A Gaussian Radial Basis Function Based Feature Selection Algorithm

Zhiliang Liu[1,2]
[1]School of Automation Engineering
University of Electronic Science and Technology of China
Chengdu, China

Ming J. Zuo
[2]Department of Mechanical Engineering
University of Alberta
Edmonton, Canada
ming.zuo@ualberta.ca

Hongbing Xu
School of Automation Engineering
University of Electronic Science and Technology of China
Chengdu, China

***Abstract*—Recently Li et al. proposed a parameter selection method for Gaussian radial basis function (GRBF) in support vector machine (SVM). In his paper cosine similarity was calculated between two vectors based on the properties of GRBF kernel function. Li's method can determine an optimal sigma in SVM and thus efficiently improve its performance, yet it is limited by only focusing on a fixed original feature space and may suffer if the space contains some irrelevant and redundant features, especially in a high-dimensional feature space. In this paper, Li's method is extended to a flexible feature space so that feature selection and parameter selection are conducted at the same time. A feature subset and sigma are determined by minimizing the objective function that considers both within-class and between-class cosine similarities. Our experimental results demonstrate that the proposed method has a better performance than Li's method and traditional SVM in terms of classification accuracy.**

Keywords-cosine similarity; feature selection; Gaussian radial basis function; parameter selection;

I. INTRODUCTION

In the machine learning area, data preprocessing is important for data mining algorithms. Dimension reduction is one of the usual objectives of preprocessing, and it aims to alleviate the curse of dimensionality and Hughes phenomenon [1], focus on relevant features only, and improve the performance of data mining algorithms, e.g. in terms of classification accuracy and computational time. Theoretical analyses and experimental studies both confirm that many algorithms scale poorly when there are a large number of irrelevant and redundant features [2]. Therefore, it is necessary and significant to do dimension reduction in the preprocessing step. There are two representative categories of dimension reduction: feature extraction and feature selection. Feature extraction is to generate a set of new features from the original feature space through some functional projection [3]. Principal component analysis (PCA), independent component analysis (ICA), and linear discriminant analysis (LDA) are three representative statistical methods in this category. Several new methods such as the nonparametric weighted feature extraction (NWFE) [4] and the LDA-based clustering feature extraction [5] have been proposed recently.

On the other hand, feature selection attempts to find a subset with M features from the original feature space of N features ($M \leq N$) so that the space is optimally reduced according to a certain criterion [6], such as classification accuracy, mutual information [7], Pearson correlation coefficient, χ^2-statistic, t-statistic, and reliefF [8]. Qu et al. used the norm of the weight vector ($\|\boldsymbol{\omega}\|$) of SVM as the performance measure and applied their feature selection method to damage degree classification of planet gears [9]. Compared to feature extraction, feature selection has the advantage of interpreting selected features intuitively because it does not change the original features.

Nowadays, researchers pay much attention to feature selection and apply it to various areas including text processing of internet documents, gene expression array analysis [8], combinatorial chemistry, and fault diagnosis [9] [10]. Feature selection is normally grouped into filters, wrappers, and embedded methods. Filters do feature selection before classification and are thus independent of the chosen classifier. Wrappers utilize an interested classifier to evaluate feature subsets according to their predictive power, such as classification accuracy. Embedded methods perform feature selection in the process of training and are usually specific to a given classifier [10]. This paper focuses on filters with the classifier independent and time-saving properties. Sequential search and exhaustive search are two searching methods for filters. The former usually evaluates features one by one according to a certain criterion mentioned above. The sequential strategy has a drawback as Guyon et al. reported that a completely useless feature can greatly improve the performance when being used together with others [10]. Exhaustive search can avoid such a problem because it looks

978-1-61284-925-6/11/$26.00 ©2011 IEEE

at features as groups rather than individuals. This, however, may be time consuming and cause over-fitting.

