Hindawi Publishing Corporation
Journal of Artificial Evolution and Applications
Volume 2009, Article ID 130498, 12 pages
doi:10.1155/2009/130498

Research Article

Conserved Self Pattern Recognition Algorithm with Novel Detection Strategy Applied to Breast Cancer Diagnosis

Senhua Yu and Dipankar Dasgupta

Department of Computer Science, University of Memphis, Memphis, TN 38152, USA

Correspondence should be addressed to Senhua Yu, senhuayu@gmail.com

Received 1 October 2008; Accepted 22 May 2009

Recommended by Marylyn Ritchie

This paper presents a novel approach based on an improved Conserved Self Pattern Recognition Algorithm to analyze cytological characteristics of breast fine-needle aspirates (FNAs) for clinical breast cancer diagnosis. A novel detection strategy by coupling domain knowledge and randomized methods is proposed to resolve conflicts on anomaly detection between two types of detectors investigated in our earlier work on Conserved Self Pattern Recognition Algorithm (CSPRA). The improved CSPRA is applied to detect the malignant cases using clinical breast cancer data collected by Dr. Wolberg (1990), and the results are evaluated for performance measure (detection rate and false alarm rate). Results show that our approach has promising performance on breast cancer diagnosis and great potential in the area of clinical diagnosis. Effects of parameters setting in the CSPRA are discussed, and the experimental results are compared with the previous works.

1. Introduction

Normally, breast cells grow and then rest in cycles. The periods of growth and rest in each cell are controlled by genes in the cell's nucleus. Genes sometimes develop abnormalities, which cause to lose their ability to control the cycle resulting in uncontrolled growth of breast cells (cancer). Breast cancer is the most common cancer and the second largest cause of cancer deaths among women [1]. Based on the estimation (from www.BreastCancer.org) in 2007, there are about 178 480 new cases of invasive breast cancer and 62 030 new cases of noninvasive breast cancer diagnosed in the United States. Early detection of this disease via accurate diagnosis and treatment can greatly improve the chances for survival.

Most breast cancers are symptomatic of lump in the breast but the majority of breast lumps are benign. Therefore, it is important to distinguish benign lumps from malignant ones. The methods for diagnosing breast cancer include mammography, fine-needle aspirates (FNAs) with visual interpretation, and surgical biopsy. The detection ability of both mammography and fine-needle aspirates with visual interpretation to correctly diagnose breast cancer is unstable [2, 3]. Surgical biopsy, although accurate, is invasive, time consuming, and costly. The anticipated course of the cancer not only determines whether chemotherapy is needed but also affects the mental state and personal goal of the patient. Therefore, developing an accurate, efficient, and inexpensive method for breast cancer diagnosis is an important and challenging goal for the treatment of this disease.

The great economic and social values of breast cancer diagnosis have attracted many researchers. Some methods such as linear programming [4, 5], neural network [6, 7], and ant colony-based system [8] were applied to breast cancer diagnosis. Over the last decade, immunity-based approaches have been applied to solve problems in a wide variety of domains such as anomaly detection, pattern recognition, data mining, computer security, adaptive control, and fault detection [9]. The majority of breast lumps are benign, which could in practice provide a large training data set of normal class. Hence, Artificial Immune System (AIS) can be applied to breast cancer diagnosis by taking advantage of one-class classification. In this paper, we describe an improved Conserved Self Pattern Recognition Algorithm (CSPRA) for breast cancer diagnosis. The details of the algorithm are described in Section 2 and Section 3 outlines

the breast cancer diagnosis problem. Section 4 describes how the algorithm can be applied to the breast cancer diagnosis problem and reports some experimental results. Section 5 provides discussions on algorithmic parameters and comparative results. The last section provides conclusive remarks and future work.

2. A Novel Detection Strategy in Conserved Self Pattern Recognition Algorithm (CSPRA)

2.1. A Brief Overview of CSPRA. The immune system plays roles by discriminating self (defined early in life) and nonself (anything that comes later), tolerating self, and attacking nonself [10]. This elegantly simple idea, known as the Self-Nonself model, has dominated the field for over 50 years. However, it has failed to explain a great number of findings such as alerted self, pregnancy, and aging [10]. Pattern Recognition Receptors (PRRs) model was published in 1989 to accommodate incompatible new findings [11]. The PRRs model suggested that Antigen Presenting Cells (APCs) can recognize evolving pathogens. The lymphocytes (T cell or B cell) would die if it recognized antigen (Signal 1) without the costimulation from APC (Signal 2) but APCs do not costimulate unless activated via encoded PRRs that recognize conserved pathogen-associated molecular patterns (PAMPs) on bacteria [11]. Inspired by the biological PRRs model, we recently proposed a novel algorithm called Conserved Self Pattern Recognition Algorithm (CSPRA) [12]. The basic steps involved in the CSPRA are as follows.

(1) Build up the "Self" training samples from the collected normal data and store the samples as a multiset S_0 of equal length strings, L over a finite alphabet.

(2) Establish a model of normal behavior by generating the set of T detectors (R_1) and a specific APC detector (R_2), respectively. The negative selection strategy [13] is employed in the generation of T detectors. However, the generation of the specific APC detector includes two major steps.

   (a) Based on the relationship between the antigen objects and the dimensions of their feature space, define the *conserved self pattern*. It can be predefined from the empirical data based on the scientists' lab results. This paper introduces a new technique for finding conserved self pattern, as described in Section 4.1.

   (b) Within the conserved self pattern consisting of the features located in loc 1, loc 2, ..., generate APC detector $R_2\{\langle \text{loc } 1, \text{min}, \text{max}, \text{mean}\rangle, \langle \text{loc } 2, \text{min}, \text{max}, \text{mean}\rangle, \ldots\}$ by calculating maximum, minimum, and mean of all of the values in the features (or descriptors) of loc 1, loc 2, ..., respectively.

(3) Monitor the system by detecting anomaly in the incoming new data in the testing data set S_1 using the generated T detectors and APC detector in R_1 and R_2.

