A Set of new kernel function for Support vector machines: An approach based on Chebyshev polynomials

Sara Zafar Jafarzadeh
Center of Excellence on Soft Computing and Intelligent Information Processing, Department of Computer Engineering,
Ferdowsi University of Mashhad,
Mashhad, Iran
sarazafarjafarzadeh@um.ac.ir

Mohammad Aminian
Department of Computer Engineering,
ShahidChamran University of Ahvaz, Iran
m.aminiyan@gmail.com

SohrabEfati
Department of Mathematics,
Ferdowsi University of Mashhad,
Mashhad, Iran
s-effati@um.ac.ir

Abstract—**In this paper, we introduce a set of new kernel functions Which is derived by combining generalized Chebyshev polynomials with other standard kernel functions. New kernel functions have significant advantages over classic support Vector Machine's (SVM) kernel functions and Chebyshev kernel. Simulation results illustrate the fact that the new set of kernel functions (in particular Chebyshev-Gaussian kernel) has noticeable improvement in decreasing error rate and support vector numbers.**

Keywords-SVM; mixing kernel; classification; Linear and non-linear modeling; orthogonal polynomials

I. INTRODUCTION

SVM is applied to solve both classification and regression problems. Its theory is based on structural risk minimization by maximizing the margin. SVM generalization performance completely depends on two major steps. The first step is, to define the constraint satisfaction problem and then solving it. The second step is, to choose the best kernel function which maps input data to higher dimensional feature space where the data can be discriminate linearly. Choosing optimal kernel function is one of the major tasks in building a SVM. A kernel function must be checked from several viewpoints such as ability to map input data to the best feature space where SVM can classify them with minimum error rate and lowest number of support vectors(SV), providing an acceptable performance The "Gaussian kernel function" requires only one parameter. Another major task is to find optimal parameters which are solved by different approaches. In the proposed method, a try and error approach is used to find optimal parameters.

In recent years, many kernel functions have been introduced for example [1-3]. In [4], a new flexible kernel function is proposed which is a proper generic alternative to the common linear, polynomial and RBF kernels. A new way to produce unlimited number of nonparametric and efficient kernels is introduced in [5].

Chebyshev kernel is proposed as a standard kernel for scalar values. It is improved by [6] in order to be applicable in multidimensional applications. Since a kernel function provides a measure of similarity between two vectors and it is defined as the inner product of two given vectors in the higher dimensional space for SVM, obviously finding a kernel function which reduced error and number of SVs is a key factor to enhance the performance. Therefore in this paper, a set of new kernel functions which is a combination of Chebyshev polynomials with some of classic kernel functions is proposed. The rest of this paper is organized as follows. In the second section we review SVM. Related works are presented in the third section. In the fourth section, validity is presented. Our proposed method is defined in the fifth section. The comparison of experimental results is provided in the seventh section. Finally, the conclusion is discussed in section eight.

II. SUPPORT VECTOR MACHINE

The fundamental of SVM can be track back to statistical learning theory[6]. However, current SVM is a deterministic supervised learning algorithm, rather than being a statistical learning method. There are several different models based on the SVM's cost function, see [6].The SVM is based on the idea of inserting a hyperplane between two (binary) classes. Inserting a hyperplane can be done in either the current data space (linear SVM), or the higher dimensional space by using the kernel functions (nonlinear SVM). The label of test data is gained from the following formula [6,7]:

$$f(\boldsymbol{x}) = sgn\left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right) \quad (1)$$

Where$\alpha_i$is the nonzero Lagrange coefficient of the associated support vectorx_i, n is the number of support vectors, f(x) is the class label of the given test data x and K(.) is the kernel function. The class labels y_i associated to the SV x_i, is a binary value, i.e., y_i A{− 1, +1}, and b is the bias value. These values are calculated by maximizing the following function:

$$w(\alpha) = \sum\nolimits_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j\, K\left(\boldsymbol{x_i}, \boldsymbol{x}_j\right)$$

$$\text{S.T}: \sum_{i=1}^{l} \alpha_i y_i = 0\ , \qquad 0 \le \alpha_i \le C \quad (2)$$

Where $\alpha_i$is the nonzero Lagrange coefficient of the associated support vectorx_i, n is the number of training samples and K(.) is the kernel function.

