Proceeding of the IEEE
International Conference on Information and Automation
Yinchuan, China, August 2013

SVM with Multiple Kernels based on Manifold Learning for Breast Cancer Diagnosis

Xiufeng Yang[1,2]， Hui Peng[1], Mingrui Shi[3]
[1]*Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China*
[2]*University of Chinese Academy of Sciences , Beijing, China*
[3]*Department of Control Science and Engineering ,Zhejiang University, Hangzhou, China*
yangxiufeng@sia.cn

***Abstract*—In this paper, we propose an efficient algorithm Support Vector Machines with multiple kernels based on Isometric feature mapping(Isomap) in the process of breast cancer classification. We use Wisconsin Diagnostic Breast Cancer (WDBC) as our original data set. The first step, we use Isomap to project high dimensional breast cancer data into a much lower dimensional space. Second , we use SVM with multiple kernels to classify the lower dimensional breast cancer data. Finally, the experimental results illustrate that the proposed algorithm has a better performance than traditional SVM for breast cancer classification.**

Keywords—Breast Cancer, Isometric feature mapping (Isomap), Support Vector Machines(SVM), multiple kernels.

I. INTRODUCTION

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. 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. [3]So , it is very necessary to introduce a kind of technology that can help doctors 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[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. when the dimensionality of the samples is very high, the efficiency of constructing separating hyperplane will degrade greatly. Many approaches have been proposed for dimensionality reduction, such as the well-known methods of principal component analysis (PCA) [8], independent component analysis (ICA) [6], and multidimensional scaling (MDS) [7]. In PCA, the main idea is to find the projection that restores the largest possible variance in the original data. ICA is similar to PCA except that the components are designed to be independent. Finally, in MDS, efforts are taken to find the low-dimensional embeddings that best preserve the pair wise distances between the original data points. All of these methods are easy to implement. At the same time, their optimizations are well understood and efficient. Because of these advantages, they have been widely used in visualization and classification. Unfortunately, they have a common inherent limitation: They are all linear methods while the distributions of most real-world data are nonlinear.

Recently, two novel methods have been proposed to tackle the nonlinear dimensionality reduction problem, namely Isomap [10] and LLE [11]. Both of these methods attempt to preserve as well as possible the local neighborhood of each object while trying to obtain highly nonlinear embeddings. So they are categorized as a new kind of dimensionality reduction techniques called local embeddings [12]. The central idea of local embeddings is using the locally linear to solve the globally nonlinear problems, which is based on the assumption that data lying on a nonlinear manifold can be viewed as linear in local areas. And both of these two methods can solve nonlinear dimensionality reduction. 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[9]. 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 Isomap and support vector machine with multiple kernel functions for breast tumor classification. Firstly , Isomap is implemented to reduce the dimension of the breast cancer data . we reduce the dimension of the original data into a lower dimension .Then, a SVM model with multiple kernels is provided to classify the lower dimensional data. The results of the experiment are shown in Receiver Operating Characteristics(ROC) curve[13] and area under ROC curve(AUC)which illustrates that the performance is better than traditional SVM model.

The paper is organized as follows. The basic theory of the Isomap is simply introduced in section2, we describe briefly the algorithm of support vector machine and construct multiple kernel functions in section 3,4. We focus on results and analysis of the experiment in section5, finally, conclusion

978-1-4799-1334-3/13/$31.00 ©2013 IEEE

are made in section 6.

II. ISOMAP

The main purpose of Isomap is to find the intrinsic geometry of the data, as captured in the geodesic manifold distances between all pairs of data points. The approximation of geodesic distance is divided into two cases. In case of neighboring points, Euclidean distance in the input space provides a good approximation to geodesic distance. In case of faraway points, geodesic distance can be approximated by adding up a sequence of "short hops" between neighboring points. Isomap shares some advantages with PCA,LDA, and MDS, such as computational efficiency and asymptotic convergence guarantees, but with more flexibility to learn abroad class of nonlinear manifolds[12]. The detailed steps of Isomap are listed as follows.

Step I: Construct neighborhood graph: Define the graph G over all data points by connecting ny two points x_i and x_j, if x_i is one of the k-nearest neighbor of x_j. Set edge lengths equal to $d_X(x_i, x_j)$.

Step II: Compute shortest paths: $d_G(x_i, x_j) = d_X(x_i, x_j)$ if x_i and x_j are neighbors; $d_G(x_i, x_j) = \infty$ otherwise. Then for each value of $K = 1,2,..., N$, in turn , replace all entries $d_G(x_i, x_j)$ by min $\{d_G(x_i, x_j), d_G(x_i, x_k) + d_G(x_K, x_j)\}$. The matrix of final value $D_G = \{d_G(x_i, x_j)\}$ will contain the shortest path distances between all pairs of points in G

Step III: Construct d-dimensional embedding: Let λ_p be the p-th eigenvalue (in decreasing order) of the matrix $\tau(D_G)$ (the operate τ is defined by $\tau(D) = HSH/2$, where S is the matrix of squared distances $\{S_{ij} = D_{ij}^2\}$, and the H is the " centering matrix" $\{H_{ij} = \sigma_{ij} - 1/N\}$, σ_{ij} is the Kronecker delta function), and V_p^i be the i-th component of the p-th eigenvector. Then set p-th component of the d-dimensional coordinate vector y_i equal to $\sqrt{\lambda_p} V_p^i$.