   The matching rule (Euclidean distance) is used for T detectors to report anomaly. The distance between APC detector, and the new sample is calculated by (1). If it is *greater* than the predefined threshold, then the anomaly is detected by APC detector. In (1), w is the number of the dimensions for the conserved pattern; m_i and n_i represent the lower and upper bounds of the ith attribute in the entire training data, respectively; $p = (p_1, p_2, \ldots, p_w)$, p_i is the value of the ith attribute for the antigen object to be examined; $d = (d_1, d_2, \ldots, d_w)$, d_i is the mean of all of the values in the ith attribute in the entire training data:

$$\text{Dist}(p, d) = \sum_{i=1}^{w} \frac{|p_i - d_i|}{m_i - n_i}. \quad (1)$$

(4) The new sample (antigen) is firstly checked with T detectors. The costimulation of APC detector is conducted if and only if *both* of the following conditions are fulfilled during the phase of T cell detection.

   (a) The affinity between the T cell detector and the new sample is very low, that is, the Euclidean distance between the T cell detector and the incoming sample is greater than predefined suspicious threshold.

   (b) The decision for abnormal (non-self) is made based on the other antigen epitope instead of the antigen peptide, where the conserved self pattern is located.

   The new sample satisfying the above conditions is named *suspicious antigen*. APC detector finally determines whether it is anomaly in this case.

Now, we extend the proposal in [12] aiming at resolving the conflicts in anomaly detection between two types of detectors in the detection phase of the algorithm. By wisely using domain knowledge and randomized method, the novel detection strategy we present in this paper has been proven to greatly enhance the efficiency for anomaly detection by the unpublished results when testing with multiple data sets. In this paper, we center on the application of this modified Conserve Self Pattern Recognition Algorithm to breast cancer diagnosis. In this section, we are going to describe the novel detection strategy, the interesting readers can refer to [12] for the details of the algorithm.

In biology, the PRRs model added additional layer of pathogen-associated molecular patterns (PAMPs) to the self-nonself model. Inspired by this metaphor, our earlier proposal in [12] combined both APCs Pattern Recognition and T cell Negative Selection to detect anomalies in new samples, which had been proven to efficiently reduce high false positive error rate that often occurred in Negative Selection Algorithm (NSA). By exploring this idea, a question is naturally raised: if the conflicts on anomaly decision happen when testing the new samples using APC detector and T cell detector, respectively, for example, T cell detector

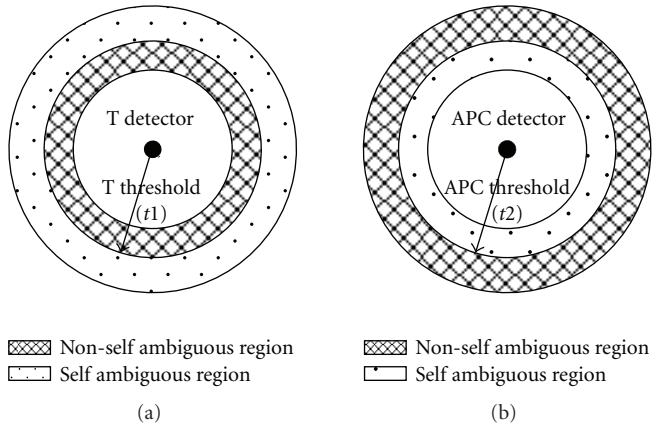

Figure 1: Interpretations of ambiguous boundary on anomaly detection.

recognizes the new sample as "Nonself" whereas the same sample is classified as "Self" by APC detector, how does the system make the final decision for the new sample? The solution in the previous proposal mirrors the biological metaphor: APCs are quiescent until they are activated via encoded PRRs that recognize conserved PAMPs [11]. The costimulation of APC detector will not be conducted until the detection from T cell detectors becomes unsure, that is, the *suspicious antigen* is encountered in the system. Although this solution shows its strength in terms of algorithmic complexity, its performance relies on the application domain since the definition of suspicious antigen is not always in accordance with a specific application. Our recent work proposes the mathematical methods to resolve the conflicts.

2.2. Ambiguous Boundary in Anomaly Detection. An important concept in our proposal is *ambiguous boundary*. Figure 1 illustrates the interpretations of ambiguous boundaries for T cell detection and APC detection, respectively. The representation for the single T detector in Figure 1(a) is different in two detection cases.

(i) If the new sample matches with *any* of the T detectors, that is, the distance between the new sample and the current T detector is less than the predefined threshold, then the new sample is detected as "non-self." The single T detector in Figure 1(a) represents the one in the set of T detectors that recognizes the new sample that is currently being tested.

(ii) The new sample is detected as "self" if it *does not* match *any* T detectors. In this case, the single T detector in Figure 1(a) is the one in the set of T detectors that is the nearest to the new sample that is currently being tested. From the viewpoint of algorithm implementation, the Euclidean distance between the new sample and each T detector is calculated while the testing sample is checked against all T detectors but only the shortest distance is returned.

Now, it becomes very straightforward to explain the ambiguous boundary for both T detector and APC detector. The *ambiguous coefficient* α in this proposal is applied to adjust the range of the ambiguous region, that is, to increase/decrease the shadow area in Figure 1.

(i) As shown in Figure 1(a), the ambiguous region for T detector is a ring-shaped region contained between the two circles (upper ambiguous boundary and lower ambiguous boundary) with the radius of $(1+\alpha_t)^* t_1$ (outer circle) and $(1-\alpha_t)^* t_1$ (inner circle), where t_1 represents the threshold for T detector activation and α_t is the ambiguous coefficient for T detector detection.

(ii) Similarly, as shown in Figure 1(b), the ambiguous region for APC detector is also a ring-shaped region contained between the two circles with the radius of $(1+\alpha_p)^* t_2$ (outer circle) and $(1-\alpha_p)^* t_2$ (inner circle), where t_2 represents the threshold for APC detector activation and α_p is the ambiguous coefficient for APC detector detection.

2.3. A Novel Detection Strategy to Resolve the Conflicts on Anomaly Detection. Once the ambiguous boundary is defined, the following rules are set to resolve the conflicts in anomaly decision between T detector and APC detector by using domain knowledge and randomized method.

