RESEARCH ARTICLE

Cancer Prediction Based on Radical Basis Function Neural Network with Particle Swarm Optimization

Xiao-Bo Yan[1], Wei-Qing Xiong[2], Liang Hu[1]*, Kuo Zhao[1]

Abstract

This paper addresses cancer prediction based on radial basis function neural network optimized by particle swarm optimization. Today, cancer hazard to people is increasing, and it is often difficult to cure cancer. The occurrence of cancer can be predicted by the method of the computer so that people can take timely and effective measures to prevent the occurrence of cancer. In this paper, the occurrence of cancer is predicted by the means of Radial Basis Function Neural Network Optimized by Particle Swarm Optimization. The neural network parameters to be optimized include the weight vector between network hidden layer and output layer, and the threshold of output layer neurons. The experimental data were obtained from the Wisconsin breast cancer database. A total of 12 experiments were done by setting 12 different sets of experimental result reliability. The findings show that the method can improve the accuracy, reliability and stability of cancer prediction greatly and effectively.

Keywords: Cancer prediction - radial basis function neural network - particle swarm optimization - breast cancer

Asian Pac J Cancer Prev, **15 (18)**, 7775-7780

Introduction

With the improvement of people's living standards and industrialization in recent years, cancer incidence and mortality are significantly higher. Research shows that "cancer prediction" can greatly reduce cancer incidence and mortality. Cancer prediction means predicting whether the body will suffer from cancer, or predicting the probability of the body suffering from cancer in the near future, with the body's own relevant indicators and the contact between body and the external environment, such as the location and number of diseased cells, whether to smoke, work environment radiation situation, whether to eat carcinogenic foods, etc. Now, there are a number of research institutions and medical institutions engaged in the study of cancer prediction. However, although researchers around the world have done a lot of exploration and research work, an effective method of cancer prediction still have not been developed. The current methods of cancer prediction are not so satisfactory that they cannot be put into reality using, since they are too complicated or have no practical value.

Based on the above situation, this paper proposes a new method of cancer prediction based on Radial Basis Function Neural Network (RBF Neural Network) Optimized by Particle Swarm Optimization (PSO) (Zhang et al., 2013). The neural network parameters to be optimized include the weight vector between network hidden layer and output layer, and the threshold of output layer neurons (Zhu et al., 2013). This is a method of computer-aided cancer prediction. The method will greatly reduce the complexity of early preconditioning large amounts of data, improve the accuracy of cancer prediction, and makes the process of cancer prediction more standard (Cheung et al., 2013). At the same time, this approach will liberate people from large amount of data analysis so that they can focus on research to improve the accuracy of cancer prediction, which will greatly improve the efficiency of cancer prediction (Reddy et al., 2012). Experimental results showed that the method of cancer prediction based on Radial Basis Function Neural Network Optimized by Particle Swarm Optimization (CRP Algorithm) can improve the accuracy, reliability and stability of cancer prediction greatly and effectively (Wang et al., 2012).

In the literature (Gao et al., 2005), the authors found that excessive calcium intake may increase the chance of prostate cancer (Che et al., 2014). Their findings of the authors in Literature (Hibi et al., 1997) showed that hypercalcemia and acute lymphoid leukemia are related. In Literature (Uhl et al., 1997), the authors found the serum calcium levels of breast cancer patients were lower than normal. Literature (Koksoy et al., 1997) studied the relationship between serum copper levels and breast cancer. In Literature (Goodman et al., 2004), the authors found that the copper imbalance of the body can cause a range of diseases, even cancer. The results of Literature (Gupta et al., 2005) showed that copper concentration in blood and gallbladder are closely related. In Literature (Mayland et al., 2004), the authors studied the serum zinc levels of the cancer patients, and they found the serum zinc levels of the cancer patients were generally lower

*[1]College of Computer Science and Technology, Jilin University, Changchun, Jilin, [2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China *For correspondence: 980513225@qq.com*

than the normal persons.