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 $\omega^T \cdot x + b = 0$ (ω represents the weight vector and b is the bias) makes the following two function valid for all elements of the training set

$$\begin{aligned} \omega^T \cdot x + b \geq 1, y_i = 1 \\ \omega^T \cdot x + b \leq -1, y_i = -1 \end{aligned} \quad (1)$$

Below we write the inequalities in the form ：

$$y_i(\omega^T \cdot x + b) \geq 1, i = 1,2..., l \quad (2)$$

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

$$f(x) = \text{sgn}(\omega^T \cdot x + b) \quad (3)$$

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:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2} \| \omega \|^2 \\ &\text{s.t.} \quad y_i(\omega^T \cdot x + b) \geq 1, i = 1,2..., l \end{aligned} \quad (4)$$

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

$$L(\omega, b, \beta) = \frac{1}{2} \| \omega \|^2 - \sum_{i=1}^{l} \beta_i [y_i(\omega^T \cdot x + b) - 1] \quad (5)$$

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

$$\begin{aligned} &\text{Maximize } Q(\beta) = \sum_{i=1}^{l} \beta_i - \frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j y_i y_j (x_i^T x_j) \\ &\text{s.t.} \quad \beta_i, \beta_j \geq 0, \sum_{i=1}^{l} \beta_i y = 0 \end{aligned} \quad (6)$$

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

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} \varepsilon_i \\ &\text{s.t.} \quad y_i(\omega^T \cdot x_i + b) \geq 1 - \varepsilon_i \\ &\qquad \varepsilon_i \geq 0, i = 1,2..., l \end{aligned} \quad (7)$$

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:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} \varepsilon_i \\ &\text{s.t.} \quad y_i(\omega^T \cdot \phi(x_i) + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1,2..., l \end{aligned} \quad (8)$$

Similar to the linear case, the problem of the optimal

separating hyperplane can be transformed to its dual problem:

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

$$\text{s.t.} \quad \sum_{i=1}^{l} \beta_i y_i = 0, \quad 0 \le \beta_i \le C \tag{9}$$

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:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \beta_i y_i K(x_i, x_j) + b \right) \tag{10}$$

IV. MUTIPLE KERNELS FOR SVM

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 way from each other have an influence on the kernel values[9].

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

$$K(x, x_j) = \left[(x, x_i) + 1 \right]^d \tag{11}$$

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

$$K(x, x_i) = e^{\frac{\left[|x - x_i|^2 \right]}{2\sigma^2}} \tag{12}$$

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 perpormance in classification. So we use the combination of RBF and Polynomial as our kernel function

$$K_d(x, x_j) = (1 - \rho)\left[(x, x_i) + 1 \right]^d + \rho e^{\frac{\left[|x - x_i|^2 \right]}{2\sigma^2}} \tag{13}$$

V. EXPERIMENTS AND ANALYSIS

A. *Experimental data*

In this paper, we use the data set that can be obtained from UCI[2], 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.

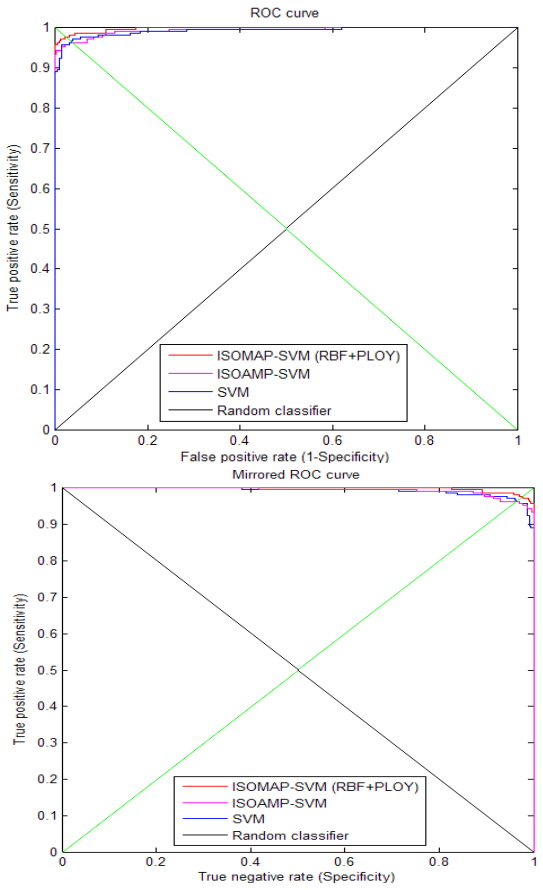

Fig. 1 ROC curve of different algorithms

B. *Experimental procedure*

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

StepII:Then Isomap is used to simplify the data set by reducing dimension, and the dimension of the original data is reduced into 5 dimension.

