Breast Cancer Classification Using Hybrid Synthetic Minority Over-Sampling Technique and Artificial Immune Recognition System Algorithm

Kung Jeng Wang, and Angelia Melani Adrian

***Abstract*—** This paper proposed a hybrid method by combining Synthetic Minority Over-Sampling Technique (SMOTE) and Artificial Immune Recognition System (AIRS) to handle the imbalanced data problem that are prominent in medical data. We used the Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) datasets to compare our proposed method with other popular classifiers i.e. AIRS, CLONALG, C4.5, and BPNN. The comparison based on the accuracy, sensitivity, specificity and *G*-mean. We confirmed that the proposed method superior to other compared classifiers. Based on the experimental results, we conclude that the proposed approach can be used as an efficient method to handle imbalanced class problem. Moreover, the combination of SMOTE with classifier algorithm can improve the classification performance. Additionally, the proposed method can serve as a supplementary tool for doctors to diagnose the malignant and benign tumors early in breast cancer disease.

***Keyword*—** SMOTE, AIRS Algorithm, Imbalanced Dataset, Classification.

I. Introduction

BREAST cancer is the second leading cause of cancer death among women in the world [1]. Breast cancer considered the most common invasive cancer in women, with more than one million cases and nearly 600,000 deaths occurring worldwide annually [2]. To reduce the number of deaths, early diagnosis and treatment have been pointed at as the most reliable approach.

The highly mortality rate of breast cancer patients annually and the huge amount of patients' data could be used to extract useful information have motivated researchers to utilize data mining techniques such as classification task to help doctors in the early prognosis and diagnosis of breast cancer.

Artificial Immune Recognition System (AIRS) is a well-known algorithm used in medical classification problem. AIRS algorithm which is inspired by the immune system had been utilized for diagnose several disease such as diabetes [3], breast cancer and liver disorder [4,5]. Researchers have revealed that AIRS is a successful supervised classification algorithm based on the test which made on standard datasets and professional area [6].

However, medical data usually categorized as class imbalanced data set. This is a problem that regularly found in the real world application that can cause serious negative effect on classification performance. Medical datasets with class imbalance pose problems when less observed patterns are of higher relevance; since most of the data mining techniques tend to generalize the patterns observed over the majority data and ignored those observed over small portions of the data.

One of the powerful approaches to dealing with the class imbalance problem is Synthetic Minority Over-Sampling Technique (SMOTE). In this technique, SMOTE generates minority class within the overlapping regions. SMOTE has been widely used to solve imbalanced dataset problems in many medical area, such as medical imaging intelligence [7] and prostate cancer staging [8].

In this paper we have proposed new method, a hybrid SMOTE and AIRS algorithm (S-AIRS) to distinguish malignant (cancerous) from benign (non-cancerous) samples. Wisconsin Breast Cancer dataset (WBC) and Wisconsin Diagnostic Breast Cancer dataset (WDBC) were used as the case studies in this paper. To justify the performance of our proposed approach, we compare our result with the standard version of AIRS algorithm, Clonal selection algorithm (CLONALG), C4.5 and Back Propagation Neural Network (BPNN). Our performance measure were evaluates based on the accuracy, sensitivity, specificity, and *G*-mean.

The rest of the paper is organized as follows. In section 2, we review the related works. Section 3, presents our proposed method. After that, in section 4, shows our experimental results. Finally, we conclude the paper in section 5.

II. Related Works

A. Synthetic Minority Over-Sampling Technique

There are two main approaches for solving imbalanced data, first, is to preprocess data by under-sampling the majority

Kung Jeng Wang, PhD is with the Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, ROC. (email: kjwang@mail.ntust.edu.tw)

Angelia Melani Adrian is currently with the Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, ROC and De La Salle University, Manado 95000, INDONESIA (Corresponding Author's e-mail: D9901806@mail.ntust.edu.tw).

instance and second is to over-sampling the minority instance. SMOTE is recognized as a well-known over-sampling method. In SMOTE, the positive class is over-sampled by creating synthetic instances in the decision regions formed by the instance and its *k*-nearest neighbors [9]. SMOTE uses different ways for sample generating in continuous and categorical features. Euclidean distance is calculated for continuous samples generating, Value Distance Metric is used for nominal features. By applying SMOTE technique, it can leads to a better generalization for the classifiers.