Literature (Karakitsos et al., 1998) distinguished between benign and malignant stomach tumor with LVQ, and the overall accuracy rate was very high. Polat et al., (2005), authors used FS-AIRS identification system to diagnose breast cancer, and received satisfactory results.

Literature (Rutkowska et al., 2004) took advantage of the multi-step classification to identify laryngeal. Literature (Pena-Reyes et al., 1999) used fuzzy genetic algorithm to diagnose breast cancer, and the accuracy rate was very high.

In Literature (Zhu et al., 2004), the authors used identify cancer with logistic regression and gene sequence, and they achieved a high classification accuracy. Literature (Liu et al., 2005) used the decision table to identify cancer, and the achieved a high sensitivity and specificity.

Literature (Wang et al., 2003) took advantage of Multilayer Perceptron and artificial neural network model to distinguish fibroids and oral cancer, and got better results. In Literature (Al-Ammar et al., 2001), the authors used supervised clustering algorithm to identify cancer, and some researchers diagnose cancer with support vector machines and gene sequence data, and achieved good results (Lee et al., 2003; Peng et al., 2003).

In Literature (Hadjiiski et al., 2004), the authors diagnosed breast cancer with computer and mammography images so that diagnostic accuracy increased from 79% to 84%. Literature (Sahiner et al., 2004) combined the breast ultrasound images and computer-aided diagnosis to achieve a higher diagnostic accuracy than radiologist (Hamilton et al., 1996; deGuzman et al., 2002).

The features, which could improve the diagnostic accuracy of computer-aided diagnosis, should be found. Literature (Hamilton et al., 1994) searched for the classification of the best features of breast cancer with genetic algorithms, and achieved good results. In Literature (Golobardes et al., 2002), the authors sorted the feature with SNR, so that the effective diagnosis of lymphoma could be found. Literature (Xin et al., 2005) sorted by high genetic characteristics with the maximum entropy model, and found the effective genetic characteristics for cancer diagnosis (Chim-Ong et al., 2014).

The rest of the paper is organized as follows: we first illustrate Radial Basis Function Neural Network, Particle Swarm Optimization and Radial Basis Function Neural Network and Particle Swarm Optimization (CRP Algorithm) in Section 2. We present our experimental results in Section 3. Finally, we discuss our study in Section 4.

Materials and Methods

RBF Neural Network

The nature of Neural Network (NN) is a mathematical model, which is built up on the basis of the structure and function of the biological neural network. The Neural Network consists of a large number of neurons, and these neurons may be layered, or not. Generally, the Neural Network, after setting the external conditions, can adjust internal structural parameters, weight parameters and threshold parameters automatically and intelligently. So, the Neural Network is adaptive.

We filter the data attributes with feature selection algorithms firstly. Then, we train the neural network with the selected data. When the training meets certain conditions, the neural network can be used for prediction. Its essence is a nonlinear mapping.

The NN does not have a standard and uniform definition. The definition, given by Hecht Nielsen, emphasizes the input information processing of NN. However, the definition, given by T.Koholen emphasizes the interaction of neural network response to the outside world.

RBF Neural Network (RBFNN) is a typical feed-forward neural network, which consists of three layers of neurons: input layer, hidden layer and output layer. The relationship between input layer and hidden layer is nonlinear, but the relation between hidden layer and output layer is linear.

The process of building a neural network is to determine the parameters of the network. The parameters of the RBFNN are input layer weight vector, the input layer neurons threshold, the output layer weight vector, the output layer neuron threshold, and the base function centers and width. These parameters can be obtained through the appropriate neural networks learning algorithms. A typical three-layer RBF neural network is shown in Figure 1

In the figure above, x_i (i= (1, 2, ⋯, n)) is the input vector; c_i (i= (1, 2, ⋯, m)) is the center of the hidden layer; w_i (i= (1, 2, ⋯, m)) is the weight vector of output layer; Y is the output of the network.

