Diagnosis of Breast Tumours and Evaluation of Prognostic Risk by Using Machine Learning Approaches

Qianfei Yuan, Congzhong Cai, Hanguang Xiao,
Xinghua Liu, and Yufeng Wen

Department of Applied Physics, Chongqing University, Chongqing 400044, China
caiczh@gmail.com

Abstract. Machine learning approaches were employed for malignant breast tumour diagnosis and evaluation of the prognostic risk of recrudescence and metastasis by using age and ten cellular attributes of Fine Needle Aspirate of Breast (FNAB) and gene microarrays data of the breast cancer patient respectively. Feature ranking method was introduced to explore the salient elements for cancer identification and simultaneous improve the classification accuracy. In this paper, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN) and Probabilistic Neural Network (PNN) combined with Signal-to-Noise Ratio (SNR) for feature ranking and filtering were applied to distinguish between the benign and malignant tumours of breast and evaluate the prognostic risk of recrudescence and metastasis. The results reveal that feature ranking method SNR can effectively pick out the informative and important features, which had significance for clinical assistant diagnosis and is useful for improving the performance of evaluation. The best overall accuracy for breast cancer diagnosis and evaluating the prognostic risk of recrudescence and metastasis achieved 96.24% and 88.81% respectively, by using SVM-Sigmoid and SVM-RBF combined with SNR under 5-fold cross validation. This study suggests that SVM may be further developed to be a practical methodology for clinical assistant differentiating between benign and malignant tumours and possible to help the inexperienced physicians avoid misdiagnosis. It also has benefit to the cured patients who are predicted as recrudescence and metastasis pay more attention to their diseases, and then reduce the mortality rate of breast cancer.

1 Introduction

Cancer begins with uncontrolled division of one cell, which results in a visible mass named tumour. Tumour can be benign or malignant. Malignant tumour grows rapidly and invades its surrounding tissues causing their damage. Breast cancer is a malignant tissue beginning to grow in the breast. The symptoms of breast cancer include breast mass, change in shape and dimension of breast, differences in the color of breast skin, breast aches and gene changes etc. At present, breast cancer is one of the most prevalent cancers, ranking second only

D.-S. Huang, L. Heutte, and M. Loog (Eds.): ICIC 2007, CCIS 2, pp. 1250–1260, 2007.

to lung cancer and is the most prevalent form of cancer among worldwide women [1]. According to the report of the World Health Organization (WHO), about 1,000,000 women would be newly diagnosed with breast cancer and over 500,000 women died from breast cancer every year. It is estimated that the incidence of this disease will increase getting along with the damaging of environment in the future.

The traditional approaches to diagnose breast cancer include physical examination, X-ray mammography, Ultrasonic diagnosis, CT scan, Fine Needle Aspirate of Breast (FNAB) etc. Some of them are complex and expensive; especially the X-ray mammography method may be injurious to the patients exposed to X-ray. At the same time, traditional ultimate diagnosis method is cytological examination which requires cytopathologists to make their diagnoses and the diagnostic accuracies appear not satisfying. Up to now, there is no effective tools to evaluate the prognostic risk of recrudescence and metastasis. In order to improve the accuracy of diagnosis and evaluate the prognostic risk, a number of computer-aided approaches have been proposed for breast cancer diagnosis and prognostic risk evaluation. For instance, Butler et al. applied Bayes classifiers combined with feature selection to diagnose breast cancer, which reached 90% accuracy by using X-Ray scatter images [2]. Palmer et al. employed SVM with Principal Component Analysis (PCA) method, and obtained 70% sensitivity and 92% specificity by utilizing multiexcitation fluorescence and diffuse reflectance spectroscopy [3]. Song et al. adopted artificial neural network by using ultrasound image of breast to predict breast cancer and got 95% sensitivity and 76.5% specificity [4]. Cosar et al. obtained satisfying result by using Breast Imaging Reporting and Data System (the fourth edition) (BI-RADS) to classify breast cancer [5]. Li et al. used multivariate logistic regression to recognize breast cancer, the sensitivity and specificity achieved 93% and 91% respectively by using serum biomarkers [6]. Quinlan obtained 94.74% accuracy by using 10-fold cross validation with C4.5 decision tree [7]. Abonyi and Szeifert applied Supervised Fuzzy Clustering (SFC) technique, and obtained 95.57% accuracy [8]. Setiono got 98.1% overall accuracy by using neuro-rule method [9]. Huang et al. employed SVM and multilayer perception neural networks (MLPs) to differentiate between the benign and malignant of breast tumour by using ultrasonic images of solid breast nodules, and found that the sensitivity is 100%, specificity is 92.05%, and the overall accuracy is 95% [10]. Delen et al. adopted the machine learning approaches to predict breast cancer survivability and obtained 93.6% accuracy [11]. van't Veer et al. applied gene expression profiling to predict clinical outcome of breast cancer [12]. Choudhary et al. employed genetic test bed to select feature and predict breast cancer [13]. All above studies demonstrate that CAD is capable of improving the radiologist's performance [14].