III. RELATED WORKS

An appropriate kernel function maps input vectors to high dimensional feature space where all input data can be linearly separated. Therefore, the inner product of each given pair of transformed vectors in the higher dimensional space can be reached by applying the kernel function onto the input vectors directly without the need of a transformation function$\phi(.)$as

$$K(\boldsymbol{x},\boldsymbol{z}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{z}) \rangle \tag{3}$$

Where K(.) is the kernel function. Some of the common kernel functions are mentioned below.

Gaussian kernel [8], [6]:

$$K(\boldsymbol{x},\boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\right) \tag{4}$$

where σ is the kernel parameter.

Wavelet kernel [9]:

$$K(\boldsymbol{x},\boldsymbol{z}) = \prod_{i=1}^{m} \cos\left(1.75\frac{x_i - z_i}{a}\right) \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2a^2}\right) \tag{5}$$

where a is the kernel parameter and m is dimensional of input vectors.

Polynomial kernel [8]:

$$K(\boldsymbol{x},\boldsymbol{z}) = \left(\frac{\langle x, z\rangle + 1}{\beta}\right)^n \tag{6}$$

Where β is the scaling parameter and n is the kernel parameter.

Chebyshevkernel[10]:

$$K(\boldsymbol{x},\boldsymbol{z}) = \prod_{j=1}^{m} \frac{\sum_{i=0}^{n} T_i(\boldsymbol{x_j}) T_i(\boldsymbol{z_j})}{\sqrt{1 - \boldsymbol{x_j z_j}}} \tag{7}$$

Where n is the kernel parameter and m is the dimension of input vectors. Also$\sqrt{1-\boldsymbol{x_j z_j}}$is the weighted function and $T_i(.)$ represents Chebyshev polynomials which is calculated by equation below:

$$\begin{aligned} &T_0(x) = 1 \\ &T_1(x) = x \\ &T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \text{ for k=2,3,4,...} \end{aligned} \tag{8}$$

Generalized Chebyshevkernel[10]:

$$K(\boldsymbol{x},\boldsymbol{z}) = \frac{\sum_{i=0}^{n} T_i(\boldsymbol{x}) T_i^T(\boldsymbol{z})}{\sqrt{a - \boldsymbol{xz}}} \tag{9}$$

Where a=m, x and z are m-dimensional vectors, n is the kernel parameter. Also$\sqrt{a-\boldsymbol{x_j z_j}}$ is the weighted function and its value is always positive and finally$T_i(.)$represents generalized Chebyshev polynomials which is calculated by equation (10)[10]:

$$\begin{aligned} &T_0(x) = 1 \\ &T_1(x) = x \\ &T_{k+1}(x) = 2xT_{k-1}^T(x) - T_{k-2}(x) \\ &\text{for k=2,3,4,....} \end{aligned} \tag{10}$$

IV. VALIDITY

A valid kernel should satisfy the Mercer Conditions [8],[7]. If the kernel does not satisfy them, SVM may not find the optimal parameters, but it is still possible to find suboptimal parameters. Besides, if the Mercer conditions are not satisfied, then the Hessian matrix for the optimization part may not be positive definite.

Mercer Theorem: To be a valid SVM kernel, for any finite function g(x), the following integral should always be non- negative for the given kernel function K(x,z) [8]:

$$\begin{aligned} &K(\boldsymbol{x},\boldsymbol{x}') = \sum_{m}^{x} a_m\, \phi_m(\boldsymbol{x})\phi_m(\boldsymbol{x}') \\ &\iint K(\boldsymbol{x},\boldsymbol{x}')g(\boldsymbol{x})g(\boldsymbol{x}')d\boldsymbol{x}dx' \end{aligned} \tag{11}$$

V. PROPOSED METHOD

Based on Mercer theorem, having two valid kernel functions, a new kernel function can be made by summation and multiplication of these two kernel functions. In our proposed method, this idea is applied to produce new kernel functions.

A. *Chebyshev-Gaussian kernel*

Based on the above idea, first we try to build a kernel by mixing Gaussian kernel with generalized Chebyshev kernel. As mentioned before, we assume that multiplying two functions will give us features and advantages of both functions almost in the same strength. So we define first kernel as (12):

$$K(\boldsymbol{x},\boldsymbol{z}) = \frac{\sum_{\mathrm{i=0}}^{\mathrm{n}} \mathrm{T_i}(\mathbf{x})\mathrm{T_i^T}(\mathbf{z})}{\sqrt{\mathrm{a} - \mathbf{xz}}} \times \sum_{\mathrm{i=1}}^{\mathrm{m}} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\right) \tag{12}$$

Where n is the generalized Chebyshev kernel parameter, σ is the Gaussian parameter and m is the dimension of input vectors' space.