In the RBFNN, the relationship between input layer and hidden layer is nonlinear. The output of the i unit of the hidden layer is as follows,

$$h_i(x)=\exp[-\|x-c_i\|\div 2\sigma_i^2] \quad i=1, 2, \cdots, m \qquad (1)$$

In the above formula (1), c_i is the i neuron center of the hidden layer; σ_i is the width of the neurons in the hidden layer. x is the input layer vector.

In the RBFNN, the relationship between hidden layer and output layer is linear,

$$f(x)=\sum_{(i=1)}^{m} h_i(x)w_i \quad i=1, 2, \cdots, m \qquad (2)$$

In the above formula (2), $h_i(x)$ is the output of the i neuron in the hidden layer; w_i is the weight between the the i neuron in the hidden layer and the output layer. m is the number of neurons of the hidden layer; f (x) is the output result of the RBFNN.

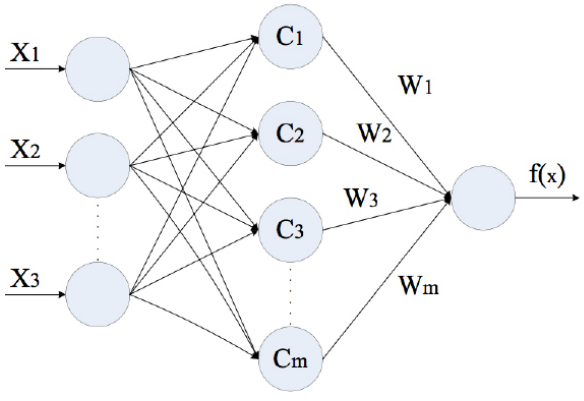

Figure 1. A Typical Three-Layer RBF Neural Network

Particle swarm optimization

Particle Swarm Optimization (PSO) is proposed in 1995 by Dr. Eberhart and Dr. Kennedy, who are inspired by the social behavior of birds and fish. It is a population-based stochastic optimization algorithm, which belongs to a cluster of intelligence.

In PSO, particles, which keep track of the current best particle, search the optimal solution, by flying in the solution space constantly. In the process of flight in the solution space, the particle maintains and records its own history optimal solution, which is the optimal solution found by the particle itself. Meanwhile, the PSO maintains and records the optimal solution of the particle group. The previous value is called the individual extreme value, which is recorded as PMAX, and it represents the nature of the individual particle. The second value is called the global extreme value, which is recorded as GMAX, and it represents the nature of the particle group.

We use the fitness function to evaluate whether the solution, found by the particles, is the optimal solution, and we calculate the fitness of the particle by this function. But, how do the particles update their position and speed, and find the optimal solution?

Assuming that particle swarm flies in n-dimensional solution space. Then, the j individual particle can be expressed as follows,

$$P_j = (P_{j1}, P_{j2}, \cdots, P_{jn})$$

The history optimal solution (PMAX) of the j particle, as well as the optimal solution of the particle group (GMAX), can be expressed as follows,

$$L_j = (L_{j1}, L_{j2}, \cdots, L_{jn})$$

The speed of the j particle can be expressed as follows,

$$V_j = (V_{j1}, V_{j2}, \cdots, V_{jn})$$

For each generation of particle swarm, their x-dimensional can be calculated according to the following equation,

$$V_{jx} = m*V_{j(x-1)} + m_1*rand_1*(PMAX - P_{j(x-1)}) + m_2*rand_2*(GMAX - P_{j(x-1)}) \quad (3)$$

$$P_{jx} = P_{j(x-1)} + V_{jx} \quad (4)$$

Equation (3) is the particle speed updating formula, and Equation (4) is the particle position updating formula. In the above formula, *m* is the inertia weight of the particle swarm; m_1 and m_1 are the acceleration constants of the particles; $rand_1$ and $rand_2$ are two random variables between 0 and 1.