Several researchers have utilized SMOTE technique to deal with medical imbalanced data problem. Gao et al. [9] combined SMOTE and particle swarm optimization (PSO) with Radial Basis Function as classifier to predict the survivability of patients' undergoes breast cancer surgery. They proved that SMOTE is effective to increase the significance of the positive class in the decision region and conclude that their proposed method offers a very competitive solution to other existing methods that deal with imbalanced class problem. Chandana et al. [8] investigate the performance of SMOTE and a combination of genetic algorithm (GA) and rough set (RS) to predict the stage of prostate cancer. They conclude that under-sampling and rough sets based features were identified to be most useful in improving overall performance of their system. Wang et al. [10] proposed SMOTE+PSO+C5 to enhance the effectiveness of classification for 5-year survivability of breast cancer patients with imbalanced data set. The result shows that this proposed method has the highest performance on the mentioned dataset. They conclude that standard classifier solely cannot improve the classification performance.

B. Artificial Immune Recognition System

An artificial immune system is a class of adaptive or learning algorithm inspired by the function of the biological immune system to solve the real-world problems. Artificial Immune System (AIS) have been applied to solve various difficult problems such as intrusion detection, data clustering, classifications and search problems. The field of AIS has been around for about fifteen years, and for the majority of its history has been concerned with feature extraction. Recently, AIS has been applied to broader domains such as function optimization, and in the case of AIRS is classification [11], [12].

The foremost work in AIS supervised learning algorithms is associated with the Artificial Immune Recognition System (AIRS) of Watkins and Timmis [6], which uses the concepts of artificial recognition balls (ARBs), resource limitation, memory cells, and hypermutation. ARBs are essentially B-cells supplemented with information on the 35 resources available in the system.

AIRS adopts a one-shot approach in that learning patterns (antigens) are allocated to the closest matching ARB in the pool of ARBs, followed by a competitive stage in which the ARBs either survive or die depending on their fitness with regard to capturing antigens of the right class. Resources are re-allocated throughout the ARBs depending on which ARBs survive or die. Memory cells are produced from the surviving ARBs. At the end of the one-shot approach, the memory cells adopt a *k*-nearest neighbor (KNN) voting method by presenting test samples to all memory cells and reporting the stimulation values returned by each memory cell. AIRS is composed of four main stages, which are initialization, memory cell identification and ARB generation, competition for resources and development of a candidate memory cell and finally memory cell introduction.

In data mining area, KNN is one of the successful techniques used in classification task [13] and has been widely applied to solve various classification problems. It's become a popular classifier due to the simplicity and high convergence speed. Despite its benefit, KNN also have drawback for the large memory requirement. KNN using Euclidean distance to calculate the difference between attributes for continuous data [14].

Remarkable performance of AIRS with conjunction of classifiers algorithms have been exposed by some researchers. Saidi et al. [13] proposed Modified AIRS2 (MAIRS2) where the KNN algorithm was replace with the fuzzy K-NN to improve the diagnostic accuracy of diabetes diseases. This combination of AIRS2 and fuzzy K-NN shows better performance in accuracy than the classical AIRS2. Polat and Güneş [4] proposed Fuzzy-AIRS, to classify three well-known medical data sets, the Wisconsin breast cancer data set (WBCD), the Pima Indians diabetes data set and the ECG arrhythmia data set. The results show that Fuzzy-AIRS can be used as an effective classifier for medical problems. Polat et al. [5] utilized Fuzzy-Logic on AIRS and used it as a classifier in the diagnosis of Breast Cancer and Liver Disorders. The result shows that classification time in Fuzzy-AIRS was reduced about 70% for both datasets.

