT) during the 24th to 28th week of gestation. The gravid women eat 50-g glucose and proceed with blood sugar test. If the value of 1-h 50-g glucose is higher than 140 mg/dl, they need a further 3-h 100g OGTT. In accordance with the criterion set by the National Diabetes Data Group, if more than one of the following evaluation standards are met, the gravid women are considered that they have GDM: (1) fasting value ≥105 mg/dl; (2) 1-h ≥190 mg/dl; (3) 2-h ≥165 mg/dl; (4) 3-h ≥145 mg/dl. These cases were collected from 1998 to 2002 and inscribed in GDM records. During this period, some attributes of the women who had GDM were gathered and are shown in Table 1.

These attributes were employed to establish a predictor of type 2 diabetes mellitus after delivery. For exploring the type 2 diabetes mellitus state, all gravid women with GDM were informed to come back for metabolic test; 558 women with GDM are recorded in the database. Among the 558 registrations, 152 women were contacted and they agreed to undergo the follow-up test for type 2 diabetes mellitus. The test results could be assorted into three groups based on the benchmark set by the Department of Health, Executive Yuan, Taiwan: (1) diabetes mellitus (DM) (fasting value ≥126 mg/dl or 2-h 75-g OGTT value ≥200 mg/dl); (2) pre-diabetes mellitus (Pre-DM) (110 mg/dl ≤ fasting value <126 mg/dl or 140 mg/dl ≤2-h 75-g OGTT value <200 mg/dl); (3) Normal (fasting value <110 mg/dl and 2-h 75-g OGTT value <140 mg/dl). From the 152 samples, ten instances were DM, 32 were Pre-DM, and 110 were Normal. This is a so-called imbalanced data wherein the majority part of the dataset comes from one or two classes and the rest from the other classes.

Table 1 Attributes gathered for GDM case

Symbol	Attribute
A_1	Motherly age
A_2	50-g oral glucose challenge test value
A_3	Fasting glucose value
A_4	1-h 100-g OGTT value
A_5	2-h 100-g OGTT value
A_6	3-h 100-g OGTT value
A_7	BMI before gestation
A_8	Weight before gestation
A_9	Increased weight during pregnancy
A_{10}	Weight of newborn
A_{11}	Family history of diabetes mellitus
A_{12}	The number of pregnancy times

Natural and artificial immune systems

The natural immune system is a layered and distributed pattern detection system against extraneous ingredients named as Antigen like viruses, microbes, and so on. It consists of several functional elements throughout the body. The immune system is handling the defense mechanism of the body by innate and adaptive immune response. Adaptive immune response is especially important for us because it has abilities like memory acquisition, diversity, recognition, etc. The adaptive immunity then become as the main line of defense in the body and has three key properties. It responds only if an invader is present and it remembers a previous contact with an invader, therefore responding faster after initial recognition. Furthermore, it can differentiate between the self and the non-self.

The immune system is composed of lymphocytes which are classified into two types such as B and T cells. T-cells mature in the thymus, while B-cells mature in the bone marrow. Both of these lymphocytes have receptors to bind invading antigens that could trigger immune response. The

receptor of T cells is title as T-cell receptor (TCR) and that of the B cells is called Antibody, which is the most ordinary model unit of immune system in Artificial Immune System (AIS) [15, 16].

The antigens and the antibodies have one-to-one correspondence. When the foreign antigens invade the body, such as above mentioned, B-cells will produce corresponding antibodies binding to the antigens. The AIS imitates this mechanism. When data enter the system like an antigen, the model will generate the corresponding datum as the antibody and store this in the memory cells to immediately respond when the same situation happens. Furthermore, the AIS can produce similar antibodies through mutation to respond to the similar data faster.

Artificial Immune Recognition System (AIRS) classification algorithm

AIRS is proposed by Andrew B. Watkins and it is a resource limited and a type of supervised learning algorithm [10]. This algorithm used immune mechanisms are resources competition, clone selection, maturation, mutation and memory cells generation. The training and test data items are viewed as antigens in the system. These antigens induce the B-cells in the system to produce artificial recognition balls (ARBs). These ARBs compete with each other for the given resource number. The ARBs with higher resources will get more chances to produce the mutated offspring to improve the system. The memory cells generated after all training antigens have been introduced are used to classify the test data items. The algorithm is composed of five stages. These stages are initialization, memory cell identification and ARB generation, competition for resources and development of a candidate memory cell, memory cell introduction, and classification. The Fig. 1 shows the flowchart of AIRS algorithm.