2.3.1. Domain Knowledge. When the antigen (the new incoming sample) is loaded into the system, we check the matching rule of Euclidean distance against each T detector. If the matching is found, we keep the matching distance in record. If the matching is *not* found after *all* T detectors are checked, we only keep the distance to the nearest T detectors in record. Similarly, the matching distance is also kept in record when the antigen is checked against the APC detector. The information from the matching distances for both T detector and APC detector obtained in the detection phase is considered as valuable domain knowledge, which is used in the proposed detection strategy to resolve the conflicts in the following cases.

Case 1. The anomaly decision from APC detector is granted but the results from T detector detection are discarded if and only if *both* of the following conditions are fulfilled:

(i) the new sample falls into the ambiguous region that is predefined by the ambiguous boundary when testing with T detector;

(ii) the new sample is *not* in the ambiguous region when testing with APC detector.

Case 2. It is opposite of Case 1. T detector finally determines whether the new sample is anomalous but APC detection is simply ignored if and only if *both* of the following conditions are fulfilled:

(i) the new sample is *not* in the ambiguous region when testing with T detector;

(ii) the new sample is in the ambiguous region when testing with APC detector.

2.3.2. Randomized Method. Randomized method is selected to resolve the conflicts for all other cases rather than Cases 1 and 2 described in Section 2.3.1. In other words, randomized method is used under one of the following conditions as (1) the new sample falls into the ambiguous region when it is matched against with either T detector or APC detector; (2) the new sample does *not* fall into the ambiguous region when it is matched against with either T cell detector or APC detector.

When implementing this algorithm, a list $L(n_1, n_2, n_3, n_4, n_5 \ldots)$ is established in which to store these ambiguous testing samples while the system loops through every testing sample. Another parameter called *pattern weight w* is introduced in the algorithm to control the randomized decision making. When the loop is terminated, the algorithm randomly picks up n^*w samples, provided that the list L contains n ambiguous samples, to form a new sublist $l(m_1, m_2, m_3, \ldots)$. For every sample in the sublist l, APC detector *solely* determines whether it is anomalous regardless of the results from T detector decision. Anomaly decision for the remaining $n^*(1 - w)$ samples in the parent list L follows T detector decision process.

2.4. Pseudocode for the Detection Algorithm. Listed below is the pseudocode for the detection algorithm that reflects the novel detection strategy we have outlined in the previous sections.

As noted in Algorithm 1, in lines 2 and 3, each sample is examined by T detector and APC detector, respectively. We keep the distance (d_t and d_p) in record when we check the matching rule for both APC detector and T detector, which are used to determine the resolution of the conflicts, as seen between lines 9 and 15. In lines 14 and 15, we save the index (*i*), T detector detection result (*TDecision*), and APC detector detection result (*APCDecision*) to the defined *struct* for each ambiguous testing sample (*ambiguousAg*), and then add *ambiguousAg* to the list of the ambiguous testing samples (*randList*). Hence, we no longer loop through all testing samples again when we determine the anomaly for the new sample using randomized method in lines 31 and 34, which greatly enhances the algorithm efficiency. As shown in line 20 in Algorithm 1, in the list of the ambiguous testing samples, the total number of the samples to be determined by APC conserved pattern recognition is affected by the pattern weight *w*. Line 23 through line 35 show how to determine the anomalies using randomized method. The pseudocode shown in Algorithm 1 is only part of the algorithm implementation showing the difference from our previous proposal in the detection phase for this algorithm. Readers interested in this information can refer to [12] for the entire pseudocode for the algorithm implementation.

3. Breast Cancer Diagnosis Problem

This breast cancer database is downloaded from the UCI machine learning repository [14], which was collected by Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison, USA [4]. The dataset is comprised of elements that consist of various scalar observations. The total number of the original samples is 699 but 16 samples with missing values are removed to construct a new dataset with 683 samples that are actually used in our experiments. The dataset contains two classes referring to benign and malignant samples. There are 444 samples in the dataset that are assigned to benign, and the other 239 samples are malignant. The original dataset contains 11 attributes including both sample id number and class label, which are removed in the actual dataset that are used in our experiments. The remaining 9 attributes represent 9 cytological characteristics of breast fine-needle aspirates (FNAs), as shown in Table 1. The cytological characteristics of breast FNAs were valued on a scale of one to ten, with one being the closest to benign and ten the most malignant.

Samples arrived periodically as Dr. Wolberg reported his clinical cases [4, 5]. The original database therefore includes the information about this chronological grouping of the data, having been removed from the data itself. Some brief statistical analysis is presented in Table 2. The calculation of class correlation in Table 2 is introduced in Section 4.1.

Prior to the experiments, we normalize the raw data by columns in the range of [0,1] with max-min normalization, as shown in

$$f(x) = \frac{x - \min(\text{column})}{\max(\text{column}) - \min(\text{column})}. \quad (2)$$

4. Application of CSPRA to Breast Cancer Diagnosis

In this section we begin by describing the data preprocessing, then further to discuss the algorithm parameters and report the experimental results showing that the algorithm is able to predict the breast cancer quite efficiently.