The SVM, K-NN and PNN have been extensively employed in many fields [4-7, 9-11, 15-25] and shown excellent generalization, learning and classification ability for the binary and multi-class classification of real problems.

In this study, several machine learning approaches, such as SVM, K-NN and PNN, combined with Signal-to-Noise Ratio (SNR), are employed to discriminate

between benign tumour group and breast cancer patients by using fine needle aspirate of breast lesions dataset, evaluate the prognostic risk of recrudescence and metastasis by using gene microarrays dataset.

2 Material and Methods

2.1 Dataset

The dataset of fine needle aspirate of breast lesions (dataset I) was posted on the website by Dr. Cross (http://www.phil.gu.se). This dataset was collected by Department of Pathology, Royal Hallamshire hospital, Sheffeld, during 1992-1993. It contains 692 specimens of fine needle aspirates of breast lumps (FNAB), the number of positive samples (malignancy) is 235 and negative samples (benign) is 457. All of the specimens were confirmed by open biopsy. Each sample includes eleven features which consist of patient age and ten attributes of cell (Table 1). All observations of the cellular attributes were made by a consultant pathologist with 10 years experience of reporting FNAB. The ten features were assigned a value of -1 for their absent or +1 for their present.

The dataset of gene microarrays (dataset II) comes from Refs [12] and [13]. It contains 295 microarrays, 115 belong to the 'good-prognosis' class (labeled 1) and the remaining 180 belong to the 'poor-prognosis' class (labeled -1). Each sample contains 70-gene prognosis profile, a prognosis signature based on gene-expression is proposed in Ref [12] that correlates well with patient survival data and other existing clinical measures.

2.2 Signal-to-Noise Ratio

Feature saliency measures provide a way to measure the relative usefulness of features and a means to rank the features. Signal-to-Noise Ratio is a value that uses the signal to compare with other background noise. Usually, it is simple and capable of fast ranking and filtering features for classifiers [26]. The definition of SNR for two classes is formulated as:

$$SNR_i = \frac{|\mu_P(i) - \mu_N(i)|}{\sigma_P(i) + \sigma_N(i)} \tag{1}$$

where SNR_i is the value of saliency metric for the i-th feature; $\mu_N(i)$ and $\mu_P(i)$ are the averages of the i-th feature in class N and class P respectively; $\sigma_N(i)$ and $\sigma_P(i)$ are the standard deviations of the i-th feature in class N and class P respectively.

2.3 Classifiers

Support Vector Machine. SVM was proposed by Vapnik and co-workers [27] based on the statistical learning theory and structural risk minimization, which was extensively used as an effective algorithm to deal with classification and

Table 1. The definition of the ten attributes of cell and age used as input features for dataset I