III. HYBRID SMOTE AND AIRS FOR BREAST CANCER CLASSIFICATION

A. Overview

Our proposed method consists of two main parts: first, data sampling by SMOTE and second, classification by AIRS algorithm. Data sampling by SMOTE provides a mechanism to eliminates the imbalanced that exists in the data by over-sampling the minority class, then using the resulting balanced data set, the AIRS algorithm in charge for the classification task. Fig. 1 shows our proposed method procedure.

B. Hybrid SMOTE and AIRS Algorithm

1) SMOTE Technique

Step 1: Take majority vote between the feature vector under consideration and its *k* nearest neighbors for the nominal feature value. In the case of a tie, choose at random.

Step 2: Assign that value to the new synthetic minority class sample. Next step is to classify the data using AIRS classifier.

2) Antigens initialization

To initialize the population first of all we normalized all

items in the data set such that the Euclidian distance between the feature vectors of any two items is in the range of [0,1]. After normalization, we calculate the affinity threshold according to (1).

$$affinity\_threshold = \frac{\sum_{i=1}^{n}\sum_{j=j+1}^{n} affinity(ag_1, ag_2)}{n(n-1)/2} \quad (1)$$

Where n is the number of training data items (antigens), ag_i and ag_j are the i^{th} *and* j^{th} training antigen in the antigen training data set.

The affinity for each antigen is calculated based on (2)

$$affinity(ag_1, ag_2) = 1 - Euclidean\_dist(ag_1, ag_2) \quad (2)$$

Then we create a random base called the memory pool (*M)* and the ARB pool (*P)* from the training data.

3) Memory cell identification and ARB generation

Clonal Expansion: For each element of *M* determines their affinity to the antigenic pattern, which resides in the same class. Select highest affinity memory cell (mc_{match}) based on (3) and (4) and *clone mc* in proportion to its antigenic affinity to add to the set of ARBs (*P*)

$$mc_{match} = \arg\max(stimulation(mc, ag)) \quad (3)$$

$$stimulation(mc, ag_2) = \begin{cases} affinity(mc, ag) if mc.class = ag.class \\ 1 - affinity \end{cases} \quad (4)$$

Affinity Maturation: Mutate each ARB descendant of mc_{match}. Place each mutated ARB into *P*.

4) Competition for resources and development of a candidate memory cell

a) *Metadynamics of ARBs*: Process each ARBs through the resource allocation mechanism. This will result in some ARB death, and ultimately controls the population. Calculate the average stimulation for each ARB, and check for stopping criteria.

$$s_i = \frac{\sum_{j=1}^{|ARB_i|} arb_j.stimulation}{|ARB_i|} \; arb_j \in ARB_i \quad (5)$$

b) *Clonal Expansion and Affinity Maturation*: Clone and mutate a randomly selected subset of the ARBs left in P based in proportion to their stimulation level.

c) *Cycle:* While the average stimulation value of each ARB of the same class as the antigen is less then given stimulation threshold then repeat from section 4 a.

5) Memory cell introduction

Metadynamics of Memory Cells: Select the highest affinity ARB from the last antigenic interaction. If the affinity of this ARB with the antigenic pattern is better than that of the previously identified best memory cell mc then add the candidate (mc-candidate) to memory set *M*. additionally, if the affinity of mc_{match} and mc-candidate is below the affinity threshold, and then remove mc_{match} from *M*.

Cycle: repeat steps in section 3-5 until all antigenic patterns have been presented.