4.1. Finding Conserved Self pattern. The first task to use CSPRA is to find the conserved self pattern based on training data. The *Pearson Product Moment Correlation Coefficient* that was developed by Karl Pearson is the most widely used

```
//detection phase
S: set of testing samples
t1: T detector threshold
t2: APC detector threshold
αt: ambiguous coefficient for T detector detection
αp: ambiguous coefficient for APC pattern detection
dt: distance between T detector and the current sample
dp: distance between APC detector and the current sample
w: pattern weight used in random decision
Decision*: bool variable for the final detecting conclusion for the new sample
TDecision*: bool variable for the detecting conclusion from T detector
APCDecision*: bool variable for the detecting conclusion from APC detector
ambiguousAg: struct for each new sample to be decided on anomaly with randomized method
randList: list of ambiguousAg
*For these bool variables, true for anomaly and false for self
(1) for every si in S = {si, i = 1, 2, ...}
(2)       TDecision = CheckWithTDetector(si, t1, dt)
(3)       APCDecision = CheckWithAPCDetector(si, t2, dp)
( 4)      if(TDecision && APCDecision)
(5)                 Decision = true;
(6)       else if((!TDecision) && (!APCDecision))
(7)                 Decision = false;
(8)       else
(9)                 if( ((dt > (1 + αt)*t1)||(dt < (1 − αt)*t1)) &&
                    ((dp > (1 − αp)*t2) && (dp < (1 + αp)*t2))
(10)                      Decision = TDecision
(11)              else if( ((dt < (1 + αt)*t1) && (dt > (1 − αt)*t1)) &&
                    ((dp > (1 + αp)*t2)||(dp < (1 − αp)*t2))
(12)                      Decision = APCDecision
(13)                else
(14)                      save i, TDecision, and APCDecision to the struct ambiguousAg
(15)                      Add ambiguousAg to the list randList
(16)                end else
(17)       end else
(18) end for
(19) int total_ambiguous = size of the list randList
(20) int total_apc_decided = (int) total_ambiguous *w
(21) APCList: list of the index of the samples to be decided by APC detector in randList
(22) TList: list of the index of the samples to be decided by T detector in randList
(23) while(size of APCList < total_apc_decided)
(24)       int val = rand()%total_ambiguous
(25)       if(val doesn't exist in APCList)
(26)                Add val to APCList
(27) end while
(28) for(int i = 0; i < total_ambiguous; i ++)
           if(i doesn't exist in APCList)
                    Add i to TList
(29) end for
(30) for(int i = 0; i < size of APCList; i ++)
(31)       Decision = randList[APCList[i]].APCDecision
(32) end for
(33) for(int i = 0; i < size of TList; i ++)
(34)       Decision = randList[TList[i]].TDecision
(35) end for
```

Algorithm 1: Detection algorithm with new decision strategy.

Table 1: Attributes in the breast cancer databases.

No.	Attribute
0	Clump thickness
1	Uniformity of cell size
2	Uniformity of cell shape
3	Marginal adhesion
4	Single epithelial cell size
5	Bare nuclei
6	Bland chromatin
7	Normal nucleoli
8	Mitoses

Table 2: Statistical analysis of attribute values in 683 samples.

Attribute no.	Mean	Standard deviation	Class correlation
0	4.44	2.82	0.7148
1	3.15	3.07	0.8208
2	3.22	2.99	0.8219
3	2.83	2.86	0.7063
4	3.23	2.22	0.6910
5	3.54	3.64	0.8227
6	3.44	2.45	0.7582
7	2.87	3.05	0.7187
8	1.60	1.73	0.4234

measure of correlation between two variables X and Y [15]. The correlation coefficient r can be simply described as the sum of the product of the Z-scores for the two variables divided by the number of scores as shown in

$$r = \frac{\sum z_X z_Y}{N}. \tag{3}$$

However, it is fairly difficult to calculate the correlation coefficient using (3). We use the computational formula that is mathematically identical but is much easier to use, as shown in (4), to calculate the Pearsonian r between the values (X) in the column of each attribute and their corresponding class labels (Y) in breast cancer data. The corresponding class correlations for different sample size in breast cancer data are listed in Table 3

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum Y)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}. \tag{4}$$

From the results in Table 3, the cytological characteristics of breast fine-needle aspirates such as uniformity of bare nuclei, uniformity of cell shape and cell size are ranked first three places in terms of class correlation. Hence, the subset from 1st, 2nd, and 5th (zero-based index) dimension of the original dimensions in training data is considered as *conserved self pattern*.

As noted, the abnormal samples (malignant samples for this application) are required if the conserved self pattern is defined by computing the class correlation. However, as shown in Table 3, the same conclusion about conserved self pattern is obtained when we calculate the class correlation using 100% currently collected samples, 50% of the samples from each class, and 25% of the samples from each class. It indicates that the conserved self pattern can be found with Pearsonian r method even if the samples from abnormal behaviors are smaller. With this observation, the Pearsonian r method to find conserved self pattern becomes more practical since it does not require the collection of a large number of the abnormal samples.

4.2. Effectiveness Measurement. It is a step forward from the previous works that we consider not only to what extent the system is able to detect the anomalies (malignant samples) but also to what extent the system possibly misclassifies the normal samples (benign samples) when we evaluate the system performance. Two measures of effectiveness for detecting anomaly are calculated as follows:

$$\begin{aligned} \text{DetectionRate} &= \frac{TP}{TP + FN}, \\ \text{FalseAlarmRate} &= \frac{FP}{FP + TN}, \end{aligned} \tag{5}$$

where TP, TN, FP, and FN are the counts of *true positives* (anomalous elements identified as anomalous), *true negatives* (normal elements identified as normal), *false positives* (normal elements identified as anomalous), and *false negatives* (anomalous elements identified as normal).

Various system parameters, that is, detector thresholds, control the sensitivity of the system. By employing various strategies to change the parameter values, different values for detection rate and false alarm rate are obtained that are used for plotting the Receiver Operating Characteristics (ROCs) curve, which reflects the tradeoff between false alarm rate and detection rate. Since high detection rate and low false alarm rate are the two goals between which balance is needed, the ROC curve is used in this paper to evaluate the influence of the various parameters on the system performance for breast cancer diagnosis.

4.3. Experimental Results. The experiments are carefully designed to objectively evaluate the performance of the algorithm and analyze the experimental results. By combining the idea of k-fold cross validation and the features of one-class classification of our method, three groups of training data, as seen in Table 5, are generated with the following schema to effectively reduce the ordering effect that perhaps exists in the original collections.

(i) *Scheme 1* for training data of 100% benign samples: the first training data are obtained by shuffling the original benign samples; the reordered samples are shuffled again to output the second training data. Repeating this step 10 times generates 10 different training data sets.

(ii) *Scheme 2* for training data of 25% benign samples: pick up 3 training data sets that are generated in the last three rounds of the scheme 1 to guarantee that the collected samples have been fully randomly reordered. Each dataset is partitioned into 4 subsamples, and thus total 12 subsamples are obtained.

Table 3: Attribute class correlation in breast cancer data.