Feature No.	Observed Feature	Definition
1	Age	–
2	Cellular dyshesion	True if the majority of epithelial cells are dyshesive, false if the majority of epithelial cells are in cohesive group.
3	Intracytoplasmic lumina	True if intracytoplasmic lumina are present in some epithelial cells, false if absent.
4	'3-dimensionality' of epithelial cells clusters	True if some clusters of epithelial cells are not flat (more than teo nuclei thick) and this is not due to artefactual folding, false if all clusters of epithelial cells are flat.
5	Bipolar 'naked' nuclei	True if bipolar 'naked' nuclei are present, false if absent.
6	Foamy macrophages	True if foamy macrophages are present, false if absent.
7	Nucleoli	True if more than three easily-visible nucleoli are present in some epithelial cells, false if three or fewer easily-visible nucleoli in all epithelial cells.
8	Nuclear pleiomorphism	True if some epithelial cells have nuclear diameters twice that of other epithelial cell nuclei, false if no epithelial cell nuclei have diameters twice that of other epithelial cell nuclei.
9	Nuclear size	True if some epithelial cell nuclei have diameters twice that of red blood cell diameters, false if all epithelial cell nuclei have diameters less than twice that of red blood cell diameters.
10	Necrotic epithelial cells	True if necrotic epithelial cells are present, false if absent.
11	Apocrine change	True if the majority of epithelial cell nuclei show apocrine change, false if apocrine change is not present in the majority of epithelial cells.

regression problems. The basic idea of applying SVM for solving classification problems can be stated briefly in three steps. In the first step, SVM transforms the original features in the input space to the feature vectors in a higher dimension feature space through a kernel function (e.g. linear kernel, polynomial kernel, RBF kernel and sigmoid kernel, etc.). And then, it constructs the optimal separating hyperplane (OSH) with maximum distance between the closest points of positives and negatives within the training set. For the last step, the class of a query sample for test is determined by the sign of the projection result of the test vector to the normal direction on OSH [28].

K-Nearest Neighbor. K-NN classifier is one of the simplest and oldest methods for performing general, nonparametric classification. First, complete the computation of distances between the test sample and all samples in the training

set. And then, the class of test sample is assigned according to a simple majority vote over the labels of its K nearest neighbors.

Probabilistic Neural Network. PNN was proposed by Specht in 1988 [29]. It is designed to improve the performance of conventional neural networks in which long computation times are required. PNN replaces the sigmoid activation function often used in neural networks with a statistically derived exponential function. The PNN is an extension of what is probably the simplest possible classifier - find the training sample closest to the test sample and assign it the same class. The PNN weight w_i is the contribution of i-th training sample which is according to its distance to the test sample. The weight for a given training sample is proportional to the Gaussian kernel:

$$w_i = \alpha e^{-(d/\sigma)^2} \tag{2}$$

where α is the coefficient of the proportion; d is the Euclidean distance and σ is a constant called the kernel width (other kernels and distance metrics are possible but less common). The weights of each class are summed respectively, the class of the test sample is assigned based on that of the largest summation value.

3 Results and Discussion

3.1 Feature Ranking

The eleven features numbered as in Table 1 were ranked by using SNR index. The ranking result is shown in Table 2. It can be seen that No.9, No.7 and No.8 features are ranked as the top three by using SNR criterion, corresponding to 'Necrotic epithelial cells', 'Nuclear pleiomorphism' and 'Nuclear size', respectively. This is consistent with the result of other researcher that 'Nuclear size' has great significance to distinguish the benign from malignant breast tumour [30]. The ranking results indicated that the above three features contain more informative and important information than other features for distinguishing between benign tumour and breast cancer. It supplies a valuable clue for cytopathologist to pay more attention to these factors in their clinical breast tumour diagnoses.

Table 2. Feature ranking result of dataset I by using SNR criterion

Feature ranking method	Feature rank										
	1	2	3	4	5	6	7	8	9	10	11
SNR	9	7	8	4	1	3	2	10	11	5	6

3.2 Classification Results and Discussions

Three evaluation terms, sensitivity (Sen), specificity (Spe) and overall accuracy (Q) [11] are introduced to estimate the performance of classifiers. They are defined as follow:

$$Sen = TP/(TP + FN) \tag{3}$$

$$Spe = TN/(TN + FP) \tag{4}$$

$$Q = (TP + TN)/(TP + FN + TN + FP) \tag{5}$$

Where TP and TN are the number of samples which are right identified as positives or negatives by the classifier in the test set, respectively; FN and FP are the numbers of samples corresponding to those cases as they are mistakenly tested as benign or malignant, respectively.

Considering imbalanced positive and negative samples in the data set, another quantity suitable for evaluating the classification accuracy of imbalanced positive and negative samples is the Matthews Correlation Coefficient MCC, which is given by:

$$MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{6}$$

Obviously, the scope of the MCC is within the range of [-1, 1]. The larger the MCC value, the better the classifier performance.