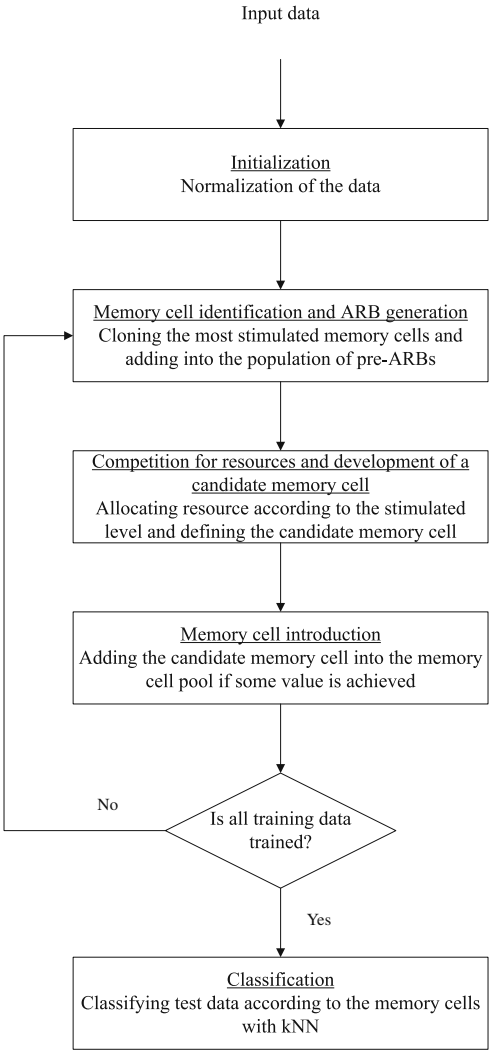

Fig. 1 Flowchart of AIRS algorithm

Initialization

The first stage of the algorithm is data pre-processing stage. In this stage, all items in the dataset are normalized to comply with the Euclidean distance which says that the distance between any two data is in the range of [0, 1]. This could be executed by a variety of means. In this study, we use Eq. 1 to deal with normalization, and the calculation of the Euclidean distance is showed in Eq. 2.

$$x_i = \frac{\left(\frac{a_i}{max_i - min_i}\right)}{\sqrt{m}} \tag{1}$$

where a_i is the original datum of ith attribute, max_i and min_i are the maximum and minimum of ith attribute of all data, respectively, and m is the number of attributes in data.

$$Euclidean\_distance = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{2}$$

where x and y represent feature vectors.

After normalization, the affinity threshold is calculated as Eq. 3

$$affinity\_threshold = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} affinity(ag_i, ag_j)}{\frac{n(n-1)}{2}} \tag{3}$$

where n is the number of training data items, ag_i and ag_j are the ith and jth training antigen in the training vector, and

Table 2 Level of the parameters in experiment

Parameters	TotalNumResources	Stimulation threshold	Mutation rate	Clone rate	Hyper mutation rate	Affinity threshold scalar
Level 1	200	0.9	0.1	10	2	0.2
Level 2	100	0.5	0.05	5	1	0.1

$affinity(ag_i, ag_j)$ represents the Euclidean distance between the two antigens' feature vector.

Memory cell identification and ARB generation

In this stage, the first step is to find out the memory cell, mc_{match}, given a specific training antigen. The mc_{match} is defined as $\arg\max_{mc \in MC} stimulation(ag, mc)$, where stimulation($x$, y) is defined as Eq. 4.

$$stimulation(x,y) = 1 - Euclidean\_distance(x,y) \tag{4}$$

Given this definition, it can be assumed that the antibody, mc_{match}, is the most stimulated memory cell by the given antigen in the set of memory cells of the same category. When mc_{match} is identified, this cell is used to create new ARBs to be introduced to the population of pre-existing ARBs. The number of new ARBs depends on the stimulation value between the memory cell and the antigen.

Competition for resources and development of a candidate memory cell

In this stage, all ARBs presently existing in the system are awarded the resource numbers according to their affinity values. The ARBs with higher affinity values will get more resources than those with lower affinity values. If the sum of the number of resources of all ARBs exceeds the allowed number, the system will remove the ARBs and their awarded resources beginning with the lowest number of resources until the sum of the number of resources of all ARBs is lower than the allowed number in the system. After this process, the stimulation values of all remaining ARBs are calculated and maximal and minimal stimulation values are determined as *maxStim* and *minStim*, respectively. The stimulation level of each ARB is recalculated as Eq. 5.