B. *Chebyshev-wavelet kernel*

The second kernel is the a product of wavelet kernel (5) and generalized Chebyshev kernel (9) We choose wavelet kernel as second multiple since it has better performance compare with other available test data-sets' kernel functions. This kernel is defined as (13):

$$\begin{aligned} &K(\boldsymbol{x},\boldsymbol{z}) \\ &= \frac{\sum_{\mathrm{i=0}}^{\mathrm{n}} \mathrm{T_i}(\mathbf{x})\mathrm{T_i^T}(\mathbf{z})}{\sqrt{\mathrm{a} - \mathbf{xz}}} \\ &\times \prod_{i=1}^{m}\left(\cos\left(1.75\frac{x_j - z_j}{a}\right)\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2a^2})\right) \end{aligned} \tag{13}$$

Where n is the generalized Chebyshev kernel parameter, a is the wavelet kernel parameter and m is the dimension of input vector.

VI. DATA NORMALIZATION

Chebyshev polynomials are orthogonal only within region [-1,1]. Because of this, all input data has to be normalized in this region. Since wavelet and Gaussian kernels are also orthogonal in the region of [-1,1], normalization will not affect the result of these two kernels.

It is necessary to normalize input data to avoid the difference between scale of each record feature value. This problem affects on the results specially in methods like SVM which are classified in discriminative classifiers category [11]. Normalization formula that is used in this paper is defined as below:

$$\mathrm{x}^{\mathrm{new}} = \frac{2(\mathrm{x} - \mathrm{Min})}{\mathrm{Max} - \mathrm{Min}} - 1 \qquad (14)$$

VII. EXPERIMENTAL RESULT

A. validation

There are several standard validation methods. In this paper, we use K-fold cross validation. The value of K parameters which are used for all of experiments is 10(10-fold cross validation). In order to have more accurate results we first apply this method on each individual dataset and save the index of split points, consequently we could use exactly the same structure for all of our tests.

Average error E is used as a performance parameter. To compare kernels, the SV number is taken into account. The E parameter is defined in equation (15):

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i \qquad (15)$$
$$E_i = \frac{n_i}{t_i}$$

Where $\mathrm{n_i}$ is the number of correct classified test data in the i-th test and t_i indicates the total number of test data.

B. simulation and results

In the simulation, we use seven different standard datasets from UCI repository. Moreover, we test our new methods and three other different SVMs that have only different kernels, with different kernel parameters and then we report the best kernel parameter in first place with respect of E and in second place SV numbers.

Breast cancer Wisconsin dataset: This dataset is currently available at*http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29* known as breast cancer Wisconsin dataset. This dataset has 569 data where each data vector has 30 features in form of real values. The dataset has only two classes which are malignant and benign. Table 1 illustrates that SVM with Chebyshev-wavelet kernel has the minimum error rate of 1.71. But from SV number viewpoint Chebyshev-Gaussian kernel has the minimum SV number. The experimental results shows that Chebyshev-Gaussian is the most appropriate kernel for this dataset.

Table 1.
breast cancer Wisconsin dataset test results with various kernel functions.

Kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 1$, $\sigma = 2$	0.4	1.93	73
Chebyshev-wavelet	$n = 1, a = 4$	2	1.71	131
Chebyshev	$n = 2$	2	2.63	88
wavelet	$a = 16$	2	2.1	196
Gaussian	$\sigma = 8$	2	2.1	197

Heart disease dataset: This dataset is currently available at*http://archive.ics.uci.edu/ml/datasets/Heart+Disease* known as heart disease dataset. This dataset has 303 data where each data vector has 75 features in form of integer, real and categorical values. The dataset has four classes. As we can see in table 2, both Chebyshev-Gaussian and Chebyshev-wavelet kernel have same results and their results are the same in error rate but their SV number is less than the other kernels.

Table 2.
Heart disease dataset test results with various kernel functions.

kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 1$, $\sigma = 0.1$	1	19.1	114
Chebyshev-wavelet	$n = 1, a = 2$	0.8	19.1	114
Generalized Chebyshev	$n = 1$	2	19.1	157
wavelet	$a = 8$	1	19.1	186
Gaussian	$\sigma = 8$	2	19.4	167

Liver disorder dataset: This dataset is currently available at *http://archive.ics.uci.edu/ml/datasets/Liver+Disorders* known as liver disorder dataset. This dataset has 345 data where each data vector has 7 features in form of integer, real and categorical values. The dataset has only two classes. For this dataset, Gaussian kernel has minimum number of SVs and minimum error rate. Table 3 shows the result of this dataset.