In this study, machine learning approaches including SVM, K-NN and PNN were applied to diagnose breast cancer via dataset I and evaluate the prognostic risk of recrudescence and metastasis via dataset II, by using 5-fold cross validation. For SVM, three types of kernel functions, such as polynomial kernel, RBF kernel and sigmoid kernel, were implemented in the classification.

The overall accuracies for 5-fold cross validation by using dataset I are shown in Table 3. The results illustrated, the overall accuracies of SVM-Polynomial, SVM-RBF, SVM-Sigmoid, K-NN and PNN achieved 96.09%, 95.80%, 96.24%, 95.37% and 95.08% respectively. All of the overall accuracies of SVM with three kernel functions are superior to those of K-NN and PNN. Surpassing all of other remaining classifiers, SVM-RBF obtained the highest accuracy (96.24%). It also demonstrates that classifier and kernel function optimization are necessary to obtain the best accuracy. Values of MCCs were roughly similar and ranged from 0.8898 to 0.9127.

Table 3. 5-fold cross validation results of dataset I for 3 classifiers (SVM, K-NN, PNN) by using the original features

classifiers	*Sen* (%)	*Spe* (%)	*Q*(%)	*MCC*
SVM-Poly	93.19	97.59	96.09	0.9127
SVM-RBF	93.19	97.15	95.80	0.9063
SVM-Sig	94.04	97.37	96.24	0.9161
K-NN	91.06	97.59	95.37	0.8963
PNN	91.06	97.15	95.08	0.8898

In order to further improve the prediction accuracy and save the computing time, SNR is introduced to find out the predominant features and filter the irrelevant features for classification. According to the feature ranking result in Table 2, we implemented sequential backward feature selection algorithm, and

the accuracies using different classifiers with 5-fold cross validation are shown in Table 4. It can be seen from Table 3 and Table 4, after ranking and filtering features, only the recognition rate of K-NN was improved from 95.37% to 96.09% when the No.6 feature was eliminated. The performances of SVM-Polynormial, SVM-RBF and SVM-Sigmoid were not improved while as the accuracy for PNN kept unchangeable (95.08%) by using either eleven original features or the selected top seven features. Values of MCCs for these classifiers were also large and ranged from 0.8833 to 0.9127 respectively.

Table 4. 5-fold cross validation results of dataset I by using the classifiers (KNN, SVM, PNN) after feature filtering based on the ranking result of SNR

		Test set			
Classifier	Input features No.	$Sen(\%)$	$Spe(\%)$	$Q(\%)$	MCC
SVM-Poly	9, 7, 8, 4, 1, 3, 2, 10, 11, 5	93.19	97.37	95.95	0.9095
	9, 7, 8, 4, 1, 3, 2, 10, 11	92.76	97.37	95.80	0.9062
	9, 7, 8, 4, 1, 3, 2, 10	91.91	97.37	95.52	0.8996
	9, 7, 8, 4, 1, 3, 2	92.34	96.71	95.23	0.8934
SVM-RBF	9, 7, 8, 4, 1, 3, 2, 10, 11, 5	91.48	97.15	95.23	0.8931
	9, 7, 8, 4, 1, 3, 2, 10, 11	91.06	97.81	95.52	0.8996
	9, 7, 8, 4, 1, 3, 2, 10	91.91	96.72	95.08	0.8901
	9, 7, 8, 4, 1, 3, 2	90.63	97.37	95.08	0.8898
SVM-Sig	9, 7, 8, 4, 1, 3, 2, 10, 11, 5	94.04	96.72	95.80	0.9067
	9, 7, 8, 4, 1, 3, 2, 10, 11	92.76	97.37	95.80	0.9062
	9, 7, 8, 4, 1, 3, 2, 10	92.34	96.72	95.23	0.8934
	9, 7, 8, 4, 1, 3, 2	92.34	96.71	95.23	0.8934
K-NN	9, 7, 8, 4, 1, 3, 2, 10, 11, 5	91.06	98.68	96.09	0.9127
	9, 7, 8, 4, 1, 3, 2, 10, 11	91.49	97.81	95.66	0.9028
	9, 7, 8, 4, 1, 3, 2, 10	88.93	99.12	95.66	0.9034
	9, 7, 8, 4, 1, 3, 2	88.93	98.90	95.52	0.9
PNN	9, 7, 8, 4, 1, 3, 2, 10, 11, 5	90.63	97.37	95.08	0.8897
	9, 7, 8, 4, 1, 3, 2, 10, 11	90.63	96.93	94.79	0.8833
	9, 7, 8, 4, 1, 3, 2, 10	91.49	96.71	94.94	0.8867
	9, 7, 8, 4, 1, 3, 2	91.91	96.71	95.08	0.8901