No.	Attribute	Class correlation		
		100% sample	50% sample	25% sample
0	Clump thickness	0.7148	0.7513	0.7184
1	Uniformity of cell size	0.8208	0.8024	0.7598
2	Uniformity of cell shape	0.8219	0.8156	0.7905
3	Marginal adhesion	0.7063	0.6848	0.6241
4	Single epithelial cell size	0.6910	0.6889	0.6665
5	Bare nuclei	0.8227	0.8029	0.7238
6	Bland chromatin	0.7582	0.6823	0.6046
7	Normal nucleoli	0.7187	0.7025	0.6937
8	Mitoses	0.4234	0.4509	0.4439

10 subsamples are randomly selected from the total 12 subsamples to make up 10 training data of 25% benign samples.

(iii) *Scheme 3* for training data of 50% benign samples: pick up 5 randomly reordered training data sets that are generated in the last five rounds of the scheme 1. Each dataset is partitioned into 2 subsamples, and thus total 10 subsamples are obtained. The 10 subsamples are the actual 10 training data of 50% benign samples.

When all the available benign samples are used to train the enhanced CSPRA, the total 683 samples, including the same benign samples used in the training phase, are considered as testing data. Although such testing case can demonstrate the system's capability to recognize known normal data, the false alarm rate could be deceptive because the resulted model may overfit the training data. Therefore, the training data are removed in our testing with training data of 50% benign samples and 25% benign samples. When 50% benign samples are used as training data, the remaining 50% benign samples and all malignant samples are combined as testing data. Similarly, if the training data are 25% benign samples, the testing data include the remaining 75% benign samples and all malignant samples.

Through repeatedly running the algorithm and observing the generated results, the parameter settings listed in Table 4 produce better results. As noted, when 25% benign samples are used as training data, 10 different randomly reordered training data with 25% benign samples, known as *subsamples,* are generated from the original collections. The 10 subsamples are used to train our proposed model, respectively. Since T detectors are randomly generated with negative selection in the CSPRA, different values for detection rate and false alarm rates are observed. Therefore, each subsample undergoes 100 repeated runs, and the statistics for these repeated experiments is summarized and recorded. When the experiments with all 10 subsamples are completed, the average values of the statistics from 10 subsamples are calculated, which are reported in Table 5. The same procedures are applied in our experiments when 50% benign samples and 100% benign samples are used as training data, respectively.

Table 4: The values of the parameters that are used in the experiments.

Parameters	Value
T detector threshold (t_1)	0.1
APC detector threshold (t_2)	0.5
T ambiguous coefficient (α_t)	0.5
APC ambiguous coefficient (α_p)	0.4
Pattern weight (w)	0.7
T detector size	300
APC detector size	1
Sliding window size	3

As shown in Table 5, the performance of the enhanced CSPRA regarding to breast cancer diagnosis is very promising, which indicates the great potential of the AIS methods in the area of clinical diagnosis. The experimental results also show that the detection accuracy with the enhanced CSPRA is very high even if we use the smaller training data. Table 6 shows the best experimental results we picked out from 100 repeated experiments for each training data of 100%, 50%, 25% benign samples by comprehensively considering the balance of the detection rate and false alarm rate. To clarify this, the best result is actually chosen among the experiments having the highest accuracy rate in the 100 repeated experiments. Because there are 10 randomly-reordered subsamples for each group of training data, the values of detection rate and false alarm rate reported in Table 6 are the average of the best results obtained from the experiments with 10 subsamples. As seen in Table 6, the detection rate is 98.74% companying with the false alarm rate of 2.23% with the training data of 100% benign samples. When training with 50% benign samples, the detection rate is 99.54% at the false alarm rate of 4.23%. When we use the smaller training data (25% benign samples), the result is also very positive: the detection rate is 99.62% and the false alarm rate is 6.82%.

By closely observing the influence of unknown normal data when using our algorithm, less normal training data only slightly increase the false alarm rate in breast cancer diagnosis. However, in a negative selection algorithm and its

Table 5: Detection rate and false alarm rate with different training data.

Training data	Detection rate (%)				False alarm rate (%)			
	Max	Min	Mean	SD	Max	Min	Mean	SD
100% benign samples	99.08	95.10	97.15	0.84	3.13	2.05	2.60	0.24
50% benign samples	99.87	96.53	98.40	0.67	7.48	3.69	5.44	0.83
25% benign sample	99.96	97.36	99.08	0.55	10.45	6.67	8.47	0.77

Table 6: Best results picked out from 100 repeated experiments for each training data of 100%, 50%, 25% benign samples.

Training data	Detection rate (%)	False alarm rate (%)
100% benign samples	98.74	2.23
50% benign samples	99.54	4.23
25% benign samples	99.62	6.82

variants, the false alarm rate usually increases dramatically when the normal training data decrease [16]. This phenomenon exactly demonstrates the strength of APC detector in the CSPRA. By comparison of canonical negative selection algorithm, the complexity of the proposed algorithm has not been increased. As shown in Table 4, the size of the APC detector is only 1. The proposed algorithm acts as adding only one robust detector (APC detector), which is used to detect conserved self pattern, to the total size of the detectors in a negative selection algorithm. The only observed side-effect for APC detector in the CSPRA is that the system produces the average false alarm rate of 2.23% even if we use 100% training normal samples whereas the false alarm for this case is zero in the negative selection algorithm. However, as already discussed, the false alarm rate is deceptive when all normal samples are used to train the system and are included in the testing data.

4.4. Influence of Various Parameters in the Algorithm. To further explore the effects of various control parameters in the algorithm on the performance of breast cancer prediction, experiments are done by changing the values for a certain parameter while the values for other parameters remain unchanged as listed in Table 4. The results reported in this section are obtained from the experiments with training data set of 50% benign samples randomly selected from the original collections.

