Neural Comput & Applic (2012) 21:1865–1870
DOI 10.1007/s00521-011-0547-0

ICIC2010

Breast cancer diagnosis based on a kernel orthogonal transform

Yong Xu · Qi Zhu · Jinghua Wang

Received: 4 November 2010 / Accepted: 31 January 2011 / Published online: 16 February 2011

Abstract Many pattern recognition and machine learning methods have been used in cancer diagnosis. In this study, we propose a kernel orthogonal transform method for breast cancer diagnosis. We test our method using the widely used Wisconsin breast cancer diagnosis (WBCD) dataset. The performance of the method is evaluated in terms of the classification accuracy, specificity, positive and negative predictive values, as well as receiver-operating characteristic curve (ROC). The experimental results show that our method classifies more accurately than all of the previous methods.

Keywords Breast cancer diagnosis · Pattern recognition · Machine learning · Kernel method

1 Introduction

Most types of cancers affect people at all ages, and the risk increases with age. It was reported that cancers caused about 13% of all human deaths in 2007 [1]. It is known that the cancer is generated from a group of cells that multiplies out of control. The group of rapidly dividing cells may form a lump or mass of extra tissue, i.e., tumors. Tumors can be classified as either cancerous (malignant) or non-cancerous (benign). Malignant tumors infract adjacent tissues and sometimes spread to other locations in the body via lymph or blood [2]. Breast cancer is a class of deadly disease and refers to a malignant tumor that has developed from cells in the breast. Breast cancer is the leading cause of death among women between 40 and 55 years of age and is the second overall cause of death among women [3]. Early diagnosis of this disease is very crucial for treatment of the patients.

We note that pattern recognition and machine learning methods have been used in disease diagnosis and there is promising performance. For the last decade, the kernel method has shown good performance in solving classification and prediction problems [4]. In this study, we propose a novel kernel-based orthogonal transform method for the diagnosis of breast cancer. As a novel kernel method, our method has the following advantages: first, an ordinary nonlinear method needs to explicitly implement the nonlinear mapping, whereas our method does not need to do so [4]. Second, our method can be applied to the cases where the dimensionality of the sample is larger than the sample number, whereas the linear orthogonal transform method described in Sect. 3.1 cannot be applied in these cases.

In this study, we also design a special classifier for breast cancer diagnosis. The designed classifier obtains a very high accuracy on the WBCD dataset with different training–testing partitions. Besides the classification accuracy, we also use the specificity, positive and negative predictive values, and receiver-operating characteristic curve (ROC) to evaluate the diagnostic performance of our method.

The rest of the paper is organized as follows. In Sect. 2, we describe the WBCD problem and provide a brief overview of related works on automatic diagnosis of breast cancer. In Sect. 3, we present the kernel-based orthogonal transform method for breast cancer diagnosis. In Sect. 4,

Y. Xu (✉) · Q. Zhu
Bio-Computing Research Center, Shenzhen Graduate School,
Harbin Institute of Technology, Shenzhen, China
e-mail: laterfall286@yahoo.com

J. Wang
Biometrics Research Centre,
The Hong Kong Polytechnic University,
Kowloon, Hong Kong

we describe the performance evaluation results of our method. The experimental results of our method on WBCD are presented in Sect. 5. Finally, we offer our conclusion in the last section.

2 Background

In 2002, Parkin et al. showed a study about global cancer statistics in which the data are collected from 20 large areas of the world [5]. According to this study, the high incidence of breast cancer in women makes it the most popular cancer both in developing and developed countries among all the cancer types, and the sum of the mortality cases of breast cancer also occupies the top position. Figure 1 shows the estimated numbers of new cancer cases (incidence) and deaths (mortality) in 2002 [5].

2.1 WBCD problem

The WBCD dataset is collected by Wolberg at the University of Wisconsin–Madison Hospitals. It consists of 683 samples taken from fine needle aspirates from human breast tissue, and each sample has nine features: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis. The measurements are assigned an integer value between 1 and 10. Each sample has its class label, which is either benign or malignant.