As a whole, the best overall accuracy (96.24%) was obtained by using SVM-Sigmoid with the original eleven features. Although the feature ranking and filtering did not improve the best overall accuracy, it explored the informative and important features to distinguish the benign from malignant breast tumour.

The dataset II was normalized ([-1, 1]) ahead. The overall accuracies for 5-fold cross validation are shown in Table 5. It can be seen from Table 5, the overall accuracies of SVM-RBF achieved 87.80%, which is superior to SVM-Polynomial (86.44%), SVM-Sigmoid (86.44%), K-NN (78.98%), PNN (78.98%). Value of MCC for SVM-RBF achieved 0.7427 and is superior to those of other classifiers. It is illustrated, SVM-RBF has excellent performance to distinguish the cured patients who are easy recrudescence and metastasis.

Table 5. 5-fold cross validation results of dataset II for 3 classifiers (SVM, K-NN, PNN) by using the original features

classifiers	*Sen* (%)	*Spe* (%)	*Q*(%)	*MCC*
SVM-Poly	83.48	88.33	86.44	0.7159
SVM-RBF	83.48	90.56	87.80	0.7427
SVM-Sig	83.48	88.33	86.44	0.7159
K-NN	86.09	74.44	78.98	0.5905
PNN	92.17	70.56	78.98	0.6137

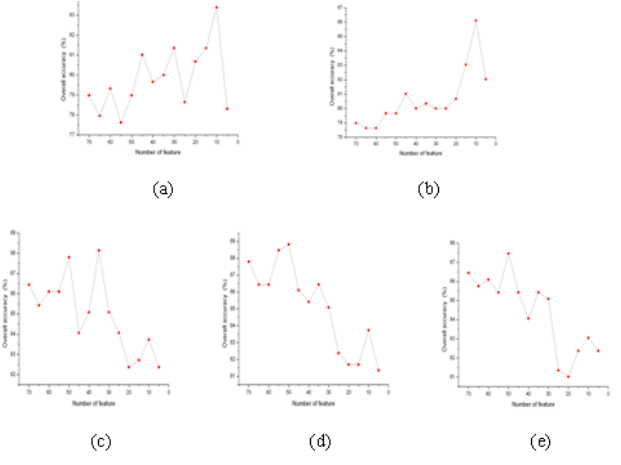

Fig. 1. The accuracy of those classifiers by using the SNR criterion feature ranking and filter under 5-fold cross validation in the computational process. (Fig.1(a), Fig.1(b), Fig.1(c), Fig.1(d) and Fig.1(e) are the results of K-NN, PNN, SVM-Polynomial, SVM-RBF and SVM-Sigmoid respectively.)

Considering each sample of dataset II contains 70 features, SNR was employed to identify the predominant features and filter the irrelevant features to further improve the prediction accuracy and save computational time. In this study, the last five features were filtered each time in the computational process. The results are depicted in Fig.1. It can be found from Fig.1, there exist some local optimal feature sets in the computational process. In order to find out the global optimal set, it is indispensable to searching in the global space.

The number of optimal feature set and the best overall accuracies of evaluating the prognostic risk of recrudescence and metastasis are shown in Table 6. It is illustrated, the best overall recognition result achieved 88.81% by SVM-RBF and its value of *MCC* also reached 0.7696 and excelled to other classifiers, although the number of optimal features attained 50 (compared the original data, it decreases 10), which is more than other classifiers.