$$arb.stim = \begin{cases} \frac{arb.stim - minStim}{maxStim - minStim}, & \text{if class of } arb = \text{class of } antigen \\ 1 - \frac{arb.stim - minStim}{maxStim - minStim}, & \text{otherwise} \end{cases} \tag{5}$$

Then, we use Eq. 6 to calculate the average value of these levels for each class and verify if any of these average values is lower than a given stimulation threshold or not. If any of the average values is lower, the ARBs belonging to that class are mutated and the generated clones are added to ARB pool. This process is continuously done until the average stimulation levels of all classes are larger than the stimulation threshold.

$$s_i = \frac{\sum_{j=1}^{|ARB_i|} arb_j.stim}{|ARB_i|}, arb_j \in ARB_i \tag{6}$$

where i=1,2,...,nc , $s=\{s_1,s_2,...,s_{nc}\}$, $|ARB_i|$ is the number of ARBs belonging to ith class and $ab_j.stim$ is the stimulation level of jth ARB of ith class.

Memory cell introduction

After achieving the criterion described above, the ARB with the highest stimulation value in the same class with the presented training antigen is taken as a candidate memory cell, $mc_{candidate}$. If the stimulation value of the $mc_{candidate}$ motivated by the training antigen is higher than the stimulation value of the mc_{match}, the candidate memory cell is added to the set of memory cells. If this test is passed, a calculation of the affinity between $mc_{candidate}$ and mc_{match} must be obtained. If the affinity between this two memory cells is lower than the product of the affinity

Table 3 ANOVA table for the overall classification recall

Source	SS	*df*	MS	*F*	*p*
TotalNumResources	0.002	1	0.002	0.750	0.387
Stimulation threshold	0.001	1	0.001	0.232	0.631
Mutation rate	0.000	1	0.000	0.123	0.726
Clone rate	0.001	1	0.001	0.224	0.636
Hyper mutation rate	0.005	1	0.005	1.784	0.182
ATS	0.000	1	0.000	0.000	0.991
Error	1.703	633	0.003		
Total	74.342	640			

Table 4 Used parameters in AIRS for Diabetes dataset

Parameter	Value
Number of resources in system	200
Stimulation threshold	0.9
Mutation rate	0.1
Clone rate	10
Hyper mutation rate	2
Affinity threshold scalar	0.2
k value for k-nearest neighbor	1

threshold and the affinity threshold scalar, $mc_{candidate}$ is replaced with mc_{match} in the set of memory cells.

Classification

After repeating step 2 to step 4 to each training antigen, the developed memory cells are ready for exploitation and for classification. The classification is executed in a k-nearest neighbor approach. That is, the classification of a datum in the system is determined by the ballot of the results of the k most stimulated memory cells.

Used parameters

One advantage of AIRS is that it is not necessary to try all combinations of all parameters to find the best one. AIRS is self-adjusting to the feature of its architecture. Based on the experience of Goodman, Boggess and Watkins [17], the setting of AIRS's parameters had a classifier with only a few percentages of accuracy less than the optimal combination of the parameters of the system. However, for ensuring that the parameters would not affect the result of the diabetes dataset, a design of experiment was established. The level of each parameter is shown in Table 2. The diabetes dataset has been divided into five portions randomly. One part has been treated as test dataset, with other parts as the training dataset. For testing the validity of this experiment, we have used ten replications in each combination of levels, thus there have been a total of 640 runs in this experiment. Table 3, ANOVA table for the overall classification recall, shows that the parameters would not influence the result significantly. Therefore, we arbitrarily set the parameters as in Table 4.

Measure for performance evaluation

With an imbalanced data as dataset, if we use the ratio which is the total predicting correctly amounts divided by the total number of test dataset, it would be affected by the amount of the majority class and result in bias. There have been many researches to analyze the imbalanced data sets [18–21]. Most of them use the sensitivity, specificity, or g-means suggested by Kubat and Matwin [22] to evaluate the classifiers. Nevertheless, there are only two classes, the majority class and the minority class, investigated in their researches. However, the equations of the sensitivity and specificity are the number of the true positives divided by the total number of the actual positives and the number of the true negatives divided by the total number of the actual negatives, respectively. They are the same with the formula of the recall. Therefore, we calculate the recall rate of the individual class and compute the mean of the recall rate of the individual class. The individual classification recall used for the dataset is defined as Eq. 7

$$recall(T_j) = \frac{\sum_{i=1}^{|T_j|} assess(t_i)}{|T_j|}, \qquad t_i \in T_j$$
$$assess(t) = \begin{cases} 1, & \text{iff } classify(t) \equiv t.c \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where T_j is the set of data items which belong to class j to be classified (the test set) and $|T_j|$ is the number of T_j, $t \in T_j$, $t.c$ is the class of item t, and $classify(t)$ returns the classification of t by the developed classifier and the classification recall of each class of the testing data is calculated and the overall recall of classification is the average of the percentage of the three classification recall. Furthermore, for test results to be more credible and to avert the variation in virtue of sampling, five-fold cross-validation is used in this application.

