International Journal Of Computer Science And Applications Vol. 1, No. 1, June 2008
ISSN 0974-1003

Penalty Parameter based Support Vector Machine (PSVM) for Breast Cancer Classification

J.Jaya
Research Scholar
Anna University
Chennai, Tamil Nadu.
09960950621
jjayaphd@yahoo.com

Dr.K.Thanushkodi
Principal
CIET,Coimbatore
Tamil Nadu
0422-2650131
Principal.ciet@kovaikalaimagal.org

ABSTRACT

Computational diagnostic tools are becoming indispensable to support modern medical diagnosis. Soft computing refers to a collection of computational techniques in computer science, artificial intelligence, machine learning and some engineering disciplines, which attempt to study, model, and analyze very complex phenomena. More conventional methods have not yielded low cost, analytic, and complete solutions so far. Earlier computational approaches could model and precisely analyze only relatively simple systems. Simplicity and complexity of systems are relative to technologies. Conventional mathematical models have also been both challenging and productive.

This paper explains how one of the soft computing techniques is useful for breast cancer diagnosis. The method suggested in this paper is Penalty Parameter based Support Vector machine (PSVM).The basic idea is to assign different weights to each class such that the training algorithm learns the decision surface according to the relative importance of data points in the training data set. Experimental results indicate that the proposed method reduces the influence of uneven classification and yields higher classification rate than standard Support Vector Machine (SVM). The proposed system identifies breast tumors with a comparatively high accuracy. The proposed system enables the physicians to adopt proper diagnosis. The main advantage of the proposed system is that the training and diagnosis procedure of PSVM are faster, more stable and reliable than that of multilayer back propagation neural networks. The PSVM is a dependable choice for the Computer Aided Diagnosis (CAD) system because it results in more precise and excellent image classification in addition to being fast.

Keywords

Ultrasound, penalty parameter based Support vector machine (PSVM), Multilayer back propagation, CAD system.

Published by Research Publications, Chikhli, India

1 INTRODUCTION

Breast cancer is one of the major causes for the increase in mortality among women especially in developed countries. Breast cancer is second most common cancer in women. The World Health Organization's International the first being agency for research on cancer in Lyon, France, estimates that more than 1, 50,000 women worldwide die of breast cancer in year[2]. In India, breast cancer accounts for 23% of all the female cancer death followed by cervical cancer (17.5%) in India. Although the incidence is lower in India than developed countries, the burden of diagnosing and treating of breast cancer in India is alarming [9]. The incidence is more among urban than rural women. It is more prevalent in the higher socio-economic groups. The average incidence rate varies from 2 to 28 per 1, 00,000 women per year in urban setting to 6 per 1, 00,000 women per year in rural India. The survey suggests that by 2020 there will be 10 million new cancer cases every year in the developing world, of which 6 million people will die. In India alone, it is estimated that 1.5 million new cancer cases will occur yearly at the start of this century.

Scientists have conducted preliminary studies on the effects of exposure to radiation in space. Since high levels of radiation are known to cause higher-than-normal mutation rates (which can lead to cancer) within a living cell, the levels of cosmic radiation experienced outside of Earth's protective atmosphere will be a large concern in the next era of longer manned spaceflights. [1]

The most frequently adopted medical imaging studies for early detection and diagnosis of breast cancers include mammography and ultrasonography. Ultrasound examination, which is non-invasive and non-radiative, is a more convenient and suitable tool for palpable tumors in daily clinical practice [1]. However, the dense tissues and especially in younger women, cause suspicious region to be almost invisible and may be easily misinterpreted as calcifications and yield a higher False positive (FP) rate that is a major problem with most of the existing algorithms. Ultrasound examination is very operator dependent. The image is non-specific for the diagnosis of benign or malignant lesions according to the echogenecity. The examination described by Stavros et al. [1] is much more extensive than the usual examinations performed at most breast imaging centers. However, the above diagnostic results are achieved by experienced radiologists. Many invasive diagnostic procedures are still required in most cases. Most of these procedures could be avoided if a more specific diagnostic test was available because the rate of positive findings in a biopsy for cancer is low [19].

Sharp diagnostic ultrasound images are often able to show soft layers of breast tissue. Ultrasound may be particularly useful in detecting abnormalities in patients with dense breasts. Density is a term used to describe breast tissue that has many glands close together. Though fairly common (especially in younger women), dense breasts may make breast masses difficult to detect on a mammogram film.

The rest of the paper is organized as follows: section 2 briefly reviews the basic theory and uneven classification problem in SVM. Section 3 deals with the formulation of PSVM. Section 4 describes CAD system. Section 5 experimental results finally the conclusion is given in session 6.