Compared Table 5 and Table 6, the accuracies of SVM-Polynomial, SVM-RBF, SVM-Sigmoid, K-NN and PNN are improved from 86.44% to 88.14%, 87.80% to 88.81%, 86.44% to 87.46%, 78.98% to 83.39% and 78.98% to 86.10%

Table 6. 5-fold cross validation results of dataset II by using the classifiers (KNN, SVM, PNN) after feature filtering based on the ranking result of SNR

Classifier	Number of optimal features	Test set			
		$Sen(\%)$	$Spe(\%)$	$Q(\%)$	MCC
SVM-Poly	35	82.61	91.67	88.14	0.7491
SVM-RBF	50	89.57	88.33	88.81	0.7697
SVM-Sig	50	85.22	88.89	87.46	0.7378
K-NN	10	81.74	84.44	83.39	0.6554
PNN	10	93.91	81.11	86.10	0.7323

respectively by using the SNR feature ranking and filtering. It is encouraging that, the PNN performance markedly increases 7.12% rising from the lowest level 78.98% to 86.10% compared to other classifiers, and the number of optimal feature set is only 10, which markedly decreases 60 compared to the original feature set.

The best overall accuracy for dataset II reached 88.81% by using SVM-RBF combined with SNR feature ranking and filtering. SNR not only improved the classification performance, but also decreased the dimensions of the feature set.

4 Conclusions

In this paper, we have investigated the issues of breast cancer diagnosis and prognostic risk evaluation of recrudescence and metastasis by using three classifiers (SVM, KNN, PNN) combined with feature ranking method (SNR), based on FNAB dataset I and gene microarrays dataset II, respectively. Feature ranking and filtering supplied the informative and important features to classify breast tumour. It provides the physicians a valuable clue to pay more attention to these relevant features in their clinical breast tumour diagnosis. Feature ranking and filtering also improved the evaluation performance to the prognostic risk of recrudescence and metastasis, and reduced the dimensions of the feature set. Thus, it also can reduce the computational cost and predigest the process of data collection. The best overall accuracies for breast cancer diagnosis and prognostic risk of recrudescence and metastasis evaluation achieved 96.24% and 88.81% by using SVM-Sigmoid and SVM-RBF respectively. It revealed that classifier and kernel function selection are necessary to get the best results. The study suggests that SVM may be further developed to be a potential practical methodology for clinical assistant breast cancer diagnosis by providing the physicians with the immediate second opinion and is also possible to help the inexperienced physicians avoid misdiagnosis. At the same time, the study also indicates that SVM has the benefits to breast cancer patients as a tool for evaluating the prognostic risk of recrudescence and metastasis. It can make the cured patients who are recognized as easy recrudescence and metastasis pay more attention to their diseases, and then reduce the mortality rate of breast cancer.

Acknowledgements. This work is partly supported by the Natural Science Foundation of Chongqing Science and Technology Committee under Grant No. CSTC, 2006BB5240.