The standard particle swarm algorithm can be divided into two categories: the global version of the standard PSO and local version of the standard PSO, updating the particle speed with the local optimal solution of the particle swarm (LMAX), which is different from the global version.

Theoretical basis of CRP algorithm

RBF Neural Network (RBFNN) has many advantages, such as less computation, training speed, adaptive, converging to the global optimum. It is mainly used in two aspects: function fitting, classification. In this paper, RBFNN will be used to predict cancer so that cancer incidence and mortality can be reduced.

In addition, this paper will use Particle Swarm Optimization (PSO) algorithm to optimize the parameters of RBFNN, including the weight vector between network hidden layer and output layer, and the threshold of output layer neurons. The PSO is a optimization algorithm based on swarm intelligence. It begins with a randomly generated solution vector, finding the optimal solution with swarm intelligence. Then, a self-defined fitness value, which is calculated by the fitness function, will be used to evaluate the quality of the solution vector. The PSO overcomes the shortcomings of Gradient Descent algorithm and Genetic algorithm, and it has many advantages, such as strong global optimization capability, easy to implement, and high accuracy solution. Therefore, the PSO is well suited for the optimization of the RBFNN parameters.

Design of CRP algorithm

The CRP Algorithm consists of two processes: optimization process and testing process. In optimization process, RBF Neural Network is optimized by PSO. In this paper, these RBFNN parameters, which are needed to be optimized, are the weight vector between network hidden layer and output layer, and the threshold of output layer neurons, and they are recorded with w and b. In testing process, the optimized RBF Neural Network is used to predict cancer.

The detailed steps of the CRP algorithm are as follows, (1) Parameter encoding. Firstly, the RBFNN parameters are encoded into digital strings, and expressed as individuals. The optimized parameters include the weight vector between network hidden layer and output layer, and the threshold of output layer neurons, and they are recorded with w and b. Assuming that RBFNN includes w and b, and each individual particle is represented by an m-dimensional vector, including w and b. (2) Particle initialization. Assuming that the number of the particle swarm is n, then n particles are generated randomly. In this step, the individual optimum and the global optimum are initialized, and they are recorded with Pbest and Gbest. (3) Update the individual optimum and the global optimum. Each particle is decoded into an m-dimensional vector, including w and b, thus constituting a RBF neural network. For each individual particle, we train the corresponding neural network with the samples. Then, we calculate the particle fitness value to update the individual optimum and the global optimum are initialized. (4) Update the position and speed of the particles. The position and speed of the particles are updated with the individual optimum and the global optimum. (5) Repeat step 3 and step 4, until the Gbest value meets the fitness requirements, or reaches a set number of iterations. (6) Predict cancer with the RBFNN optimized by the Gbest value from step 5.

The flow chart of the CRP algorithm is shown in Figure 2 as follows.

Results

Experimental data

This experiment uses the breast cancer data of Wisconsin database. The number of the experimental data is 255. Each data contains 11 attributes, including Sample Code, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single