2 STANDARD SVM

2.1 Formulation of SVM

We consider briefly how the SVM binary pattern recognition problem is formulated [20,21]The main idea behind SVM technique is to derive a unique separating hyper plane (i.e., the optimal margin hyper plane) that maximizes the margin between the two classes. Given *l* training data points.

$$\{(x_i, y_i)\}_{i=1,}^{l} x_i \in R^N, y_i \in \{-1,1\}$$

The support vector technique requires the solution of the following optimization problem.

$$\min \phi(w) = \frac{1}{2} w^T w + C\sum_{i=1}^{l} \in_i \quad (1)$$

Subject to

$$y_i[(\dot{w}, \phi(x_i)] + b) \geq 1 - \in_i, i = 1,....,l$$
$$\in_i \geq 0, i = 1,....,l \quad (2)$$

Where the training vectors x_i is mapped into a higher dimensional space by the function. C is a user-specified positive parameter, which controls the tradeoff between classification violation and margin maximization. The existing common method to solve (1) is through its dual, a finite quadratic programming problem

$$W(\alpha) = \sum_{i=1}^{N} \alpha_t - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j [(\Phi(x_i), \Phi(x_i), \Phi(x_j)])$$

(3)

subject to

$$\sum_{i=1}^{N} y_t \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1,....,l \quad (4)$$

Where α is the Lagrangian parameter. Note the kernel trick $K(x,y) = (\Phi(x), \Phi(y))$ is used in the last equality in(3). The Kuhn-Tucker conditions of SVM are defined by

$$a_i[y_i[(\dot{w}, \phi(x_i)] + b) - 1 + \in_i] = 0, i = 1,....,l$$
$$(C - \alpha_i) \in_i = 0, i = 1,....,l \quad (5)$$

The point x_i with the corresponding $\alpha_i > 0$ is called a support vector. The optimal value of weight vector $\omega_o = \sum_{i=1}^{l} \alpha_i y_i \Phi(x_i) = \sum_{i=1}^{ls} \alpha_i y_i \Phi(x_i)$ is obtained where *l*s, is the number of support vectors. The optimal value of bias b_o can be computed from the Kuhn-Tucker conditions (5). Once the optimal pair (ω_o, b_o) is determined, the decision function is obtained by

$$f(x) = [sign[(\dot{w}_0, \phi(x_i)] + b_0)$$
$$= sign(\sum_{i=1}^{l_n} \alpha_i y_i K(x, x_i) + b_0) \quad (6)$$

The sign of the result will indicate which class the test data set belongs to. If the sign of f(*x*) is greater than zero (+ve value) then the test data points (*x*) belongs to '+1' class and if the sign of is less than zero (-ve value) then the data points (*x*) belongs to '-1' class.

2.2 Uneven Classification

If the Training set has uneven class size which results in classification biases towards the class with large training size. The main cause lie in that the penalty of misclassification for each training sample is considered equally, the larger the training class size is , the smaller the corresponding upper limit of classification error rate and there is a bias towards the class with the large training size. The biasing behavior is usually not desirable, particularly in the case of fault diagnosis and disease diagnosis where there is usually a lack of samples belonging to abnormal class.

3 PSVM

The basic idea PSVM is to assign each data point a different weight according to its relative importance in the class such that different data point has different contribution to the learning of the decision surface. Suppose the weights are given, then the training data set becomes

$\{(x_i, y_i, W_i)\}_{i=1}^{l}$, $x_i \in R^N$, $y_i \in \{-1, 1\}$, $W_i \in R$

where the scalar $0 < W_i \leq 1$ is the weight assigned to data point x_i. Where $C_i = CW_i$

where C ,is the penalty parameter.

In the normal SVM a constant value is used through out for calculating the hyper plane. Value of penalty parameter is taken arbitrarily (Such that the value must be grater than zero). For less value maximum margin is obtained and if the value tends to infinity a better classification is obtained. In normal SVM value of constant C is same through out the data set.

In the formulation of PSVM, the SVM wants to maximize the margin of separation and minimize the classification error such that good generalization ability can be achieved. Unlike the penalty term in standard SVM, where the value of C is fixed and

all training data points are equally treated during the training process, PSVM weighs the penalty term in order to reduce the affect of less important data points (such as outliers and noises). The constrained optimization problem is formulated as

Minimize $\Phi(\omega) = \omega^T\omega + C \sum^{l}_{i=1} W_i \xi_i$

In normal SVM α_i is calculated by using quadratic programming by substituting a lower and an upper limit values such that the upper limit depends on the value of C, the penalty parameter in the SVM algorithm. In PSVM if any data points is an outlier then dynamic value of C is used to calculate α_i , so that the data point which is an outlier will get less value of C compared to that of other data points used in training.