References

1. Parkin, D.M.: Epidemiology of Cancer: Global Patterns and Trends. Toxicology Letters**102-103** (1998) 227-234
2. Butler, S.M., Webb, G.I., Lewis, R.A.: A Case Study in Feature Invention for Breast Cancer Diagnosis Using X-Ray Scatter Images. Lecture Notes in Artificial Intelligence **2903** (2003) 677-685
3. Palmer, G.M., Zhu, C.F., Breslin, T.M., Xu, F.S., Gilchrist, K.W., Ramanujam, N.: Comparison of Multiexcitation Fluorescence and Diffuse Reflectance Spectroscopy for the Diagnosis of Breast Cancer (March 2003). IEEE Transactions on Biomedical Engineering **50** (2003) 1233-1242
4. Song, J.H., Venkatesh, S.S., Conant, E.A., Arger, P.H., Sehgal, C.M.: Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-aided Diagnosis of Breast Masses. Academic Radiology **12** (2005) 487-495
5. Cosar, Z.S., Cetin, M., Tepe, T.K., Cetin, R., Zarali, A.C.: Concordance of Mammographic Classifications of Microcalcifications in Breast Cancer Diagnosis - Utility of the Breast Imaging Reporting and Data System (fourth edition). Clinical Imaging **29** (2005) 389-395
6. Li, J.N., Zhang, Z., Rosenzweig, J., Wang, Y.Y., Chan, D.W.: Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer. Clinical Chemistry **48** (2002) 1296-1304
7. Quinlan, J.R.: Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research **4** (1996) 77-90
8. Abonyi, J., Szeifert, F.: Supervised Fuzzy Clustering for the Identification of Fuzzy Classifiers. Pattern Recognition Letters **24** (2003) 2195-2207
9. Setiono, R.: Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis. Artificial Intelligence in Medicine **18** (2000) 205-219
10. Huang, Y.L., Chen, D.R.: Support Vector Machines in Sonography Application to Decision Making in the Diagnosis of Breast Cancer. J. Clin. Imaging **29** (2005) 179-184
11. Delen, D., Walker, G., Kadam, A.: Predicting Breast Cancer survivability: a Comparison of Three Data Mining Methods. Artificial Intelligence in Medicine **34** (2005) 113-127
12. van't Veer, L.J., Dai, H.Y., van de Vijver, M.J. et al: Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. Nature **415** (2002) 530-536
13. Choudhary, A., Brun, M., Hua, J.P. et al: Genetic Test Bed for Feature Selection. Bioinformatics **22** (2006) 837-842
14. Hadjiiski, L., Sahiner, B., Chan, H.P.: Advances in Computer-aided Diagnosis for Breast Cancer. Curr. Opin. Obstet. Gynecol **18** (2006) 64-70
15. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z.: SVM-Prot: Web-based Support Vector Machine Software for Functional Classification of a Protein from its Primary Sequence. Nucleic Acids Res. **31** (2003) 3692-3697
16. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, Y.Z.: Enzyme Family Classification by Support Vector Machines. Proteins **55** (2004) 66-76

17. Cai, C.Z., Wang, W.L., Sun, L.Z., Chen, Y.Z.: Protein Function Prediction via Support Vector Machine Approach. Mathematical Biosciences **185** (2003) 111-122
18. Tan, S.B.: Neighbor-weighted K-nearest Neighbor for Unbalanced Text Corpus. Expert Systems with Applications **28** (2005) 667-671
19. Yang, Z.R., Platt, M.B., Platt, H.D.: Probabilistic Neural Networks in Bankruptcy Prediction. Journal of Business Research **44** (1999) 67-74
20. Thubthong, N., Kijsirikul, B.: Support Vector Machine for Thai Phoneme Recognition, International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems **9** (2001) 803-813
21. Drucker, H., Wu, D.H., Vapnik, V.N.: Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Networks **10** (1999) 1048-1054
22. Liong, S.Y., Sivapragasam, C.: Flood Stage Forecasting with Support Vector Machines. J. Am. Water Resour. As. **38** (2002) 173-186
23. Wang, Y.J., Chua, C.S., Ho, Y.K.: Facial Feature Detection and Face Recognition From 2D And 3D Images. Pattern Recogn Lett. **23** (2002) 1191-1202
24. Doniger, S., Hofmann, T., Yeh, J.: Predicting Cns Permeability Of Drug Molecules: Comparison Of Neural Network and Support Vector Machine Algorithms. J. Comput. Biol. **9** (2002) 849-864
25. Cao, L.J., Tay, F.: Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting. IEEE T. Neural Network **14** (2003) 1506 -1518
26. Hong, J.H., Cho, S.B.: Lymphoma Cancer Classification using Genetic Programming with SNR Features. Lecture Notes on Computer Science **3003** (2004) 78-88
27. Vapnik, V.: The Nature of Statistical Learning Theory. Springer: New York(1995)
28. Cai, C.Z., Wang, W.L., Chen, Y.Z.: Support Vector Machine Classification Of Physical and Biological Datasets. International Journal of Modern Physics C **14** (2003) 575-585
29. Specht, D.F.: Probabilistic Neural Networks for Classification, Mapping or Associative Memory. IEEE Int. Conf. Neural Network **1** (1988) 525-532
30. Wolberg, W.H., Street, W.N., Heisey, D.H., Mangasarian, O.L.: Computerized Breast Cancer Diagnosis and Prognosis from Fine-Needle Aspirates. Arch. Surg. **130** (1995) 511-516