In 2010 a GRBF parameter sigma selection method is proposed by Li [11]. It employs $\cos\theta$ (cosine similarity) as a similarity measure of samples, where θ is the angle of each pair vectors in the GRBF kernel space. Li's method uses the gradient descent method, a one-dimensional optimization algorithm, to find the optimal sigma by minimizing his objective function. The method is efficient to improve the performance of GRBF-based classifiers, such as SVM. Yet Li's method deals with the complete original feature space and may suffer if it contains some irrelevant and redundant features, especially in a high-dimensional feature space. That is, the optimal sigma may be trapped in a local minimizer. Considering this issue, Li's method is extended to a flexible feature space so that a feature subset and sigma can be optimized at the same time. The proposed method takes $\cos\theta$ as a similarity measure of samples as well and introduces a variable related to feature subsets into the objective function that considers both within-class and between-class similarities. Eventually, the problem becomes an optimization problem of two variables. An optimization algorithms, e.g. genetic algorithm (GA), particle swarm optimization (PSO), Nelder-Mead simplex method and simulated annealing (SA) [12], can be used to solve the optimization problem.

The remaining paper is organized as follows. Section II provides a brief introduction of the GRBF kernel function and its properties. Our method of feature selection is then presented. Section III conducts experiments and result analysis based on three benchmark datasets. The details of how to implement experiments and experimental results are provided in this section. Section IV includes the summary and conclusion of this paper.

II. The Proposed Method

A. Gaussian Radial Basis Function

Gaussian radial basis function (GRBF), also called Gaussian kernel function, is one of the most commonly used kernel functions, which can be expressed as

$$\kappa(\mathbf{x},\mathbf{z},\sigma)=\exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right), \quad (1)$$

where $\mathbf{x},\mathbf{z}\in\mathbb{R}^n$ are two samples that are n-dimensional column vectors in the original feature space, $\|\cdot\|$ denotes the Euclidean distance between two samples, $\sigma\in(0,\infty)$ is called the width of features. The GRBF kernel function has the following properties [11] [13].

a. The norm of each sample in the kernel space is one: $\kappa(\mathbf{x},\mathbf{x},\sigma)=1$. That is to say all samples stand on a circle if the kernel space is two-dimensional and on a spherical surface if the kernel space is three-dimensional.

b. The cosine value of the angle between two samples in the kernel space is equal to their kernel function value: $\cos\theta=\kappa(\mathbf{x},\mathbf{z},\sigma)$, where θ is the angle between two samples and has a range of $0\le\theta\le\pi$. Because the kernel function values of GRBF has a range of $0<\kappa(\mathbf{x},\mathbf{z},\sigma)\le 1$, It is obvious that $0<\cos\theta\le 1$, and then $0\le\theta<\pi/2$.

The property of $0\le\theta<\pi/2$ always holds for any two samples in the kernel space, and it implies that the angle between any two samples is in a range of $\pi/2$ rad. For example, all samples in the kernel space should be in a quarter of a circle if the kernel space is two-dimensional as shown in Fig. 1(i) and an eighth of a spherical surface if the kernel space is three-dimensional as shown in Fig. 1(ii). Based on the property (a), it is reasonable to measure the similarity of two samples by their angles. If the angle is smaller, these two samples are more alike, e.g. A1 and A2 in Fig. 1(i); otherwise, they are not as similar to each other, e.g. A1 and C1 in Fig. 1(i). Because $\cos\theta$ is monotonically decreasing with respect to θ when $0\le\theta<\pi/2$, $\cos\theta$ can be taken to measure the similarity of samples, which is equal to $\kappa(\mathbf{x},\mathbf{z},\sigma)$. That is to say, $\kappa(\mathbf{x},\mathbf{z},\sigma)$ becomes large when two samples are close and small when they are far away.

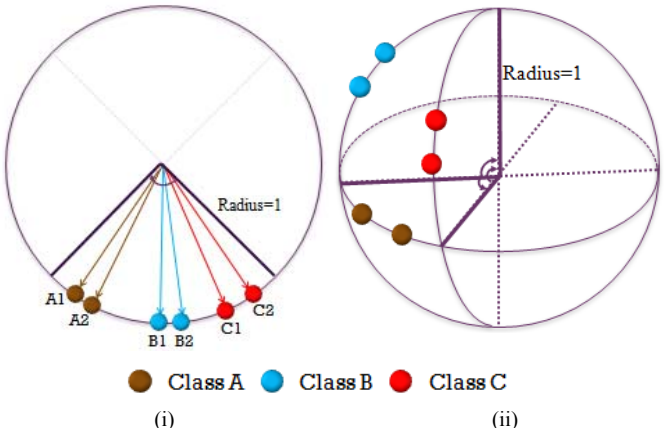

Figure 1. Schematic diagram of samples in the kernel space: (i), two-dimensional GRBF kernel space; (ii), three-dimensional GRBF kernel space