4 OVERVIEW OF THE CAD SYSTEM

A computer-aided diagnosis (CAD) system would be expected to be helpful in diagnosing breast cancer because of the difficulty of such diagnoses. [5–9] applied textural features in breast ultrasound images to differentiate between benign and malignant tumors with neural network classifiers.

Textural variation in the ultrasound image has been deemed a useful characteristic for distinguishing benign and malignant tumors [4]. The CAD utilizes a multilayer perception (MLP) neural network to perform a good diagnostic result. However, the training process is prolonged and diagnostic performance normally relies on the initial parameter setting, i.e., number of neurons, learning rate and moment value are hard to decide.

The selections of initial parameters will affect the results drastically. Whereas, the support vector machine (SVM) reveals the feasibility and superiority to extract higher order statistics. The SVM has become extremely popular in terms of classification and prediction.

This study employs the SVM model as a classifier instead of MLP for identifying benign and malignant lesions in the ultrasound image. The proposed diagnosis system can classify the ultrasound images of a breast more accurately and efficiently. The SVM is a reliable choice for the new proposed system because it is fast and excellent in ultrasound image classification.

5 EXPERIMENTAL RESULTS.

5.1 Data Set

The effectiveness of the proposed PSVM algorithm to classification problem is tested on Wisconsin data base. The objective of breast cancer diagnosis is to distinguish malignant from benign breast cytology. The whole dataset have 569 samples, of which there are 357 benign samples (positive class) and 212 malignant samples (negative class). These samples are used to train ,test the PSVM. The result is given below.

Data set	Class	α_i *value*
0.5	-1	0.00000
1.0	-1	93.7327
1.5	+1	**Treated as C value=246.199**

Data set	Class	α_i *value*
0.5	-1	18.1089
2.0	-1	164.2079
1.0	+1	0.00000
2.0	+1	0.00000
2.5	+1	29.95

Table 1. Calculation of α_i *value*

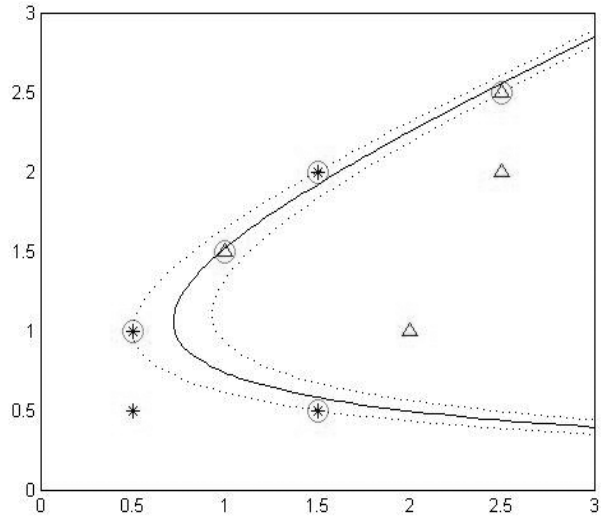

Fig. 1. SVM output With Val C = 1

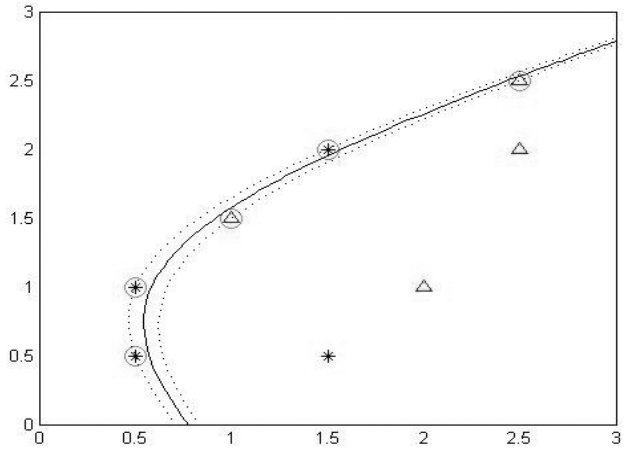

Fig. 2. SVM Output, With C=150

6 SUMMARY AND CONCLUSION

Ultrasound has become one of the major imaging modalities for the diagnosis of breast lesions. Improved imaging techniques permit the management of detected breast lesions to become less invasive. In this paper, PSVM is proposed to deal with the uneven classification problem in traditional support vector machine (SVM) for two-class data classification. Based on the data set,

different weight values are generating for main training data points. Experimental results show that the proposed PSVM can reduce the effect and yield higher classification rate than standard SVM does when the training data set is uneven. The testing time can similarly be reduced by removing the less meaningful support vectors with smaller weights. Though the pruning procedure may induce the loss of training and testing accuracy, this topic is desirable to be further investigated. The proposed CAD system diagnoses breast tumors using texture features within the ultrasound image very well and the CAD system is expected to be a helpful tool.