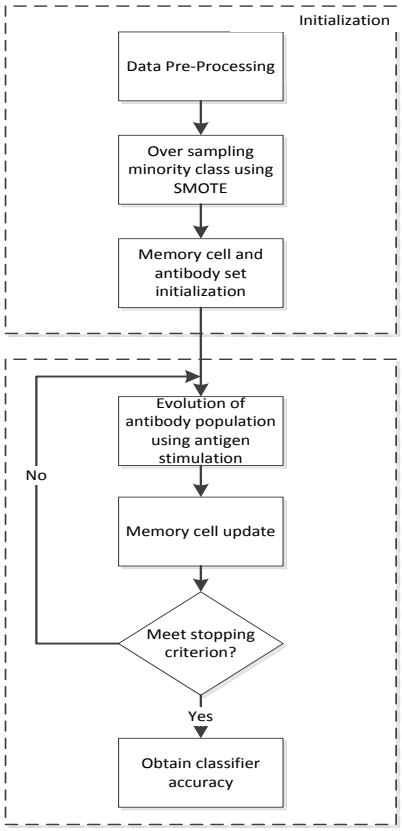

Fig. 1 Hybrid SMOTE+AIRS for breast cancer diagnosis

IV. Experimental Results

A. Dataset and Experiments Setup

Two medical datasets were used to evaluate our proposed method i.e. WBC and WDBC. These datasets were downloaded from UCI machine learning repository. We adopted 5-folds cross validation strategy to guarantee impartial comparison of the classification results and avoiding generating random results. Table I summarizes the datasets' characteristic.

Our proposed method was coded in Java and run in a Core 2 Duo P8400 (2.2 GHz) PC equipped with 4 GB of RAM under Windows 7 environment. WEKA software is utilized to perform SMOTE on the imbalanced dataset. The parameters we used for SMOTE are set to WEKA's default setting which are: the number of *k* in nearest neighbour: 5, the percentage of instance to create: 100% and the seed for random sampling: 1.

While parameter we used for AIRS are as follows: stimulation threshold: 0.9, mutation rate: 0.1, clonal rate: 10, hypermutation rate: 2, affinity threshold: 0.2, *k*-nearest neighbour: 5. We used 5-folds cross validation as the stopping criteria.

B. Performance evaluation

In order to evaluate the effectiveness of the hybrid SMOTE and AIRS (S-AIRS), we compare the result of our proposed

method with the standard version of AIRS, CLONALG, C4.5 and BPNN according to some measures. These measures are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Sensitivity(SE) = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity(SP) = \frac{TN}{TN + FP} \quad (8)$$

$$G - mean = \sqrt{sensitivity \times specificity} \quad (9)$$

Where *TP* denotes true positives, *TN* denotes true negatives, *FP* denotes false positives, and *FN* denotes false negatives. These values are often displayed in a confusion matrix as presented in Table II.

C. Result and Discussion

In Table III we present the classification accuracy and the comparison for other compared methods. This table shows S-AIRS method superior to other methods in terms of classification accuracy. Table IV and Table V show the comparison of sensitivity and specificity and G-mean respectively. The result shows that our proposed method achieved the highest G-mean for all compared methods.

S-AIRS sensitivity, specificity and G-mean are slightly higher than AIRS. The result shows that accuracy and sensitivity, specificity and G-mean are in the same trend, this is due the SMOTE contribution in this dataset.

Based on the result, SMOTE technique can enhance classifier performance with higher accuracy. It could be common sense that classification algorithm can give higher accuracy if applied on balance dataset, but due to the real case data used to be imbalance data, then utilize over sampling method is worth to try.

TABLE I
DATASET CHARACTERISTICS

Dataser	Majority Instance	Minority Instance	Oversampled Rate
WBC	458	241	100
WDBC	357	212	100

TABLE II
CONFUSION MATRIX

		Predicted Class	
		Benign	Malignant
Actual Class	Benign	TP	FN
	Malignant	FP	TN

V. CONCLUSION

In this paper, we present a method for tackle the class imbalanced problem called SAIRS. Our proposed method is consists of two steps, first is the over-sampling minority class by using SMOTE technique and second the classification task performed by AIRS algorithm.

To evaluate the performance of the proposed approach, we use two WBC and WDBC dataset from UCI databank. The experimental result revealed that the proposed method S-AIRS outperformed other classifier i.e.: Clonal selection, C4.5 and BPNN. Moreover, the results shows that the combination of SMOTE and classifier algorithm can enhance the classifier performance. This result can lead us to conclude that classifier can perform better on the balanced dataset.