2.2 Related works

There have been a lot of literatures that apply pattern recognition or machine learning methods to WBCD. Quinlan used C4.5 decision tree method to solve WBCD and obtained a classification accuracy of 94.74% with 10-fold cross-validation [6]. Abonyi and Szeifert employed supervised fuzzy clustering (SFC) technique and obtained an accuracy of 95.57% [7]. Nauck and Kruse reached 95.06% with neuro-fuzzy techniques [8]. Hamilton et al. obtained an accuracy of 96% using the RIAC method [9]. Ster and Dobnikar applied linear discriminant analysis (LDA) to WBCD and obtained an accuracy of 96.80% [10], while Bennet reached an accuracy of 98.53% using least square SVM [11]. Moreover, Akay combined F test feature selection and support vector machines for breast cancer diagnosis and achieved a high accuracy of 99.51% with the 80–20% training-test partition [12].

3 Breast cancer diagnosis based on a kernel orthogonal transform

In this section, we present the linear orthogonal transform algorithm [13] and then describe our proposed breast cancer diagnosis based on a kernel orthogonal transform. The proposed kernel orthogonal transform is established on the basis of the linear orthogonal transform presented in Sect. 3.1.

3.1 A linear orthogonal transform method [13]

The linear orthogonal transform method can be described as follows: first, we consider the WBCD problem as a two-class classification problem. We refer to malignant and benign as the first and second classes, respectively. We assume that training samples $x_1, x_2, \ldots, x_{n_1}$ are from the first class and the others $x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2} (n_1 + n_2 = n)$ are from the second class. Each sample x_i consists of f components, i.e., $x_i = \left[x_i^1, x_i^2, \ldots, x_i^f\right]^T$. Let

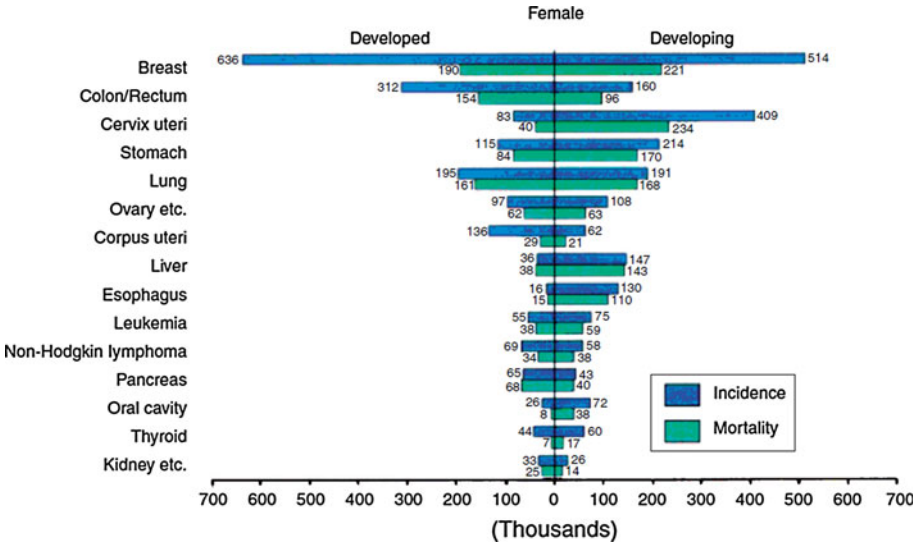

Fig. 1 Estimated numbers of new cancer cases (incidence) and deaths (mortality) in 2002 [5]

$$P = [x_1, x_2, \ldots, x_{n_1}]^T, \tag{1}$$

$$\text{and } Q = [x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2}]^T. \tag{2}$$

If $f > n_1$, then there must be c ($c = f - \text{rank(P)}$) orthogonal vectors $\alpha_1, \alpha_2, \ldots, \alpha_c$ that satisfy $P\alpha = \bar{0}$, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_c]$ ($\bar{0}$ is the $n_1 \times c$ matrix in which each element is zero). It is clear that an arbitrary group of c-dimensional orthogonal vectors $\beta_1, \beta_2, \ldots, \beta_d$ satisfies

$$P\alpha\beta = \bar{0}, \tag{3}$$

where $\beta = [\beta_1, \beta_2, \ldots, \beta_d]$. Superficially, we can obtain α by solving the linear system $P\alpha = \bar{0}$; however, there is no unique solution.