B. GRBF-based Feature Selection Algorithm

When selecting a proper feature subset, two principles have to be considered for the proposed GRBF-based feature selection algorithm: (i) Samples from the same class have large GRBF values; (ii) Samples from different classes have small GRBF values. We firstly define within-class cosine C_w and between-class cosine C_b as

$$C_w(\mathbf{s},\sigma)=\frac{1}{\sum_{i=1}^{L} l_i^2}\sum_{i=1}^{L}\sum_{\mathbf{x}\in i}\sum_{\mathbf{z}\in i}\kappa(\mathbf{x}\cdot\mathbf{s},\mathbf{z}\cdot\mathbf{s},\sigma), \quad (2)$$

$$C_b(\mathbf{s},\sigma)=\frac{1}{\sum_{i=1}^{L}\sum_{j=1,j\neq i}^{L} l_i l_j}\sum_{i=1}^{L}\sum_{\substack{j=1\\ j\neq i}}^{L}\sum_{\mathbf{x}\in i}\sum_{\mathbf{z}\in j}\kappa(\mathbf{x}\cdot\mathbf{s},\mathbf{z}\cdot\mathbf{s},\sigma), \quad (3)$$

where L is the number of classes; l_i the number of samples in class i; i and j are indices of classes; $\mathbf{s}\in\{0,1\}^n$ is an n-dimensional column vector, and $\mathbf{x}\cdot\mathbf{s}$ is the dot product between

x and **s**; by the dot product, the binary vector **s** can enable or disable a feature by setting the corresponding value to one or zero, respectively.

C_w, $0 < C_w \le 1$, is the average cosine value of samples in the same class, and C_b, $0 < C_b \le 1$, is the average cosine value of samples in different classes. Both C_w and C_b are scalar. Ideally, C_w=1 because all samples in the same class are totally overlapped, and $C_b \to 0$ because all samples from different classes tend to be orthogonal with each other. According to the above two principles, the objective function is defined as

$$J(\mathbf{s},\sigma) = \boldsymbol{\omega}^{\mathrm{T}}\mathbf{C} = \begin{bmatrix}\omega_w & \omega_b\end{bmatrix}\begin{bmatrix}1\text{-}C_w \\ C_b\end{bmatrix}, \qquad (4)$$

where $(\cdot)^{\mathrm{T}}$ is the transpose of a vector or a matrix, $\boldsymbol{\omega}$ is a two-dimensional weighting vector for C_w and C_b, and $\omega_w + \omega_b = 1$. The feature selection problem becomes a constrained optimization problem as

$$[\mathbf{s}^*, \sigma^*] = \underset{\mathbf{s}\in\{0,1\}^n, \sigma\in(0,\infty)}{\arg\min} \; J(\mathbf{s},\sigma). \qquad (5)$$

We use genetic algorithm to solve the problem, and then the optimal $\mathbf{s}^*$ and σ^* can be determined. For example, suppose $\mathbf{s}^*$ = [1, 0, 1, 0, 0]$^{\mathrm{T}}$ and σ^*=1, it means Feature 1 (F1) and Feature 3 (F3) are selected to form the new feature subset. The subset of {F1, F3} and σ^* are then used for GRBF-based classifiers as shown in Fig. 2.

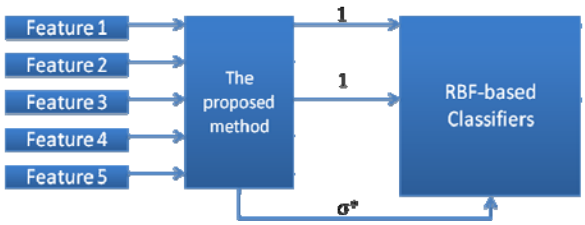

Figure 2. An example of how to use the proposed method for classification

III. Experiments and Results

When doing experiments, we choose SVM as a classifier to evaluate the performance of the proposed method. SVM is implemented by *svmtrain* and *svmclassify* from the Bioinformatics Toolbox, and GA is implemented by *ga* from the Global Optimization Toolbox of MATLAB. In the following, the proposed method (method 1) is compared with two other relevant methods: SVM with sigma selection (method 2) and the traditional SVM (method 3). Classification accuracy is used as the performance measure of these methods. It is defined as $N_c/(N_c+N_f)\times 100\%$, where N_c is the number of samples that are correctly classified, and N_f is the number of those falsely classified. The GRBF is used as the kernel function in SVM in these three methods, and default values for other parameters are used. There is no feature selection in method 2 or method 3, so they use all features in the original feature space. We take three benchmark datasets, the Parkinson, Ionosphere and Sonar datasets from the University of California Irvine (UCI) repository [1] to test the three methods. Their profiles are summarized in TABLE I.

