Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 3240–3247
Expert Systems with Applications
www.elsevier.com/locate/eswa

Support vector machines combined with feature selection for breast cancer diagnosis

Mehmet Fatih Akay *

Department of Electrical and Electronics Engineering, Cukurova University, Adana 01330, Turkey

Abstract

Breast cancer is the second largest cause of cancer deaths among women. At the same time, it is also among the most curable cancer types if it can be diagnosed early. Research efforts have reported with increasing confirmation that the support vector machines (SVM) have greater accurate diagnosis ability. In this paper, breast cancer diagnosis based on a SVM-based method combined with feature selection has been proposed. Experiments have been conducted on different training-test partitions of the Wisconsin breast cancer dataset (WBCD), which is commonly used among researchers who use machine learning methods for breast cancer diagnosis. The performance of the method is evaluated using classification accuracy, sensitivity, specificity, positive and negative predictive values, receiver operating characteristic (ROC) curves and confusion matrix. The results show that the highest classification accuracy (99.51%) is obtained for the SVM model that contains five features, and this is very promising compared to the previously reported results.

Keywords: Breast cancer diagnosis; Support vector machines; Feature selection

1. Introduction

Cancer is a group of diseases in which cells in the body grow, change, and multiply out of control. Usually, cancer is named after the body part in which it originated; thus, breast cancer refers to the erratic growth of cells that originate in the breast tissue. A group of rapidly dividing cells may form a lump or mass of extra tissue. These masses are called tumors. Tumors can either be cancerous (malignant) or non-cancerous (benign). Malignant tumors penetrate and destroy healthy body tissues.

The term, breast cancer, 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 (exceeded only by lung cancer) (http://www.imaginis.com/breasthealth/breast_cancer.asp, Last Accessed August 2007). According to the World Health Organization, more than 1.2 million women will be diagnosed with breast cancer each year worldwide. Fortunately, the mortality rate from breast cancer has decreased in recent years with an increased emphasis on diagnostic techniques and more effective treatments. A key factor in this trend is the early detection and accurate diagnosis of this disease (West, Mangiameli, Rampal, & West, 2005).

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patients and decisions of experts are the most important factors in diagnosis. However, expert systems and different artificial intelligence techniques for classification also help experts in a great deal. Classification systems can help minimizing possible errors that can be done because of inexperienced experts, and also provide medical data to be examined in shorter time and more detailed.

SVM have been proposed as an effective statistical learning method for classification (Vapnik, 1989). They rely on so called support vectors (SV) to identify the decision

* Corresponding author.
E-mail address: mfakay@cu.edu.tr

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.01.009

boundaries between different classes. SVM are based on a linear machine in a high dimensional feature space, nonlinearly related to the input space, which has allowed the development of somewhat fast training techniques, even with a large number of input variables and big training sets. SVM have been used successfully for the solution of many problems including handwritten digit recognition (Scholkopf et al., 1997), object recognition (Pontil & Verri, 1998), speaker identification (Wan & Campbell, 2000), face detection in images (Osuna, Freund, & Girosi, 1997), and text categorization (Joachims, 1999).

When using SVM, three problems are confronted: how to choose the kernel function and optimal input feature subset for SVM, and how to set the best kernel parameters. These problems are crucial because the feature subset choice influences the appropriate kernel parameters and vice versa (Frohlich et al., 2003). Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier to in order to have a good predictive and less computationally intensive model (Zhang, 2000). With a small feature set, the explanation of rationale for the classification decision can be more readily realized.

In this study, SVM with feature selection was used to diagnose the breast cancer. WBCD taken from the University of California at Irvine (UCI) machine learning repository was used for training and testing experiments (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin, Last Accessed August 2007). It was observed that the proposed method yielded the highest classification accuracies (98.53%, 99.02%, and 99.51% for 50–50% of training-test partition, 70–30% of training-test partition, and 80–20% of training-test partition, respectively) for a subset that contained five features. Also, other measures such as the confusion matrix, sensitivity, specificity, positive predictive value, negative predictive value and ROC curves were used to show the performance of SVM with feature selection.

The rest of the paper is organized as follows. Section 2 summarizes the methods and results of previous research on breast cancer diagnosis. Section 3 reviews basic SVM concepts. Section 4 describes the proposed method. Section 5 presents experimental results from using the proposed method to diagnose breast cancer. Finally, Section 6 concludes the paper along with outlining future directions.

2. Related work on breast cancer diagnosis

There has been a lot of research on medical diagnosis of breast cancer with WBCD in literature, and most of them reported high classification accuracies. In Albrecht, Lappas, Vinterbo, Wong, and Ohno-Machado (2002), a learning algorithm that combined logarithmic simulated annealing with the perceptron algorithm was used and the reported accuracy was 98.8%. In Pena-Reyes and Sipper (1999), the classification technique used fuzzy-GA method reaching a classification accuracy of 97.36%. In Setiono (2000), the classification was based on a feed forward neural network rule extraction algorithm. The reported accuracy was 98.10%. (Quinlan, 1996) reached 94.74% classification accuracy using 10-fold cross-validation with C4.5 decision tree method. (Hamiton, Shan, & Cercone, 1996) obtained 94.99% accuracy with RIAC method, while (Ster & Dobnikar, 1996) obtained 96.8% with linear discreet analysis method. The accuracy obtained by Nauck and Kruse (1999) was 95.06% with neuron-fuzzy techniques. In Goodman, Boggess, and Watkins (2002), three different methods, optimized learning vector quantization (LVQ), big LVQ, and artificial immune recognition system (AIRS), were applied and the obtained accuracies were 96.7%, 96.8%, and 97.2%, respectively. In Abonyi and Szeifert (2003), an accuracy of 95.57% was obtained with the application of supervised fuzzy clustering technique. In Polat and Gunes (2007), least square SVM was used and an accuracy of 98.53% was obtained.

3. Support vector machines

3.1. Linear SVM

Consider the problem of separating the set of training vectors belonging to two linearly separable classes,

$$(\mathbf{x}_i, y_i), \quad \mathbf{x}_i \in R^n, \quad y_i \in \{+1, -1\}, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i$ is a real-valued n-dimensional input vector and y_i is a label that determines the class of $\mathbf{x}_i$. A separating hyperplane is determined by an orthogonal vector $\mathbf{w}$ and a bias b, which identifies the points that satisfy

$$\mathbf{w}.\mathbf{x} + b = 0. \tag{2}$$

The parameters $\mathbf{w}$ and b are constrained by

$$\min_i |\mathbf{w}.\mathbf{x}_i + b| \geqslant 1. \tag{3}$$

A separating hyperplane in canonical form must satisfy the following constraints,

$$y_i(\mathbf{w}.\mathbf{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, n. \tag{4}$$

The hyperplane that optimally separates the data is the one that minimizes

$$\Phi(\mathbf{w}) = \frac{1}{2}(\mathbf{w}.\mathbf{w}). \tag{5}$$

Relaxing the constraints of (4) by introducing slack variables $\xi_i \geqslant 0$, $i = 1, 2, \ldots, n$, (4) becomes

$$y_i.(\mathbf{w}.\mathbf{x}_i + b) \geqslant 1 - \xi_i, \quad i = 1, 2, \ldots, n. \tag{6}$$

In this case, the optimization problem becomes

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}(\mathbf{w}.\mathbf{w}) + C\sum_{i=1}^{n} \xi_i \tag{7}$$

with a user defined positive finite constant C. The solution to the optimization problem in (7), under the constraints of