Figures 2, 3, 4, and 5 show that both T detector threshold and APC detector threshold strongly affect the system performance. Both detection rate and false alarm rate increase when the T detector threshold increases. As shown in Figure 3, the detection rate rises rapidly before the value for T detector threshold reaches 0.1. However, when the T detector threshold is over 0.1, the false alarm rate starts a slow increment. APC detector threshold influences both the detection rate and false alarm rate along the opposite direction as seen in Figure 5; that is, both detection rate and false alarm rate decrease when the APC detector threshold increases. We find that the false alarm rate drops dramatically when the APC threshold gradually increases to 0.5 whereas

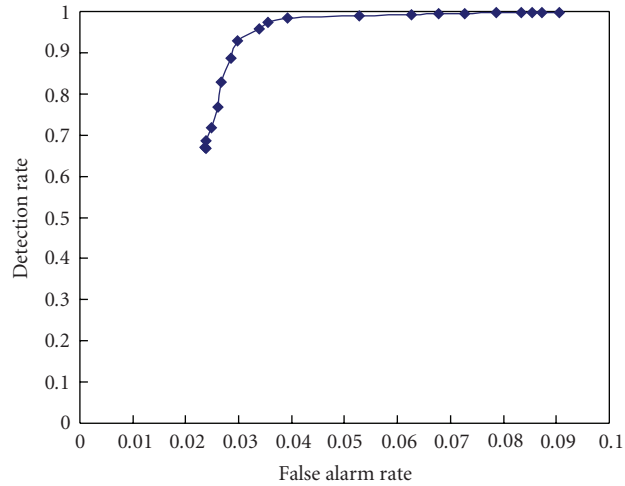

Figure 2: Performance (ROC) curve obtained by changing T detector threshold with the training data of 50% normal samples.

the decrement of the detection rate speeds up when the APC threshold is greater than 0.5.

The influence of the parameter of pattern weight on the system performance is also studied. However, the curves of both detection rate and false alarm rate are close to being a horizontal line in Figure 6, which indicates that the influence of the parameter of pattern weight is very slight when the breast cancer data are tested with the proposed algorithm. The influence of T detector size on the system performance is also limited based on the experimental results reported in Figure 7. The detection rate almost arrives at 100% when the size of the T detectors increases to 250. The larger T detector set (greater than 250) does not help the detection rate. Instead, it sometimes lowers the detection rate. The Figure 7 also illustrates that the size of the T detectors has little impact on the false alarm rate.

5. Discussion

5.1. The Sensitivity of Parameters. To make the algorithm reliable and usable on the application domain, the tradeoff between the number of the control parameters and the algorithm performance is worthy of being considered when we design the algorithm. Less control parameters make the algorithm simpler but less flexible. On the contrary, if more control parameters are provided, the algorithm becomes more flexible and thus makes it possible to maximize the algorithm performance by tuning various parameters specific for the application. Providing a good value for the

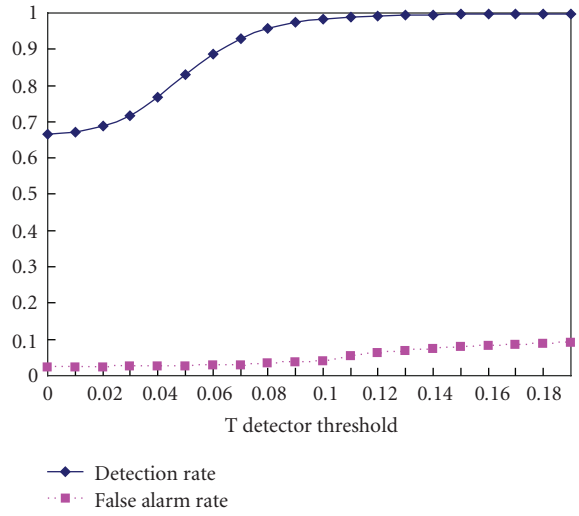

Figure 3: Influence of T detector threshold on both detection rate and false alarm rate with the training data of 50% normal samples.

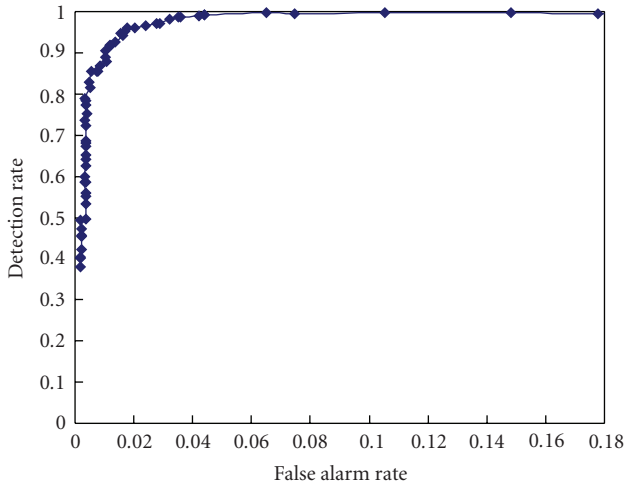

Figure 4: Performance (ROC) curve obtained by changing APC detector threshold with the training data of 50% normal samples.

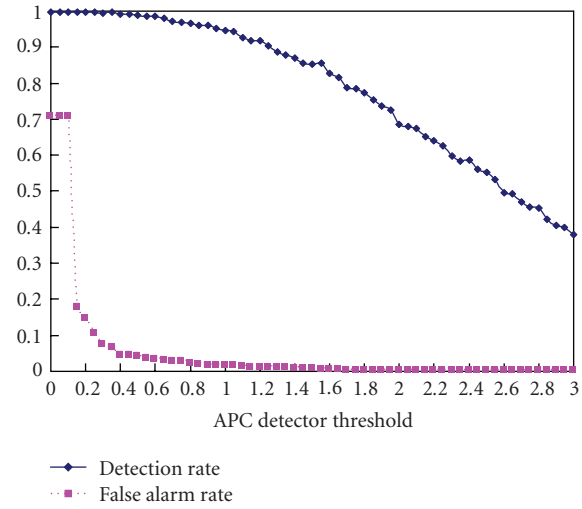

Figure 5: Influence of APC detector threshold on both detection rate and false alarm rate with the training data of 50% normal samples.

certain parameter, however, may not be very intuitive. More control parameters make the algorithm operation more complicated, so it is hard to make the algorithm work well if we have poor intuition on the parameters. The results reported above show that the performance of the modified CSPRA is very sensitive to the threshold values of both T detector and APC detector, so both thresholds are key parameters to balance between sensitivity and generation and are required to be tuned with the application data to obtain better performance. The result in Figure 7 indicates that a good value for the parameter of T detector size could easily be found because the detection rate quickly reaches 100% with lower false alarm rate when the size of the T detector increases. We find in our experiments that the influence of the parameters of both ambiguous coefficient and pattern weight on the algorithm performance is minor; so the values for both ambiguous coefficient and pattern weight could be set to 0.5 by default when we start tuning the proposed algorithm for a specific application. When the other parameters have been optimized, the ambiguous coefficient and pattern weight can be fine tuned to enlarge the algorithm performance. Better understanding the sensitivity of the parameters on the algorithm performance will direct us to first tune the parameters with higher sensitivity while ignoring the parameters with lower sensitivity by setting these parameters with the default values. Therefore, it not only reduces the workloads of tuning multiple parameters but also makes full use of the flexibility of multiple parameters.