TABLE I. Summary of Benchmark Datasets

No.	Dataset	Classes	Instances	Features
1	Parkinson	2	195	22
2	Ionosphere	2	351	34
3	Sonar	2	208	60

Normalization is firstly conducted to prevent features in greater numeric ranges dominating those in smaller numeric ranges. K-fold cross-validation is then employed for data partition (K=3 in the experiment). In each run out of five, our programs are executed K times, and the results are averaged over these K runs. The same procedure is repeated with all the three datasets. The obtained feature subsets and the sigma values with these three methods are summarized in TABLE II, and classification accuracy is provided in Fig. 3-5.

From the results presented, the proposed method uses fewer features and has higher classification accuracy than the other two methods. In the Parkinson dataset, our method selects five of 22 features. It is more accurate to use only 23% features of the original space. Dimensions are reduced to 50% and 45% in the Ionosphere dataset and the Sonar dataset, respectively. Computational time of test is thus saved with a reduced feature space, and only important features are retained in the optimal feature subset. Sigma selection of GRBF greatly affects the performance of SVM in the Sonar dataset since the traditional SVM has quite low classification accuracy in Fig. 5. It indicates that a proper sigma is critical to the performance of SVM. Method 2 shows its effectiveness in Fig. 3-5 in comparison with Method 3, as expected. Performance is further improved by using the proposed method over that of method 2.

TABLE II. Summary of Experiment Results

Dataset	Method 1		Method 2			Method 3	
	σ^*	$\mathbf{s}^*$	$\boldsymbol{\omega}$	σ^*	**s**	σ	**s**
Parkinson	0.6545	{F1,F3,F12,F18,F19}	[0.4 0.6]$^{\mathrm{T}}$	3.6125	{F1-F22}	1	{F1-F22}
Ionosphere	2.1635	{F1,F3,F5-F9,F11,F14,F15, F20-F25,F28,F31}	[0.4 0.6]$^{\mathrm{T}}$	3.9915	{F1-F34}	1	{F1-F34}
Sonar	1.8212	{F8,F10-F16,F19,F22,F23, F26,F27, F32,F34-F37,F46-F51,F55,F59,F60}	[0.4 0.6]$^{\mathrm{T}}$	5.8804	{F1-F60}	1	{F1-F60}

Note: F# is the feature # in the corresponding dataset; star (*) denotes the optimal value.