Step III: we design the SVM models based on RBF kernel function . Then construct SVM models with multiple kernels with combination of RBF and polynomial kernel function and use these models to classify the data set , the results are recorded.

C. *Analysis of the experimental results* .

In this experiment , ten times ten-fold cross validation is run. That is, in each time, the original data set is randomly divided into ten equal-sized subsets while keeping the proportion of the instances in different classes. Then, in each fold, one subset is used as testing set and the union of the remaining ones is used as training set. After ten folds, each subset has been used as testing set once. The average result of these ten folds is recorded.

Isomap with $K = 10$ is used to reduce the original data into 5 dimension and we use SVM and SVM with combination of RBF and Ploy kernel to classify the lower dimensional data . In SVM model the parameters of RBF kernel are $\sigma = 2.8, c = 100$. And in multiple kernel model the parameters are set $\rho = 0.05$, $\sigma = 5$, $c = 120$, $d = 4$. ROC

curve is used to illustrate the performance , the result is shown in figure 1. we also implement the AUC to contrast the performance, and the results are shown in figue2. And the dimension of the original data is reduced from 1 to 5 and the AUC varies . if the dimension equals to 1, SVM is better than other two models, and when the dimension of the data is reduced to 5, We can see that AUC of our proposed algorithm is larger than ISOMAP-SVM and SVM.

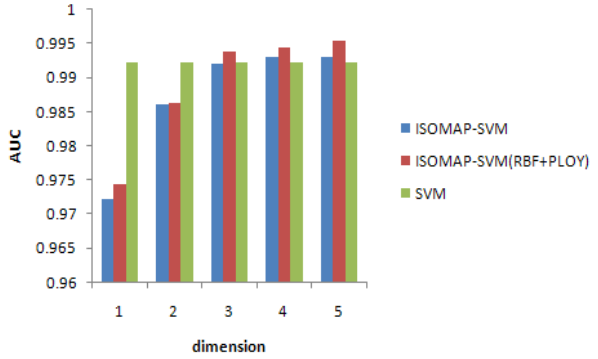

Fig.2 AUC of different dimension of the orginal data

Finally , we set dimension equal to 5 , and the accuracy of the testing data set is shown in table I. The accuracy of ISOMAP-SVM(RBF+PLOY) is much higher than other algorithms.

TABLE I Accuracy of testing data

Method	kernel	Test accuracy
SVM	RBF	97.6331%
ISOMAP-SVM	RBF	96.4497%
ISOMAP-SVM	(1-ρ) Poly +ρ RBF	98.224 %

According to the results of the experiment, the AUC of the model and the accuracy of the testing data show that ISOMAP-SVM model with combination kernels of Poly and RBF has better performance than traditional SVM model with single kernel for Breast Cancer classification.

VI. CONCLUSION

SVM is a very good method to solve the nonlinear and high dimensional data classification problems. Isomap can reduce the high-dimensional data into a much lower dimension space. In this paper, fist we use Isomap to reduce the dimension of the breast cancer data into a lower space and then use SVM to classify the data , and the performance of the algorithm is shown in ROC curve and AUC. The experiment results show that the ISOMAP-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

[1] World Health Organization FactSheet ,Cancer ,http://www. Who.int /media center/factsheets/fs297/en/.

[2] Asunction, A&Newman,D.J.(2007).UCI Machine Learning Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California ,Department of Information and Computer Science.

[3] W. Nick Street, William H. Wolberg, O.L. Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis," 1993 International Symposium on Electronic Imaging : Science and Technology, San Jose, California, vol. 1905, pp.861-870.

[4] Cortes, C. and Vapnik, V. N.(1995) "Support-Vector Networks," Machine Learning 20(3):273-297.

[5] Vapnik, V.N."An overview of statistical learning theory," IEEE Transactions on Neural NETWORKS 10(5): 988-999.

[6] P. Comon, "Independent component analysis: A new concept," Signal Process., vol. 36, no. 3, pp. 287–314, 1994.

[7] T. Cox and M. Cox, Multidimensional Scaling. London, U.K.:Chapman & Hall, 1994.

[8] I. T. Jolliffe, Principal Component Analysis. New York: Springer,1986.

[9] Smits,Guido F. "Improved SVM regression using mixtures of kernels," Proceedings of the 2002 International Joint Conference on Neural Networks. Vol.3 ,pp.2785 - 2790 .

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290,

[11] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by local linear embedding," Science, vol. 290, no. 5500, pp. 2323–2326,2000

[12] Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART B: CYBERNETICS, VOL. 35, NO. 6, DECEMBER 2005.

[13] Spackman.Signal detection theory: valuable tools for evaluating inductive learning[C]/ /Proceedings of the Sixth International Workshop on Machine Learning, 1989.