PCA and SVM Based on Multiple Kernels for Breast Cancer Classification

Xiu-feng Yang  
Shenyang Institute of Automation, Chinese Academy of Sciences  
University of Chinese Academy of Sciences  
Beijing, China  
e-mail: yangxiufeng@sia.cn

Hui Peng, Xiao-feng Zhou, Yong-lai Zhang  
Shenyang Institute of Automation, Chinese Academy of Sciences  
Shenyang, China  
e-mail: zhouxf@sia.cn

Abstract—In this paper, we propose an efficient algorithm based on Principal Component Analysis (PCA) and Support Vector Machines (SVM) with multiple kernels in the process of breast cancer classification. The first step, we use PCA to project high dimensional breast cancer data into much lower dimensional space. Second, we use SVM with multiple kernels to classify the lower dimensional breast cancer data. Finally, the experimental and analytical results show that the proposed algorithm has a better performance than traditional SVM for breast cancer classification.

Keywords—breast cancer; Principal Component Analysis (PCA); Support Vector Machines (SVM); multiple kernels.

I. INTRODUCTION

In 2008, the world health organization (WHO) [1] investigation shows that there are about 76 million people die of cancer. Breast cancer is the most common tumor-related disease all over the world, breast cancer is one of the diseases that can cause greatly high rate mortality, seriously threatening women health. Breast cancer cells can be divided into benign tumor and malignant tumor. Fine needle aspirations (FNAs) provide a way to examine a small amount of tissue from the tumor. And Some doctors and medical institutions use( FNAs) has successfully distinguish the tumors [2]. But, due to the different characteristics of malignant tumors, professional doctors always rely on their experience to judge whether the cancer is a malignant or a benign tumor. So, it is very necessary to introduce a kind of technology that can accurately diagnose the breast cancer.

In recent years, Support Vector Machine (SVM) proposed by Vapnik has become a kind of very important classification techniques in the field of pattern recognition. SVM can solve the linear and nonlinear problems, and has showed very good performance in classification. Different from the traditional classification algorithm which is based on empirical risk minimization, the SVM is a kind of technology based on the principle of structure risk minimization [3,4,5]. SVM classifier can map the inseparable input vector into high-dimensional space by kernel function, and construct an optimal separating hyperplane in the high-dimensional space to solve the classification problems, which is shown in Fig. 1.

However, kernel functions of support vector machine are divided into two categories: local kernel function and global kernel function. Local kernel function has a good learning ability, while global kernel function has a good prediction ability [8]. Choosing different kernel functions of SVM has great influence on the performance of SVM model. So, we have to choose a good SVM model which has both better learning ability and prediction ability.

In this paper, we propose an algorithm combining PCA and support vector machine with multiple kernel functions for breast tumor classification. Firstly, PCA is implemented to reduce the dimension of the breast cancer data. This method can extract good features from the input data without adding the computing complexity. We use PCA to the data set and we reduce the dimension of the original data into 2 and 3 dimension, and the result is shown in Fig. 2(red is Benign tumors, green is Malignant tumors.). Then, a SVM model with multiple kernels is provided to classify the lower dimensional data. The results of the experiment show
that the correct classification ratio is increased by the proposed algorithm.

The paper is organized as follows. The basic theory of the PCA is simply introduced in section 2, we describe briefly the algorithm of support vector machine and construct multiple kernel function in section 3, we focus on results and analysis of the experiment in section 4, finally, conclusion are made in section 5.

II THEORY OF PCA

PCA is performed to simplify the related variables of the original dataset by reducing the dimensionality. This technique is able to convert raw data into a new dataset of relevant variables, which reduces the dimensionality of the original dataset and achieves the goal of simplifying the dataset. Steps of feature extraction are as follows:

Step I: Standardize the original data:

\[ y_j = \frac{x_j - \bar{x}_j}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2}} \]  

(1)

Where \( x_{ij} \) is the primary data index-\( i \) of sample-\( j \);

(\( i = 1,2...,n; j = 1,2...,m \)), \( \bar{x}_j \) is \( \sum_{i=1}^{n} x_{ij} \) divided by \( n \).