If we apply the transform $\alpha\beta$ to both of the first and second classes, then the samples of the first class and second class will be transformed into $P\alpha\beta$ and $Q\alpha\beta$, respectively. The linear orthogonal transform method aims to obtain the β that satisfies the following formula:

$$J = (Q\alpha\beta)^T(Q\alpha\beta), \text{s.t.} : P\alpha = \bar{0}. \tag{4}$$

If we can determine the α and β that satisfy (3) and (4), then the transform results of the samples from the first and second classes will have great difference. As a result, a classification procedure can achieve a good classification performance. In (4),

$$(Q\alpha\beta)^T(Q\alpha\beta) = \beta^T(Q\alpha)^T(Q\alpha)\beta \tag{5}$$

Equation (5) shows that $\beta_1, \beta_2, \ldots, \beta_d$ should be the eigenvectors corresponding to the first d largest eigenvalues of matrix $(Q\alpha)^T(Q\alpha)$. It is clear that the transform $\alpha\beta$ is linear and possesses the property of orthogonality (i.e., $P\alpha\beta = \bar{0}$, $\alpha_1, \alpha_2, \ldots, \alpha_c$ are orthogonal to each other, and $\beta_1, \beta_2, \ldots, \beta_d$ are orthogonal to each other), so it is referred to as a linear orthogonal transform.

3.2 Kernel orthogonal transform method

In this subsection, we present our kernel orthogonal transform method derived from the linear orthogonal transform shown in Sect. 3.1. Above all, let ϕ be a nonlinear mapping and $\phi(x_1), \phi(x_2), \ldots, \phi(x_{n_1+n_2})$ be the mapping results of training samples $x_1, x_2, \ldots, x_{n_1+n_2}$. The space spanned by $\phi(x_1), \phi(x_2), \ldots, \phi(x_{n_1+n_2})$ is referred to as feature space [14]. We define

$$\Phi_p = [\phi(x_1), \phi(x_2), \ldots, \phi(x_{n_1})]^T \tag{6}$$

$$\Phi_Q = [\phi(x_{n_1+1}), \phi(x_{n_1+2}), \ldots, \phi(x_{n_1+n_2})]^T \tag{7}$$

We assume that in feature space, there exist $r_1, r_2, \ldots, r_{c'}$ that satisfy

$$\Phi_P R = \bar{0},\ R = [r_1, r_2, \ldots, r_{c'}]. \tag{8}$$

According to the reproducing kernel theory, each r_i can be represented by a linear combination of the samples in feature space, i.e.,

$$r_i = \sum_j w_{ij}\phi(x_j). \tag{9}$$

On substituting (9) into (8), we obtain

$$K_P[w_1, w_2, \ldots, w_{c'}] = 0, \tag{10}$$

where $K_q = \begin{bmatrix} k(x_1,x_1) & \ldots & k(x_1,x_n) \\ k(x_2,x_1) & \ldots & k(x_2,x_n) \\ \ldots & \ldots & \ldots \\ k(x_{n_1},x_1) & \ldots & k(x_{n_1},x_n) \end{bmatrix}$ and $w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$. $k(x_1, x_1)$ denotes $(\phi(x_2))^T\ \phi(x_2)$. If $v_1, v_2, \ldots, v_{d'}$ is an arbitrary group of orthogonal vectors, then we have

$$K_P[w_1, w_2, \ldots, w_{c'}][v_1, v_2, \ldots, v_{d'}] = 0. \tag{11}$$

Let $W = [w_1, w_2, \ldots, w_{c'}]$ and $V = [v_1, v_2, \ldots, v_{d'}]$. If we apply the transform WV to the samples of the first and second classes in feature space, the samples from the two classes will be transformed into PWV and QWV, respectively.

If WV satisfies the following formulae

$$(K_Q WV)^T(K_Q WV), \text{ s.t.} : K_P W = \bar{0}, \tag{12}$$

where $K_Q = \begin{bmatrix} k(x_{n_1+1},x_1) & \ldots & k(x_{n_1+1},x_n) \\ k(x_{n_1+2},x_1) & \ldots & k(x_{n_1+2},x_n) \\ \ldots & \ldots & \ldots \\ k(x_{n_1+n_2},x_1) & \ldots & k(x_{n_1+n_2},x_n) \end{bmatrix}$, then WV is what we want. The W and V that satisfy (10) and (12) will enable in feature space the transform results of the two classes to be as much different as possible. As a result, it is very helpful for us to obtain a high classification accuracy. In (12),

$$(K_Q WV)^T(K_Q WV) = V^T(K_Q W)^T(K_Q W)V \tag{13}$$

It is easy to know optimal $v_1, v_2, \ldots, v_{d'}$ should be the eigenvectors corresponding to the first d' largest eigenvalues of matrix $(K_Q W)^T\ K_Q W$. The above method is our kernel orthogonal transform method. Our method can be directly applied to the cases where the dimensionality of the sample is larger than the sample number, whereas the linear orthogonal transform method described in Subsect. 3.1 cannot do so.