Results and discussion

To determine the validity of AIRS, it was compared with some well-known classification methods, logistic regression (LR), decision tree (C 4.5), support vector machines (SVM), in this study. In this comparison, ARIS was built

Table 5 Classification accuracy of AIRS for diabetes classification problem with classification accuracies obtained by other methods

Testing criterion Algorithms	DM	Pre-DM	Normal	Overall
Logistic regression	0.100 (0.224)	0.033 (0.075)	0.882 (0.119)	0.338 (0.063)
Decision tree with C4.5	0.500 (0.354)	0.252 (0.170)	0.736 (0.081)	0.496 (0.132)
Support vector Machine	0.000 (0.000)	0.033 (0.075)	0.964 (0.059)	0.332 (0.009)
AIRS	0.600 (0.224)	0.538 (0.283)	0.745 (0.150)	0.628 (0.040)

with MATLAB (http://www.mathworks.com). LR was utilized by Weka software (http://www.cs.waikato.ac.nz/~ml/weka/). C 4.5 is design for decision tree analysis, and in this study, a decision tree was created by using See 5 software (http://www.rulequest/com/index.html). For the fairness of the comparison, we didn't prune any tree and use full model to execute this classification task in this study. LIBSVM was employed to deal with classification problems through the SVM [23]. The one-against-one strategy was assumed in LIBSVM to extend SVM to multi-class applications and the reasons were definitely listed in [24]. Furthermore, LIBSVM has a powerful parameter selection mechanism with cross-validation through collateral network search under the kernel of the radial basis function type. Using AIRS, the classification recall of predicting DM, Pre-DM, and Normal is 0.6, 0.538, and 0.745, respectively, and the results AIRS compared with other classifiers are showed in Table 5. The value with no bracket is the average recall rate and the value with parentheses is the standard deviation of the recall rate of the five-fold testing data. Although the classification recall of normal of AIRS is not the highest, this is not the focus of the diabetes dataset. The doctors want to find out the potential diabetics through this dataset. The information that can be gathered from the results can help the doctors to easily predict through AIRS whether the pregnant woman is likely to have diabetes, because AIRS could achieve the highest recall among the four algorithms in the table. In particular, SVM has been accepted as one of the most forceful classification techniques for imbalanced data, and it is heartening that AIRS could outperform SVM for this dataset.

Conclusion

In an era of information explosion, how to make clinical decision quickly and correctly is of paramount importance for practicing physicians. A lot of diseases should be diagnosed and treated early and diabetes is no exception. Particularly, due to the change of living-style, there are more and more people suffering diabetes mellitus which is a disease with high costs and risk. In this study, we make use of AIRS to determine that what type of pregnant women is likely to acquire diabetes mellitus and for the doctors to accurately predict the conditions of the pregnant women in a systematic way. According to the result above, AIRS is more workable than other traditional methods. Besides, SVM is a parameter sensitive approach. Even if LIBSVM, which has the parameter selection mechanism, was utilized, AIRS, which is self-adjusting to the feature of its architecture, still have better performance in this study. AIRS could reach the highest recall on DM and Pre-DM. Thus, if an expectant woman is determined that she might suffer diabetes mellitus or pre-diabetes mellitus by this constructed model, the obstetrician–gynecologist should pay more attention to this woman during the period of antenatal examination. Further, the obstetrician should offer more appropriate hygienic instruction and needed surveillance and treatment to the pregnant woman. Finally, it is hoped that there are more interesting results in the future researches.

Acknowledgement This work was supported in part by the National Science Council, Taiwan, under grant NSC-98-2221-E-007-071-MY3.