5.2. Comparison with Previous Works. The typical methods applied to breast cancer diagnosis in previous works include linear programming [4, 5], neural network [6, 7], and ant colony-based system [8]. In this section we compare our approach against these previous works. To make the comparison more appropriate, our experiments also calculate the accurate rate of the breast cancer prediction using (6). The corresponding results are reported in Table 7

$$\text{AccurateRate} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{6}$$

The earliest work on breast cancer diagnosis [4] employed linear programming as the basic computational tool to determine two planes in an n-dimensional real space, as close together as possible, so that only the region between contains points from both sets of benign samples and malignant samples. Classification of the two sets is achieved by checking whether each point lies outside the region between the first pair of parallel planes. The validity of the method was tested with 369 samples. When 50% of the samples (185 samples) were used as a training set, 6.5% of the testing samples (the remaining 50% of the samples)

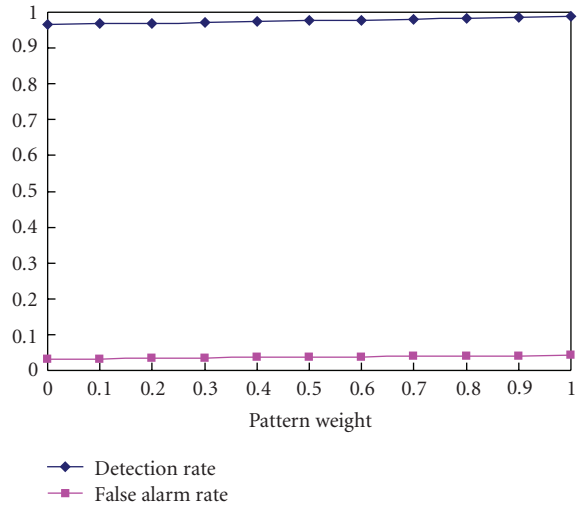

Figure 6: Influence of pattern weight on both detection rate and false alarm rate with the training data of 50% normal samples.

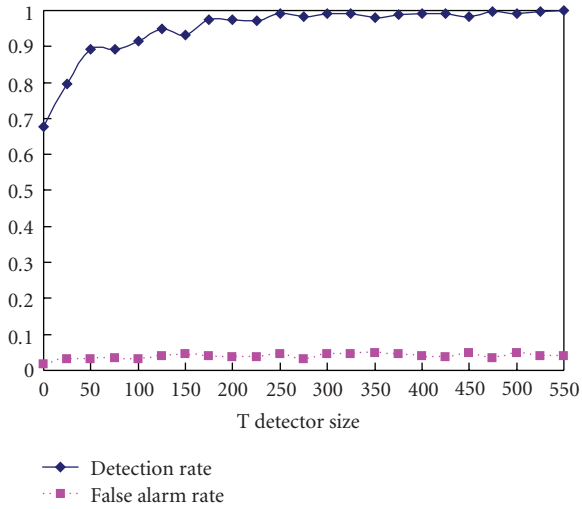

Figure 7: Influence of T detector size on both detection rate and false alarm rate with the training data of 50% normal samples.

were misclassified. This work was later extended in [5] by replacing 9 cytological characteristics for each sample with 30-dimensional feature vector that are generated by Xcyt system. The work in [5] used 569 vectors to train a linear programming-based diagnostic system, and the highest predicted accuracy, estimated with cross-validation, was 97.5% with three features (extreme area, extreme smoothness, and mean texture), a subset of the 30-dimensional feature vector. Compared to these previous works, we tested the larger dataset (683 samples) and achieved a better result (the highest accurate rate is 97.72% with training data of 50% of benign samples). Moreover, the requirement of ample malignant samples that are hardly collected in the clinic is mandatory in the training phase for these previous works [4, 5], whereas our approach only requires normal samples (the majority of the clinical cases) to train the system.

Table 7: Accurate rate with different training data.

Training data	Max (%)	Min (%)	Mean (%)	SD
100% benign samples	98.11	96.43	97.31	0.36
50% benign samples	97.72	95.21	96.55	0.49
25% benign samples	95.87	93.44	94.69	0.51

Abbass presented an evolutionary artificial neural network approach based on the pareto differential evolution algorithm augmented with local search for breast cancer diagnosis [6]. He used the same dataset as the one used in this paper. He chose the first 400 instances as the training set and the remaining 283 as the test set and the average test accuracy for his method is 98.1%. Although Abbass's method produces a higher accuracy rate than that obtained in our experiments, the larger training data used in his experiments could potentially overtrained the classifier. Because Abbass used the order of collections (*first* 400 instances) to define the training set, there could be ordering effects. Our experiment design, as described in Section 4.3, takes into account the problem with ordering effect, so our experiments are more reliable and reasonable.

In [7], the Wisconsin breast cancer database is used to train three different feedforward artificial neural network, including Cancer-Bin, Cancer-Norm, and Cancer-Cont networks. Cancer-Bin is trained with binary input patterns; Cancer-Norm is trained with normalized input patterns; Cancer-Cont is trained with the original data set. Three different rule extraction techniques (Binarized Rule Extraction, Partial Rule Extraction, and Full Rule Extraction), along with the rule ordering and integration mechanism are used to extract rules from these networks. Total 683 samples are evenly divided into a training set and a test set. The dimensionality of the breast-cancer input space is reduced from 9 to 6 inputs. When using the test set, the match rates for the trained networks of Cancer-Bin, Cancer-Norm, and Cancer-Cont are 96.20%, 95.91%, and 95.61%, respectively. By comparison with our method presented in this paper, the system used in [7] was apparently complicated. Based on the results in Table 7, our proposed method outperformed the hybrid symbolic-connectionist system in [7].