3.3 Classification procedure

The classification procedure designed for our method is as follows: first, the transform result y of testing sample x is

calculated using $y = [k(x, x_1), k(x, x_2), \ldots, k(x, x_n)]WV$. If y is an m-dimensional vector, i.e., $y = [y_1 \ldots y_m]$, then we use $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} |y_i|$ to represent the transform result of testing sample x. According to Sect. 3.2, the transform result of the sample from the first class will approximate the zero vector. Thus, if x is from the first class, then $\bar{y}$ will also approximate zero. Our devised classification rule is as follows:

$$\text{diagnosis decision}\,(x_i) = \begin{cases} \text{malignant}, & \text{if } \bar{y} < \theta \\ \text{benign}, & \text{otherwise} \end{cases} \tag{14}$$

θ is the threshold.

4 Performance evaluation

In this section, we present the performance evaluation indices including accuracy, sensitivity, specificity, positive predictive and negative predictive values. These indices will be used to evaluate our method. The classification accuracy of the dataset is measured using

$$\text{accuracy} = \frac{\sum_{i=1}^{N} \text{assess}(x_i)}{N}, \tag{15}$$

$$\text{assess}\,(x_i) = \begin{cases} 1, & \text{if diagnosis decision of } x_i \text{ is correct} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

N is the number of testing samples of the dataset. We will also show the accuracy of our performed k-fold cross-verification (CV) experiment.

The confusion matrix contains four classification performance indices: true positive, false positive, false negative, and true negative as shown in Table 1. These four indices are also usually used to evaluate the performance of the two-class classification problem.

Sensitivity, specificity, positive predictive value, and negative predictive value are defined as follows

$$\text{Sensitivity}\,(\%) = \frac{TP}{TP + FN} \times 100 \tag{17}$$

$$\text{Specificity}\,(\%) = \frac{TN}{FP + TN} \times 100 \tag{18}$$

Table 1 The four classification performance indices included in the confusion matrix

Actual class	Predicted class	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

$$\text{Positive predictive value}\,(\%) = \frac{TP}{TP + FP} \times 100 \tag{19}$$

$$\text{Negative predictive value}\,(\%) = \frac{TN}{TN + FN} \times 100 \tag{20}$$

The ROC curve is also usually used for experimental evaluation. The ROC curve simultaneously reflects the values of true positives and false positives. Each point on the ROC plot represents a sensitivity–specificity pair corresponding to a particular decision threshold.

5 Experiments and results

We conducted experiments on the WBCD dataset. We adopted the Gaussian kernel function $k(x_i, x_j) = \exp(-|x_i - x_j|^2/\sigma)$ in our method. The setting of σ is presented in Table 2. For the setting of θ in (14), we calculated the feature extraction results of all the training samples using the method presented in Subsect. 3.3. Then, we found the maximum value (denoted as a) from the feature extraction results of the training samples of the first class. Similarly, we also found the minimum value (denoted as b) of the feature extraction results of the training samples of the second class. Then we selected a value, shown in Table 2, from the interval $(\min(a, b), \max(a, b))$ for θ. Because, when θ is selected in the above interval, our method achieves higher classification accuracy on the training set than the other choices. We believe this interval for θ is also appropriate for getting good classification performance on the testing set. For the 50–50, 70–30, and 80–20% training–testing partitions, our method achieved accuracies of 98.53, 99.51, and 100%, respectively. Table 3 shows that our method obtained a higher accuracy than all of the previous methods mentioned in Sect. 2. We also performed

Table 2 Setting of σ and decision threshold θ

Case	σ	θ
50–50% training–testing partition	5	0.035
70–30% training–testing partition	5	0.035
80–20% training–testing partition	5	0.035
Fold 1 in 10-fold cv	300	0.003
Fold 2 in 10-fold cv	3	0.04
Fold 3 in 10-fold cv	55	0.5
Fold 4 in 10-fold cv	12	0.9
Fold 5 in 10-fold cv	8	0.0005
Fold 6 in 10-fold cv	35	0.2
Fold 7 in 10-fold cv	88	0.08
Fold 8 in 10-fold cv	88	0.15
Fold 9 in 10-fold cv	74	0.27
Fold 10 in 10-fold cv	12	0.8

Table 3 Classification accuracies obtained using our method and the previous methods mentioned in Sect. 2