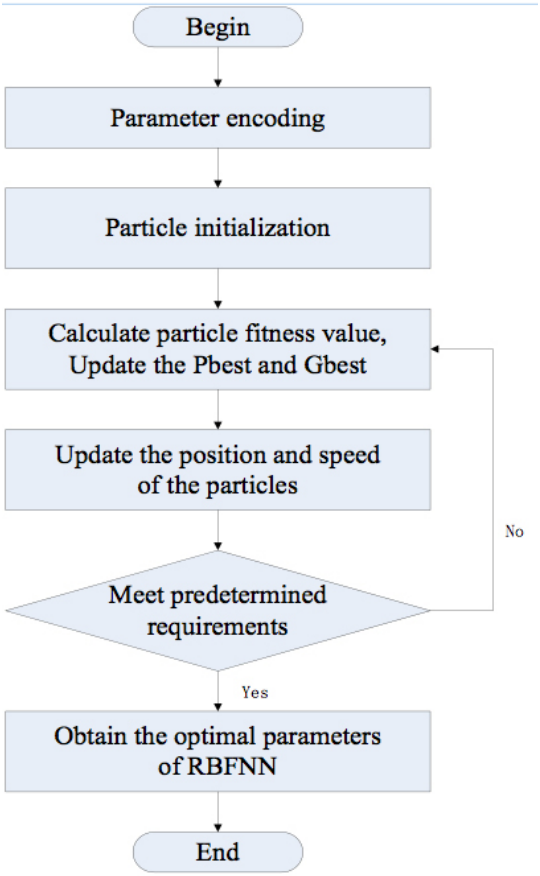

Figure 2. CRP Algorithm

Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class. In addition to the Sample Code, the other attributes values are between 1 and 10.

In this experiment, the experimental data is divided into two parts: training data set and test data set. The training data set consists of the previous 150 data; the test data set consists of the latter 105 data.

Data normalization

In order to process experimental data effectively, and improve the accuracy greatly, the experimental data will be normalized. All the experimental data will be normalized between 0 and 1. In this paper, the experimental data will be normalized with the following method,

$$Y = (2*X\text{-}Max\text{-}Min \div Max\text{-}Min) \quad (5)$$

In the above formula (5), Max is the maximum value of a property, and Min is the minimum value of a property. X is the original value, and Y is the value after normalization.

Evaluation indicators

This paper will evaluate the experimental results with reliability value and accuracy value.

Reliability value is the difference between the predicted values and expected values. The smaller the reliability value is, the more credible the predicted results are.

Accuracy value, in the case of meeting the reliability requirement, is the ratio between the number of the correctly predicted data and the number of the total test data.

Fitness function usually meets the following conditions: normative, single-valued, continuous, small amount of computation. This paper selects the mean squared error (MSE) as the fitness evaluation standards in the process of neural network optimization. MSE, as the fitness value of the j particle, is as follows,

$$MSE=1/N \sqrt{(\sum_{i=1}^{N} (x_i\text{-}y_i)^2)} \quad (6)$$

In the above formula (6), x_i represents the expected value, and y_i represents the predicted value of the neural network.

Experimental results

This experiment used 150 experimental data to train the RBF Neural Network, and optimized the RBFNN with PSO.

Then, 105 test data were used to predict cancer with the optimized RBFNN. The experimental results are shown in Figure 3, Figure 4, and Table I as follows,

The comparison of cancer prediction results before and after optimization is shown in Table II.

Analysis and evaluation

The analysis and evaluation of this experiment are as follows,

1) From Table 1 and Figure 4, it can be known that the reliability increases with the reducing of reliability value. And the higher the reliability is, the lower the prediction accuracy is.