References

1. Madhavan, A., Beena, K. R., and Sanal, M. G., A pilot study on the usefulness of body mass index and waist hip ratio as a predictive tool for gestational diabetes in Asian Indians. *Gynecol. Endocrinol.* 24:701–707, 2008.
2. Seshiah, V., Balaji, V., Balaji, M. S., Panneerselvam, A., Thamizharasi, M., and Arthi, T., Glycemic level at the first visit and prediction of GDM. *J. Assoc. Phys. India*. 55:630–632, 2007.
3. Akinci, B., Celtik, A., Yener, S., and Yesil, S., Is fasting glucose level during oral glucose tolerance test an indicator of the insulin need in gestational diabetes? *Diabetes Res. Clin. Pract*. 82 (2):219–225, 2008.
4. Campos, M. A. A., Reichelt, A. A. J., Façanha, C., Forti, A. C., and Schmidt, M. I., Evaluation of a 1-h 75-g oral glucose tolerance test in the diagnosis of gestational diabetes. *Braz. J. Med. Biol. Res*. 41(8):684–688, 2008.
5. Cheung, N. W., and Helmink, D., Gestational diabetes: the significance of persistent fasting hyperglycemia for the subsequent development of diabetes mellitus. *J. Diabetes Its Complicat*. 20 (1):21–25, 2006.
6. Lee, A. J., Hiscock, R. J., Wein, P., Walker, S. P., and Permezel, M., Gestational diabetes mellitus: clinical predictors and long-term risk of developing type 2 diabetes: a retrospective cohort study using survival analysis. *Diabetes Care*. 30(4):878–883, 2007.
7. Cho, N. H., Lim, S., Jang, H. C., Park, H. K., and Metzger, B. E., Elevated homocysteine as a risk factor for the development of diabetes in women with a previous history of gestational diabetes mellitus: a 4-year prospective study. *Diabetes Care*. 28(11):2750–2755, 2005.
8. Carr, D. B., Newton, K. M., Utzschneider, K. M., Tong, J., Gerchman, F., Kahn, S. E., and Heckbert, S. R., Modestly elevated glucose levels during pregnancy are associated with a higher risk of future diabetes among women without gestational diabetes mellitus. *Diabetes Care*. 31(5):1037–1039, 2008.
9. Smirnakis, K. V., Chasan-Taber, L., Wolf, M., Markenson, G., Ecker, J. L., and Thadhani, R., Postpartum diabetes screening in women with a history of gestational diabetes. *Obstet. Gynecol.* 106(6):1297–1303, 2005.
10. Watkins, A. B. AIRS: A resource limited artificial immune classifier. Master thesis, Mississippi State University, 2001.
11. Polat, K., Güneş, S., and Tosun, S., Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing. *Pattern Recogn*. 39(11):2186–2193, 2006.
12. Polat, K., Yosunkaya, Ş., and Güneş, S., Comparison of different classifier algorithms on the automated detection of obstructive sleep apnea syndrome. *J. Med. Syst*. 32(3):243–250, 2008.

13. Polat, K., Yosunkaya, Ş., and Güneş, S., A new approach to diagnosing of importance degree of obstructive sleep apnea syndrome: pairwise AIRS and fuzzy-AIRS classifiers. *J. Med. Syst*. 32(6):489–497, 2008.
14. Polat, K., and Güneş, S., Principles component analysis, fuzzy weighting pre-processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer. *Expert Syst. Appl*. 34(1):214–221, 2008.
15. Nasaroui, O. F., Gonzalez, F., and Dasgupta, D., The fuzzy artificial immune system: motivation, basic concepts, and application to clustering and web profiling. *International Joint Conference on Fuzzy Systems*, pp. 711–717, 2002.
16. Şahan, S., Kodaz, H., Güneş, S., and Polat, K., A new classifier based on attribute weighted artificial immune system. *Lect. Notes Comput. Sci*. 3280:11–20, 2004.
17. Goodman, D. E. Jr., Boggess, L. C., and Watkins, A. B., Artificial immune system classification of multiple-class problems. *The Proceeding of the Artificial Neural Networks in Engineering*, pp. 179–183, 2002.
18. Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., and Moss, R. H., A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph*. 31(6):362–373, 2007.
19. Wu, G., and Chang, E. Y., KBA: kernel boundary alignment considering imbalanced data distribution. *IEEE Trans. Knowl. Data Eng*. 17(6):786–794, 2005.
20. Batuwita, R., and Palade, V., MicroPred: effective classification on pre-miRNAs for human miRNA gene prediction. *Bioinformatics*. 25 (8):989–995, 2009.
21. Tang, Y., and Chawla, N. V., SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man Cybern. Part B: Cybern*. 39(1):281–288, 2009.
22. Kubat, M., and Matwin, S., Addressing the curse of imbalanced training set: one-sided selection. *Proceeding 14th International Conference on Machine Learning (ICML'97)* 179–186, 1997.
23. Chang, C. C., Lin, C. J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
24. Hsu, C. W., and Lin, C. J., A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* 12 (2):415–425, 2002.