Author	Method (dataset partition)	Classification accuracy (%)
Setiono	Neuro-rule 2a (train: 50%-test: 50%)	98.10
Our method	Kernel-based orthogonal transform (train: 50%-test: 50%)	98.53
Mehmet Fatih Akay	F-score plus SVM (train: 70%-test: 30%)	99.02
Our method	Kernel-based orthogonal transform (train: 70%-test: 30%)	99.51
Mehmet Fatih Akay	F-score plus SVM (train: 80%-test: 20%)	99.51
Our method	Kernel-based orthogonal transform (train: 80%-test: 20%)	100
Quinlan	C4.5 (10-fold CV)	94.74
Hamilton et al.	RIAC (10-fold CV)	96.00
Ster and Dobnikar	LDA (10-fold CV)	96.80
Nauck and Kruse	NEFCLASS (10-fold CV)	95.06
Abonyi and Szeifert	Supervised fuzzy clustering (10-fold CV)	95.57
Our method	Kernel-based orthogonal transform (10-CV)	98.53

Table 4 Confusion matrixes obtained using our method on different training–testing partitions

Training–testing partition	Actual class	Number of predicted "benign"	Number of predicted "malignant"
50–50% training–testing partition	Benign	260	2
	Malignant	3	81
70–30% training–testing partition	Benign	45	0
	Malignant	1	160
80–20% training–testing partition	Benign	35	0
	Malignant	0	102

The meaning of the training–testing partition is as follows: 80–20% training–testing partition means that the 80% of the samples were used as training samples and the remaining 20% of the samples were used as testing samples. Predicted "benign" and predicted "malignant" are referred to as the sample that is classified into benign and malignant, respectively

Table 5 Sensitivity, specificity, positive predictive value, and negative predictive value obtained using our method on different training–testing partitions

Measure	Accuracy with 50–50% training–testing partition (%)	Accuracy with 70–30% training–testing partition (%)	Accuracy with 80–20% training–testing partition (%)
Sensitivity (%)	99.23	100	100
Specificity (%)	96.30	93.75	100
Positive predictive value (%)	98.86	97.83	100
Negative predictive value (%)	97.59	100	100

10-fold cross-validation experiments on the WBCD dataset. Table 3 also shows that the classification accuracy of the 10-fold cross-validation experiment of our method is also higher than all of the previous methods. Tables 4 and 5 present the confusion matrixes, sensitivity, specificity, positive predictive value, and negative predictive value. Figures 2, 3, and 4 show the ROC curve of our method with different training–testing partitions. The closer the ROC plot is to the upper left corner, the higher the area under curve (AUC) and the classification accuracy is. With 80–20% training–testing partition, our method obtained a perfect classification result. In this case, the value of AUC reached the upper boundary 1, and the ROC plot passed through the upper left corner. In other words, both sensitivity and specificity are 100%.

6 Conclusion

In this work, we propose a kernel-based orthogonal transform method. We apply this method to the breast cancer diagnosis problem. Compared to previous breast cancer diagnosis researches, our diagnosis result is very promising. According to the results, our method is the most suitable automatic model for classifying WBCD data and it is very helpful for the oncologists to make the ultimate

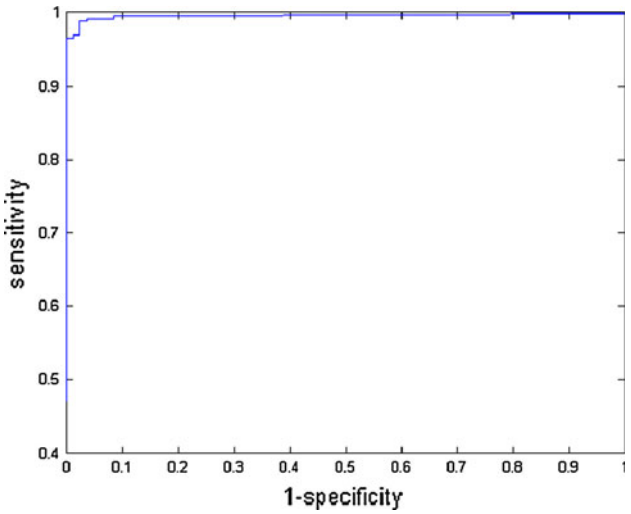

Fig. 2 ROC curve with 50–50% training–testing partition (AUC = 0.9952)