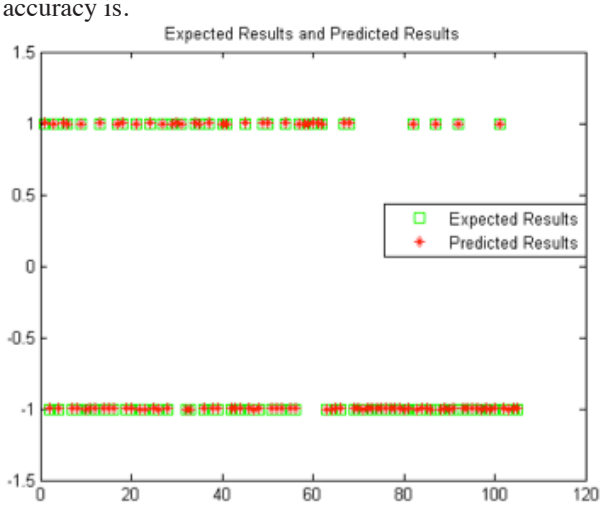

Figure 3. Experimental Results

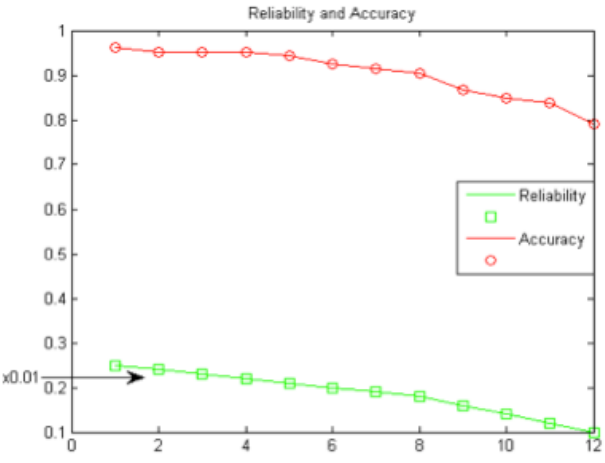

Figure 4. Reliability and Accuracy

Table 1. Comparison of Reliability and Accuracy

Experiments	Reliability	Accuracy
1	0.0025	96.190%
2	0.0024	95.238%
3	0.0023	95.238%
4	0.0022	95.238%
5	0.0021	94.286%
6	0.0020	92.381%
7	0.0019	91.429%
8	0.0018	90.476%
9	0.0016	86.667%
10	0.0014	84.762%
11	0.0012	83.810%
12	0.0010	79.048%

Table 2. Comparison of Cancer Prediction Before and After Optimization

Experiments	Reliability	Accuracy
1b	0.0023	94.286%
1a	0.0023	95.238%
2b	0.0022	93.333%
2a	0.0022	95.238%
3b	0.0021	91.429%
3a	0.0021	94.286%

1b, 2b, 3b: Before Optimization; 1a, 2a, 3a: After Optimization

2) From Table 2, we can know that cancer prediction accuracy increases significantly after optimization.

3) According to the experimental results, we can see that the CRP algorithm can improve the accuracy, reliability and stability of cancer prediction greatly and effectively.

Discussion

Based on the analysis of the current cancer situation, this paper studied the RBF Neural Network and Particle Swarm Optimization, and proposed Cancer Prediction based on RBF Neural Network Optimized by PSO (CRP Algorithm). In the CRP Algorithm, RBF Neural Network is optimized by PSO, so that it can predict cancer more accurately. The RBFNN parameters, which are needed to be optimized, are the weight vector between network hidden layer and output layer, and the threshold of output layer neurons. This paper did experiments with the breast cancer data of Wisconsin database, and the experimental results show that the method can improve the accuracy, reliability and stability of cancer prediction greatly and effectively.

Acknowledgements

This work was supported in part by the National High Technology Research and Development Program of China (Grant No. 2011AA010101), the National Natural Science Foundation of China (Grant No. 61103197 and 61073009), the Science and Technology Key Project of Jilin Province (Grant No. 2011ZDGG007), the Youth Foundation of Jilin Province of China (Grant No. 201101035), and the Fundamental Research Funds for the Central Universities of China (Grant NO.200903179).

References

Al-Ammar AS, Barnes RM (2001). Supervised cluster classification using the original n-dimensional space without transformation into lower dimension. *J Chemometrics*, **15**, 49-19.

Cheung MR (2013). Assessing the impact of socio-economic variables on breast cancer treatment outcome disparity. *Asian Pac J Cancer Prev*, **14**, 7133-4.

deGuzman MC, Prabhu N, Cramer N (2002). Automated breast cancer diagnosis based on fine needle aspiration. *Analytical Quantitative Cytology Histology*, **24**, 305-9.

Gao X, LaValley MP, Tucker KL (2005). Prospective studies of dairy product and calcium intakes and prostate cancer risk: a meta-analysis. *J Nat Cancer Institute*, **97**, 1768-10.