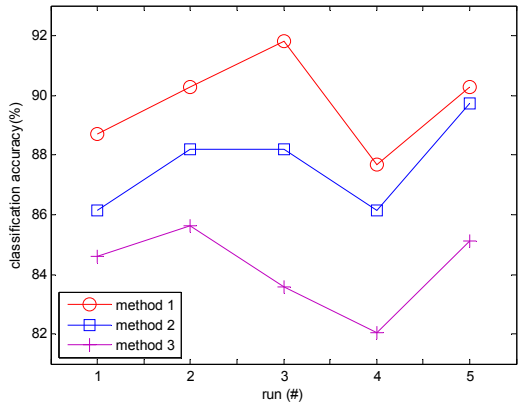

Figure 3. Classification accuracy of the Parkinson dataset

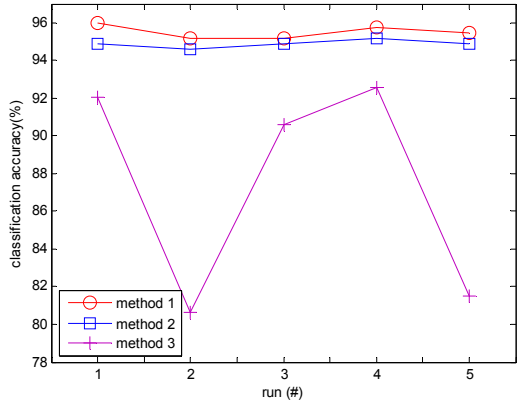

Figure 4. Classification accuracy of the Ionosphere dataset

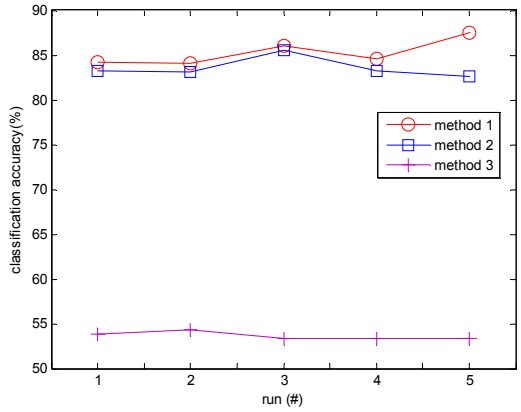

Figure 5. Classification accuracy of the Sonar dataset

IV. CONCLUSIONS

In this paper we proposed a GRBF-based feature selection algorithm that can tackle feature selection and parameter selection of the GRBF kernel function simultaneously. By utilizing the properties of GRBF, the proposed method employs cosθ as a measure of similarity of classes, where θ is the angle of each pair vectors in the GRBF kernel space. The objective function considering both within-class and between-class is minimized to obtain an optimal combination of a feature subset and sigma. Three benchmark datasets from the UCI repository are used to evaluate the proposed method (method 1), Li's method (method 2), and the traditional SVM (method 3). Results demonstrate that the proposed method possesses higher classification accuracy with a reduced feature subset. The proposed method could be applied to measurement systems to determine most significant variables for condition monitoring. Cost of measurement can be reduced with a reduced feature subset. In the future, we will validate generalization of the proposed method with other GRBF-based classifiers.

ACKNOWLEDGMENT

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC).

REFERENCES

[1] P.-H. Hsu, "Feature extraction of hyperspectral images using wavelet and matching pursuit", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 2, pp. 78-92, June 2007.

[2] P. Langley, *Elements of Machine Learining*, Morgan Kaufmann, 1996.

[3] N. Wyse, R. Dubes, and A. K. Jain, "A critical evaluation of intrinsic dimensionality algorithms," *Pattern Recognition in Practice*, pp. 514-524. Morgan Kaufmann Pulishers, Inc., 1980.

[4] B.-C. Kuo, and D. A. Landgrebe., "Nonparametric weighted feature extraction for classification," *Geoscience and Remote Sensing, IEEE Transactions on* , vol.42, no.5, pp. 1096- 1105, May 2004

[5] C.-H. Li, B.-C. Kuo, and C.-T. Lin, "LDA-based clustering algorithm and its application to an unsupervised feature extraction," *Fuzzy Systems, IEEE Transactions on* , vol.19, no.1, pp.152-163, Feb. 2011.

[6] M. Dash, and H. Liu, "Feature selection methods for classification", *Interlligent Data Analysis: An International Journal*, vol. 1, no. 3, 1997.

[7] X. Zhao, M. J. Zuo, and T. Patel, "EMD, ranking mutual information and PCA based condition monitoring", *ASME 2010 International Design Engineering Technical Conferences*, Montreal, Canada, Aug. 15-18, 2010.

[8] P. Yang, B. Zhou, Z. Zhang, and A. Y Zomaya, "A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data," *BMC Bioinformatics 11 (suppl. 1)*, S5, 2010.

[9] J. Qu, Z. Liu, M. J. Zuo, and H.-Z. Huang, "Feature selection for damage degree classification of planetary gearboxes using support vector machine," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*. accepted for publication.

[10] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol 3, pp 1157-1182, March 2003.

[11] C.-H. Li; C.-T. Lin; B.-C. Kuo; and H.-S. Chu, "An automatic method for selecting the parameter of the RBF kernel function to support vector machines," *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, pp.836-839, July 25-30, 2010.

[12] E. K. P. Chong, and S. H. Żak, "An introduction to optimization (3rd edition)," John Wiley & Sons Inc., Hoboken, New Jersey, 2008.

[13] S. T. John, and C. Nello, "Kernel Methods for Pattern Analysis," Cambridge, U.K.: Cambridge Univ. Press, 2004.

[14] A. Frank, and A. Asuncion. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.