An ant colony-based system was also applied to breast cancer diagnosis [8]. The artificial Ant Colony System first specifies how ants construct or modify a solution for the problem domain, that is, the discovery of classification rules, in a probabilistic way and then update the pheromone trail considering the evaporation rate and the quality of the current solution. The predictive accuracy obtained with this method is 95.47%. In addition to the poor performance compared to our proposed method, the ant colony system was computationally expensive, especially when the search space (number of predicting attributes) is too large [8].

6. Conclusions

In this paper, a novel strategy on anomaly detection for the Conserved Self pattern Recognition Algorithm (CSPRA) is

proposed. To explore the new detection strategy, we first put forward the concept of ambiguous boundary which determines a ring-shaped region where unsure antigen objects reside. The novel detection strategy employs the domain knowledge and randomized method to resolve the conflicts in anomaly detection between two types of detectors (T detectors and APC detectors) during the detection phase of the algorithm. In particular, when the conflicts emerge, there are two cases that can be observed in the detection phase: (1) the new sample falls not only into the ambiguous region of the T detector but also into the ambiguous region of APC detector; (2) the new sample falls into neither the ambiguous region of the T detector nor the ambiguous region of APC detector. The randomized method is selected to resolve the conflicts if one of the above cases is observed. Domain knowledge is used to arrive at the final anomaly decision for a new testing sample for the observed cases other than the cases suitable for randomized method.

The improved CSPRA is applied to breast cancer diagnosis, and the promising results reported in this paper show that the potential usage of the CSPRA is an efficient and reliable technique to diagnose the breast cancer. The total 683 clinic samples collected by Dr. Wolberg are experimented with our proposed method after max-min normalization. The conserved self pattern is discovered by computing Pearsonian r between the values for each column of the attribute and the class labels. The experimental results are evaluated by considering not only the capability of the system to detect the anomalies (malignant samples) but also the extent of the system to misclassify the normal samples (benign samples). When only 25% of the normal benign samples (111 samples) are used to train the proposed model, the result is still very promising with the detection rate of 99.62% and the false alarm rate of 6.82% in the best case. Importantly, the best detection can remain unchanged because the set of T detectors and the indices of the randomized method can be on-line recorded and reused for future detection.

The influence of various control parameters on the performance of predicting breast cancer is also studied. The results indicate that the performance of the modified CSPRA is very sensitive to the threshold parameters of both T detector and APC detector. By comparison with the results in the literature from the previous works on breast cancer diagnosis, our method outperforms the other methods in addition to the advantage of its one-class classification. However, there is still a long way to apply our proposed method to the actual diagnosis in the clinic. There may be two reasons for immaturity of the proposed approach. One, the available breast cancer data are unbalanced between the two cases. The number of benign samples is almost double than that of malignant samples. Because of this limitation, we can not exclude that the promising performance of our method is due to overfitting towards benign cases. Two, the available data set is too small and the variation of the attributes is too low. For some attributes in raw breast cancer data, the same values are observed in most samples. Therefore, we still question that these observations in the source data might contribute to the good results generated by our method and the previous works. These problems could be lessoned when we can obtain a much larger number of samples balanced over two cases. Although relatively limited, the encouraging results of the CSPRA on the available breast cancer data still give weight to further study this technique with more data sets and real world applications and compare this technique with other AIS methods and machine learning algorithms, which constitutes the direction of our future work.

References

[1] E. Marshall, "Search for a killer: focus shifts from fat to hormones.," *Science*, vol. 259, no. 5095, pp. 618–621, 1993.

[2] S. W. Fletcher, W. Black, R. Harris, B. K. Rimer, and S. Shapiro, "Report of the international workshop on screening for breast cancer," *Journal of the National Cancer Institute*, vol. 85, no. 20, pp. 1644–1656, 1993.

[3] R. W. M. Giard and J. Hermans, "The value of aspiration cytologic examination of the breast: a statistical review of the medical literature," *Cancer*, vol. 69, no. 8, pp. 2104–2110, 1992.

[4] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 23, pp. 9193–9196, 1990.

[5] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," *SIAM News*, vol. 23, no. 5, pp. 1–18, 1990.

[6] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 25, no. 3, pp. 265–281, 2002.

[7] I. Taha and J. Ghosh, "Characterization of the Wisconsin breast cancer database using a hybrid symbolic-connectionist system," in *Proceedings of the Intelligent Engineering Systems through Artificial Neural Networks Conference (ANNIE '96)*, St Louis, Mo, USA, November 1996.

[8] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, "An ant colony based system for data mining: applications to medical data," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '01)*, pp. 791–798, San Francisco, Calif, USA, 2001.

[9] D. Dasgupta, "Advances in artificial immune systems," *IEEE Computational Intelligence Magazine*, vol. 1, no. 4, pp. 40–43, 2006.

[10] P. Matzinger, "The danger model: a renewed sense of self," *Science*, vol. 296, no. 5566, pp. 301–305, 2002.

[11] C. A. Janeway Jr., "Approaching the asymptote? Evolution and revolution in immunology," in *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 54, pp. 1–13, 1989.

[12] S. Yu and D. Dasgupta, "Conserved self pattern recognition algorithm," in *Proceedings of the 7th International Conference in Artificial Immune System (ICARIS '08)*, vol. 5132, pp. 279–290, Phuket, Thailand, August 2008.

[13] S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri, "Self-nonself discrimination in a computer," in *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, pp. 202–212, IEEE Computer Society Press, Los Alamitos, Calif, USA, 1994.

[14] UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems, University of California, Irvine, Calif, USA, http://archive.ics.uci.edu/ml.

[15] D. Moore, *Basic Practice of Statistics*, W. H. Freeman, San Francisco, Calif, USA, 2006.

[16] Z. Ji, D. Dasgupta, Z. Yang, and H. Teng, "Analysis of dental images using artificial immune systems," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '06)*, pp. 528–535, Vancouver, Canada, July 2006.