Golobardes E, Llora X, Salamo M, Marti J (2002). Computer aided diagnosis with case-based reasoning and genetic algorithms. *Knowledge-Based Systems*, **15**, 45-8.

Goodman VL, Brewer GL, Merajver SD (2004). Copper deficiency as an anti-cancer strategy. *Endocrine-Related Cancer*, **11**, 255-9.

Gupta SK, Singh SP, Shukla VK (2005). Copper, zinc, and Cu/Zn ratio in carcinoma of the gallbladder. *J Surg Oncology*, **91**, 204-5.

Hadjiiski LM, Chan HP, Sahiner B, et al (2004). Improvement of radiologists characterization of malignant and benign breast masses in serial mammograms by computer-aided diagnosis: an ROC study. *Radiology*, **233**, 255-11.

Hamilton PW, Anderson N, Bartels PH, Thompson D (1994). Expert system support using Bayesian belief networks in the diagnosis of fine needle aspiration biopsy specimens of the breast. *J Clin Pathol*, **47**, 329-8.

Hamilton PW, Anderson NH, Diamond J, et al (1996). An interactive decision support system for breast fine needle aspiration cytology. *Anal Quantitative Cytology Histology*, **18**, 185-6.

Hibi S, Funaki H, Ochiai-Kanai R, et al (1997). Hypercalcemia in children presenting with acute lymphoblastic leukemia. *Int J Hematology*, **66**, 353-5.

Karakitsos P, Ioakim-Liossi A, Pouliakis A, et al (1998). A comparative study of three variations of the learning vector quantizer in the discrimination of benign from malignant gastric cells. *Cytopathology*, **9**, 114-2.

Koksoy C, Kavas GO, Akcil E, et al (1997). Trace elements and superoxide dismutase in benign and malignant breast diseases. *Breast Cancer Res Treat*, **45**, 1-6.

Lee Y, Lee CK (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132-8.

Liu J, Li M (2005). Finding cancer biomarkers from mass spectrometry data by decision lists. *J Computational Biology*, **12**, 971-9.

Mayland C, Allen KR, Degg TJ, Bennett M (2004). Micronutrient concentrations in patients with malignant disease: effect of the inflammatory response. *Annals Clin Biochemistry*, **41**, 138-4.

Pena-Reyes CA, Sipper M (1999). A fuzzy-genetic approach to breast cancer diagnosis. *Artif Intell Med,* **17**, 131-25.

Peng SH, Xu QH, Ling XB, et al (2003). Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters*, **555**, 358-5.

Polat K, Sahan S, Kodaz H, et al (2005). A new classification method for breast cancer diagnosis: feature selection artificial immune recognition system (FS-AIRS). *Lecture Notes Computer Sci*, **3611**, 830-9.

Rutkowska D, Klimala JK (2004). A multi-stage classification

method in application to diagnosis of larynx cancer. *LNAI*, **3070**, 1037-6.
Sahiner B, Chan HP, Roubidoux MA, et al (2004). Computerized characterization of breast masses on 3-D ultrasound volumes. *Med Phys*, **31**, 744-11.
Uhl L, Maillet S, King S, Kruskall MS (1997). Unexpected citrate toxicity and severe hypocalcemia during apheresis. *Transfusion*, **37**, 1063-3.
Wang CY, Tsai T, Chen HM, Chen CT, Chiang CP (2003). PLS-ANN based classification model for oral submucous fibrosis and oral carcinogenesis. *Lasers Surg Med*, **32**, 318-9.
Xin J, Bie RF (2005). Classification Analysis of SAGE Data Using Maximum Entropy Model. *LNAI*, **3614**, 1037-4.
Zhang SC, Jin W, Liu H, et al (2013). RPSA gene mutants associated with risk of colorectal cancer among the Chinese population. *Asian Pac J Cancer Prev*, **14**, 7127-5.
Zhu J, Hastie T (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**, 427-17.