Step II: Calculate the correlation matrix \( R = (r_{ij})_{m \times m} \):

\[ r_{ij} = \frac{1}{n} \sum_{p=1}^{n} (x_{pi} - \bar{x}_i)(x_{pj} - \bar{x}_j) \sigma_i \sigma_j \]  

(2)

The eigenvalues of the correlation matrix is:

\( E_1, E_2, ..., E_m \) \( (E \geq 0) \) and the corresponding eigenvectors is \( E_{v1}, E_{v2}, ..., E_{vm} \).

Step III: Calculate contribution rate:

The contribution rate of each index Can be concluded by

\[ \sum_{j=1}^{m} E_j / \sum_{j=1}^{m} E_j \] is accumulative contribution rate .

When the cumulative rate meets the requirement of accuracy , the top indexes will be selected as the principal components.

III SVM

Supposing that we have a sample data set \( S = \{(x_i, y_i), i = 1,2,...,l\}, x_i \in R, y_i \in [-1,1] \) corresponding to the two classes. If \( x_i \in R \) belong to the first class then \( y_i = 1 \), if \( x_i \in R \) belong to the second class then \( y_i = -1 \), learning goal is to construct a decision function to classify the test data as correct as possible. In view of the training sample data sets, we will discuss two cases : linear or nonlinear.

A. Linear Case

Supposing that there exists a hyperplane with function \( \omega^T \cdot x + b = 0 \) ( \( \omega \) represents the weight vector and \( b \) is the bias ) makes the following two function valid for all elements of the training set

\[ \omega^T \cdot x + b \geq 1, y_i = 1 \]
\[ \omega^T \cdot x + b \leq -1, y_i = -1 \]  

(3)

Below we write the inequalities in the form:
\[ y_i \left( \omega^T \cdot x + b \right) \geq 1, i = 1, 2, \ldots, l \]  \hspace{1cm} (4)

The optimal hyperplane can separate the training data with a maximal margin. The decision function is as follows:

\[ f(x) = \text{sgn} \left( \omega^T \cdot x + b \right) \]  \hspace{1cm} (5)

To solve the optimal hyperplane we need to maximize the margin of separation \( \frac{2}{\| \omega \|} \), this equal to minimize \( \frac{1}{2} \| \omega \|^2 \). Then, in the case of linear separation, the linear SVM for optimal separating hyperplane is equal to solve the following optimization problem:

Minimize \( \frac{1}{2} \| \omega \|^2 \),

s.t. \( y_i \left( \omega^T \cdot x + b \right) \geq 1, i = 1, 2, \ldots, l \)  \hspace{1cm} (6)

Using the Lagrange multiplier method to solve the Quadratic programming problem:

\[
L(\omega, b, \beta) = \frac{1}{2} \| \omega \|^2 - \sum_{i=1}^{l} \beta_i \left[ y_i (\omega^T \cdot x + b) - 1 \right] \hspace{1cm} (7)
\]

Where \( \beta_i \) is the Lagrange multiplier, according to the classical Lagrangian duality, (7) can be transformed to its dual problem [12]:

Maximize \( Q(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j y_i y_j K(x_i, x_j) \),

s.t. \( \beta_i, \beta_j \geq 0, \sum_{i=1}^{l} \beta_i y_i = 0 \)  \hspace{1cm} (8)

in the case that training set is linear non-separate, slack variable \( \varepsilon_i \geq 0, i = 1, 2, \ldots, l \) is introduced to solve the linear non-separate problem. The optimal separating hyperplane becomes the optimization problem,

Minimize \( \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} \varepsilon_i \),

s.t. \( y_i \left( \omega^T \cdot x + b \right) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1, 2, \ldots, l \)  \hspace{1cm} (9)

Where \( C \) is Penalty parameters.

**B. Nonlinear Case**

When the training set is nonlinear, training samples will be mapped into a higher dimensional linear space from input space by a nonlinear mapping function \( \phi(\cdot) \). The optimal separating hyperplane will be constructed in the higher dimensional space, so the function of the separating hyperplane is: \( \omega^T \cdot \phi(x) + b = 0 \) the problem of the optimal separating hyperplane can be described as follows:

Minimize \( \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} \varepsilon_i \),

s.t. \( y_i \left( \omega^T \cdot \phi(x_i) + b \right) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1, 2, \ldots, l \)  \hspace{1cm} (10)

Similar to the linear case, the problem of the optimal separating hyperplane can be transformed to its dual problem:

Maximize \( Q(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j y_i y_j K(x_i, x_j) \),

s.t. \( \sum_{i=1}^{l} \beta_i y_i = 0, 0 \leq \beta_i \leq C \)  \hspace{1cm} (11)

where \( K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \) is called kernel function which satisfies the Mercer’s conditions. The final decision function is as follows:

\[ f(x) = \text{sgn} \left( \sum_{i=1}^{l} \beta_i y_i K(x_i, x_j) + b \right) \hspace{1cm} (12) \]

**IV CONSTRUCTION OF MULTIPLE KERNELS**

The type of kernel function is directly determines many properties of SVM model, different kernel function of SVM for the same data set can produce different classification results. In local kernels only the data that are close or in the proximity of each other have an influence on the kernel values, while in the global kernels only the data that are far away from each other have an influence on the kernel values[9][10][11].

A typical representation of the global kernel function is polynomial kernel function:

\[ K(x, x_j) = \left( \langle x, x_j \rangle + 1 \right)^d \hspace{1cm} (13) \]

A typical representation of the local kernel function is Radial Basis kernel function (RBF):

\[ K(x, x_j) = e^{-\frac{\| x - x_j \|^2}{\sigma^2}} \hspace{1cm} (14) \]

The quality of SVM is not only determined by its learning ability, but also determined by its generalization ability.

Global kernel function has better generalization performance than local kernel functions. However, in terms of learning ability, local kernel function is much better. And SVM with single kernel rarely has these two properties. So, we consider that by combining the good characteristics of
two kernels, the SVM will have better performance in classification. So we use the combination of different kernels in SVM algorithm.

V EXPERIMENTS AND ANALYSIS

A. Experimental Data

In this paper, we use the data set that can be obtained from UCI, to show how our algorithm works. The data set has 569 samples. Each sample represents a real patient and has 30 medical features. 357 samples are benign tumors, 212 samples are malignant tumors. Ten real-valued features are computed for each cell nucleus: radius (mean of distances from center to points on the perimeter), texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension.

B. Experimental Procedure

Step I: First of all, in order to get better accuracy, input set are normalized to the range of [0, 1].

Step II: Then PCA is used to simplify the data set by reducing dimension. Each principle component contribution rate and the accumulative contribution rate of the first ten principal components are shown in the Table I. We can see that the accumulative contribution rate of the first ten principle components has reached 94.821%.

Step III: optimize penalty by PSO, we design the SVM models based on linear kernel function, RBF kernel function and polynomial function and record the results of each experiment.

<table>
<thead>
<tr>
<th>Principle component</th>
<th>Contribution</th>
<th>Accumulative contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>53.097%</td>
<td>53.097%</td>
</tr>
<tr>
<td>P2</td>
<td>17.283%</td>
<td>70.38%</td>
</tr>
<tr>
<td>P3</td>
<td>7.114%</td>
<td>77.494%</td>
</tr>
<tr>
<td>P4</td>
<td>6.411%</td>
<td>83.904%</td>
</tr>
<tr>
<td>P5</td>
<td>4.086%</td>
<td>90.99%</td>
</tr>
<tr>
<td>P6</td>
<td>3.071%</td>
<td>92.061%</td>
</tr>
<tr>
<td>P7</td>
<td>1.581%</td>
<td>91.624%</td>
</tr>
<tr>
<td>P8</td>
<td>1.191%</td>
<td>93.833%</td>
</tr>
<tr>
<td>P9</td>
<td>0.988%</td>
<td>94.821%</td>
</tr>
</tbody>
</table>

C. Analysis of the Experimental Results.

According to the results of the experiment, we can see that PCA&SVM model with combination kernels of linear and RBF get the worse performance. PCA&SVM with kernels of Poly and RBF has the better performance than traditional SVM models with single kernel.

VI CONCLUSION

SVM is a very good method to solve the nonlinear and high dimensional data classification problems. PCA can reduce the high-dimensional data into a much lower dimension space. Although data is reduced into a low dimension space, the classification model can still maintained a high classification accuracy rate. In this paper, first we use PCA to reduce the dimension of the breast cancer data into a lower space and then use SVM to classify the data, and the experiment results show that the SVM with combination kernels has a higher classification correct rate than traditional SVM with single kernel.

ACKNOWLEDGMENT

This work is supported by major program of the National Natural Science Foundation of Beijing, China (grant No. 7110001).

REFERENCES