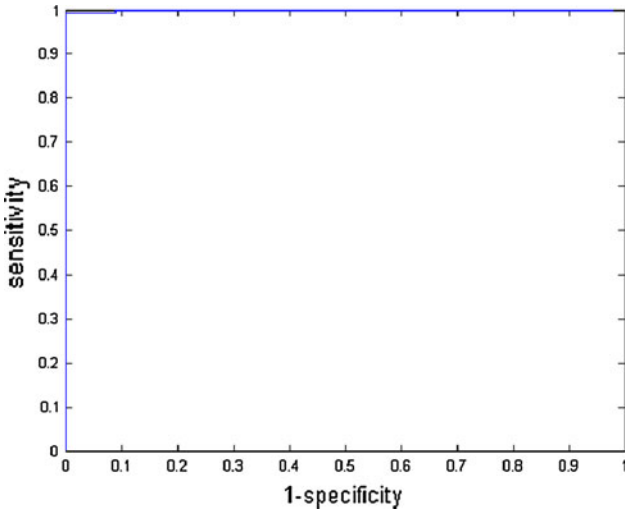

Fig. 3 ROC curve with 70–30% training–testing partition (AUC = 0.9996)

diagnosis decision. In the future, we will address the issue of automatically determining a proper threshold for the classification procedure of our method.

Acknowledgments This article is partly supported by Program for New Century Excellent Talents in University (Nos. NCET-08-0156 and NCET-08-0155), NSFC under grants No. 61071179, 60803090, 60902099, and 61001037, as well as the Fundamental Research Funds for the Central Universities (HIT.NSRIF. 2009130).

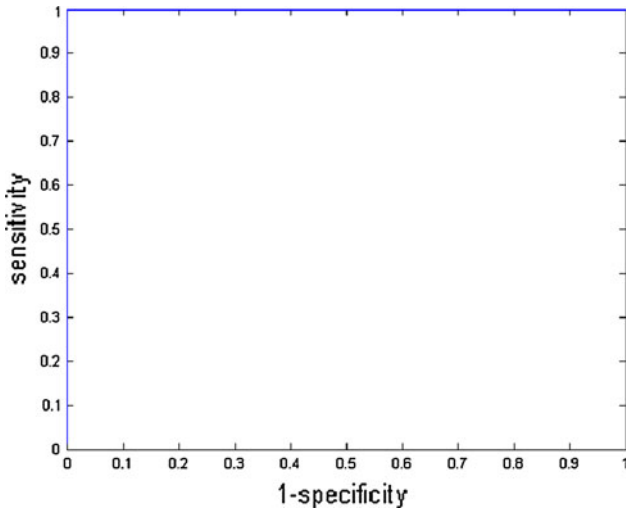

Fig. 4 ROC curve with 80–20% training–testing partition (AUC = 1.0000)

References

1. http://www.who.int/mediacentre/factsheets/fs297/en/
2. Kıyan T, Yıldırım T (2003) Breast cancer diagnosis using statistical neural networks, XII. In: TAINN symposium proceedings, E(8)
3. West D, Mangiameli P, Rampal R, West V (2005) Ensemble strategies for a medical diagnosis decision support system: a breast cancer diagnosis. Eur J Oper Res 162:532–551
4. Xu Y, Yang JY, Yang J (2004) A reformative kernel fisher discriminant analysis. Pattern Recogn 37:1299–1302
5. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics. CA Cancer J Clin 55:74–108
6. Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artific Intell Res 4:77–90
7. Abonyi J, Szeifert F (2003) Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognit Lett 24:2195–2207
8. Nauck D, Kruse R (1999) Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med 16:149–169
9. Hamilton HJ, Shan N, Cercone N (1996) RIAC: a rule induction algorithm based on approximate classification, Technical Report CS 96-06, University of Regina
10. Ster B, Dobnikar A (1996) Neural networks in medical diagnosis: comparison with other methods. In: Proceedings of the international conference on engineering applications of neural networks, pp 427–430
11. Bennet KP, Blue JA (1997) A support vector machine approach to decision trees. Math Report, No. 97-100, Rensselaer Polytechnic Institute
12. Akay MF (2009) Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 36:3240–3247
13. Wang J, You J, Li Q, Xu Y (2011) orthogonal discriminant vectors for face recognition across pose, Pattern Recognition (in press)
14. Schölkopf B et al (1999) Input space versus feature space in kernel-based methods. IEEE Trans Neural Netw 10(5):1000–1017