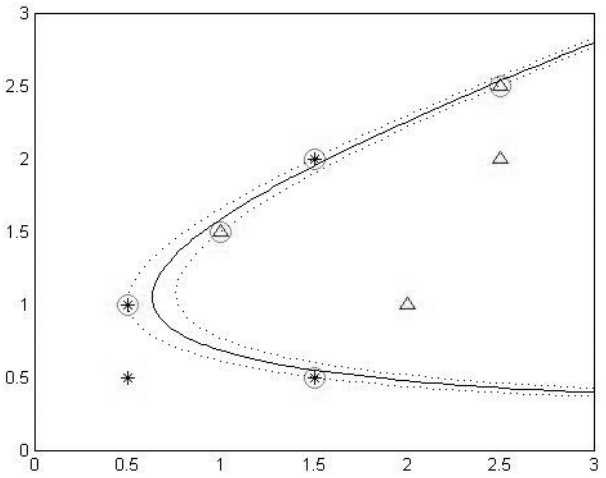

Fig. 3. PSVM output with C = 246.3

7 REFERENCES

[1] American Cancer Society (2003) Breast cancer facts and figures 2001–2002. American Cancer Society, Atlanta, Georgia.

[2] Bassett,LW., Liu,TH., Giuliano,AE., and Gold,RH. (1991) The prevalence of carcinoma in palpable vs impalpable, mammographically detected lesions. AJR Am J Roentgenol 157(1):21–24.

[3] Chen, D., Chang, RF., and Huang, YL. (2000) Breast cancer diagnosisusing self organizing map for sonography. Ultrasound Med Biol 26(3):405–411.

[4] Chen, DR., Chang, RF., and Huang, YL. (1999) Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. Radiology 213(2):407–412.

[5] Chen, DR., Chang, RF., Huang, YL., Chou, YH., Tiu, CM., and Tsai, PP. (2000) Texture analysis of breast tumors on sonograms. Semin Ultrasound CT MR 21(4):308–316.

[6] Chen, DR., Chang, RF., Kuo, WJ., Chen, MC., and Huang, YL. (2002) Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks. Ultrasound Med Biol 28(10):1301–1310.

[7] Christianini, N., and Shawe-Taylor, J. (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, UK.

[8] Detounis, S. Computer aided detection and second readingutility and implementation in a high volume breast clinic,Appl. Radiol. 3 (9) (2004) 8-15.

[9] El Naqa, I., Yang, YY., Wernick, MN., Galatsanos, NP., and Nishikawa, RM. (2002) A support vector machine approach for detection of microcalcifications. IEEE Trans Med Imaging 21(12):1552–63.

[10] http://weboflife.nasa.gov/currentResearch/currentResearchGeneralArchives/breastCancer.htm.http://weboflife.nasa.gov/faq.htm.

[11] Kim, KI., Jung, K., Park, SH., and Kim, HJ. (2002) Support vector machines for texture classification. IEEE Trans Pattern Anal Mach Intell 24(11):1542–1550.

[12] Dorigo,M., Maniczza,V., and Colorni,A., positive feedback as a search strategy.Technical report(16),politecnico di Milano,Italy,1991.

[13] M.K;J:,Siddiqui , M. Anand, P.K. Mehrotr, R. Sarangi, N. Mathur, Biomonitoring of organochlorines in women with benign and malignant breast disease, Environ. Res. 1 (2004)1-8.

[14] Sathish Kumar., Neural Networks –A Classroom Approach, McGraw-Hill companies publication.

[15] Song, MH., Breneman, CM., Bi, JB., Sukumar, N., Bennett, KP., and Cramer, S., et al (2002) Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. J Chem Inf Comput Sci 42(6):1347–1357.

[16] Song, Q., Hu, WJ., and Xie, WF., (2002) Robust support vector machine with bullet hole image classification. IEEE Trans Syst Man Cyber C Appl Rev 32(4):440–448.

[17] Stavros, AT., Thickman,D., Rapp, CL., Dennis, MA., Parker, SH., and Sisney,GA., (1995) Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology 96(1):123–134.

[18] Sun, YF., Fan, XD., and Li, YD., (2003) Identifying splicing sites in eukaryotic RNA: support vector machine approach. Comput Biol Med 33(1):17–29.

[19] Yang, MH., Roth, D., and Ahuja, N. (2002) A tale of two classifiers: SNoW vs. SVM in visual recognition. Comput Vis—ECCV 2353(Pt IV):685–699.