Based on this work, we determined that the proposed method can be applied as a diagnosis tool in conjunction with other medical test for early detection of breast cancer. The result can serve as supplementary tools to diagnose and design the treatment.

TABLE III
CLASSIFICATION ACCURACY

Classifier	AIRS	S-AIRS	Clonal G	C4.5	BPNN
WBC	0.9585	**0.9691**	0.9442	0.9399	0.9571
WDBC	0.9344	**0.9652**	0.5993	0.9385	0.9531

TABLE IV
CLASSIFICATION SENSITIVITY AND SPECIFICITY

Dataset	AIRS		S-AIRS		Clonal G		C4.5		BPNN	
	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP
WBC	1.0	1.0	**1.0**	**1.0**	0.9	1.0	1.0	0.9	1.0	0.9
WDBC	0.9	1.0	**1.0**	**1.0**	0.5	0.7	0.9	1.0	0.9	1.0

TABLE V
CLASSIFICATION *G*-MEAN

Classifier	AIRS	S-AIRS	Clonal G	C4.5	BPNN
WBC	0.9593	**0.9703**	0.9482	0.9341	0.9500
WDBC	0.9156	**0.9609**	0.5468	0.9323	0.9512

REFERENCES

[1] U.S. Cancer Statistics Working Group, "United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report," Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute, 2012.

[2] Lyon IAfRoC, "World Cancer Report," International Agency for Research on Cancer Press, 2003, pp. 188-193.

[3] M. Saidi, M. Chikh, N. Settouti, "Automatic identification of diabetes diseases using a Modified Artificial Immune Recognition System2 (MAIRS2)," in *Proceedings of the Int. Conf. on Computer Science and its Applications*, 2011.

[4] K. Polat, S. Günes, "An improved approach to medical data sets classification: Artificial immune recognition system with fuzzy resource allocation mechanism," *Expert Systems*, vol. 24 (4), 2007, pp. 252-270.

[5] K. Polat, S. Sahan, H. Kodaz, S. Günes, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism," *Expert Systems with Applications*, vol. 32 (1) , 2007, pp. 172-183.

[6] A. Watkins, J. Timmis, L. Boggess, "Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm," *Genetic Programming and Evolvable Machines*, vol. 5 (3), 2004, pp. 291-317.

[7] W. Juanjuan, X. Mantao, W. Hui, W., Z. Jiwu, "Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding," in *Int. Conf. on Signal Processing Proceedings*, 2007, 4129201.

[8] S. Chandana, H. Leung, K. Trpkov, "Staging of prostate cancer using automatic feature selection, sampling and Dempster-Shafer fusion," *Cancer Informatics*,vol. 7, 2009, pp. 57-73.

[9] M. Gao, X. Hong, S. Chen, C.J. Harris, "On combination of SMOTE and particle swarm optimization based radial basis function classifier for imbalanced problems," in *Proceedings of the Int. Joint Conf. on Neural Networks*, 2011, 6033353, pp. 1146-1153.

[10] K.J. Wang, B. Makond, K.H. Chen, "A combination algorithm for 5-year survivability of breast cancer patient," in *Proceeding of Translational Bioinformatics Conf*., 2012.

[11] J. Brownlee, "Artificial immune recognition system (AIRS) A review and analysis," CISCP, Faculty of ICT, Swinburne University of Technology, Australia, Technical Report 1-02, 2005.

[12] J. Timmis, M. Neal, J. Hunt, "An artificial immune system for data analysis," *Biosystems* vol. 55, 2000, pp. 143-150.

[13] M. Saidi, M. Chikh, N. Settouti, "Automatic identification of diabetes diseases using a Modified Artificial Immune Recognition System2 (MAIRS2)," in *Proceedings of the Int. Conf. on Computer Science and its Applications*, 2011.

[14] M. Shouman, T. Turner, R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *Int. J. of Information and Communication Technology Education*, vol. 2 (3), 2012. pp. 220 – 223.