Table 3.
Liver disorder dataset test results with various kernel functions.

kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 1$, $\sigma = 8$	2	22.6	291
Chebyshev-wavelet	$n = 2, a = 2$	2	25.8	305
Generalized Chebyshev	n =1	2	26.4	305
Wavelet	$a = 8$	1	27.2	301

Gaussian	σ =2	0.8	29.3	294

Diabetes dataset: This dataset is currently available at*http://archive.ics.uci.edu/ml/datasets/Diabetes* known as diabetes dataset. This dataset has 768 data where each data vector has 8 features form of real values. The dataset has two classes. Table 4 demonstrates the result of this dataset. As we can see all the kernels almost have the same results, but Chebyshev-Gaussian kernel has better results.

Table 4.
Diabetes dataset test results with various kernel functions.

Kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 1$, $\sigma = 16$	4	27.1	621
Chebyshev-wavelet	$n = 1, a = 32$	4	27.2	621
Generalized Chebyshev	$n = 1$	4	28	631
Wavelet	$a = 8$	2	27.7	646
Gaussian	$\sigma = 2$	2	27.7	632

Haberman's survival dataset: This dataset is currently available at *http://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival*known as Haberman'ssurvival. This dataset has 306 data where each data vector has 3 features form of integer values. The dataset has two classes. Table 5 illustrate that generalized Chebyshev kernel has the best performance in compare to the other kernels.

Table 5.
Haberman's survival dataset test results with various kernel functions.

Kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 1$, $\sigma = 2$	1	25.5098	272
Chebyshev-wavelet	$n = 1, a = 2$	2	25.4	270
Generalized Chebyshev	$n = 1$	0.8	24.8	178
Wavelet	$a = 1$	0.8	25.5	270
Gaussian	$\sigma = 8$	0.8	24.5	276

Ionosphere dataset: This is another dataset currently available at http://archive.ics.uci.edu/ml/datasets/Ionosphere known as Ionosphere dataset. This dataset has 351 data where each data vector has 34 features form of real and integer values. The dataset has two classes. Both this dataset and the next one, in contrast with first five datasets, do not belong with diagnosis area. Table 6 shows that Chebyshev-Gaussian kernel has the best performance. So in this case Chebyshev-Gaussian kernel is the best kernel.

Table 6.
Ionosphere dataset test results with various kernel functions.

kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 1$, $\sigma = 1$	1	3.1	103
Chebyshev-wavelet	$n = 1, a = 2$	1	3.4	227
Generalized Chebyshev	$n = 2$	0.8	8.5	113
wavelet	a =8	2	4.8	222
Gaussian	$\sigma = 1$	2	5.4	244

Connectionist bench (Sonar, Mines vs. Rocks) dataset: This dataset is currently available at http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks) known as Sonar dataset. This dataset has 208 data where each data vector has 60 features form of real values. The data has two classes. Table 7 illustrates the results of 7th dataset. Crystal clearly in our last result Chebyshev-Gaussian kernel also shows the best performance.

Table 7.
Sonar dataset test results with various kernel functions.

Kernel	Kernel parameter	C(penalty) parameter	E (error rate %)	SV number
Chebyshev-Gaussian	$n = 2$, $\sigma = 32$	0.4	10.46	74
Chebyshev-wavelet	$n = 2, a = 4$	0.1	11.5	98
Generalized Chebyshev	$n = 4$	0.1	15.3	80
wavelet	$a = 8$	1	15.4	181
Gaussian	$\sigma = 8$	1	15.8	174

VIII. CONCLUSION AND FUTURE WORK

In this paper we introduce two new kernel functions. Although these new kernels show almost better performance in comparison with some of the other kernels, in our simulation phase we cannot neglect the cost of adding new parameter to kernel functions. So there are some issues for this new method. First, building the new kernel function by combining other individual kernels with the aim of finding more accurate kernels. Second, we have to solve the complexity of execution phase by analysis the kernels with different values of their parameters to reach a set of optimal values.

REFERENCES

[1] J. Zhao, G. Yan, B. Feng, W. Mao, and J. Bai, "An adaptive support vector regression based on a new sequence of unified orthogonal polynomials," *Pattern Recognition*, 2012.

[2] M. Ha, C. Wang, and J. Chen, "The support vector machine based on intuitionistic fuzzy number and kernel function," *Soft Computing,* pp. 1-7, 2013.

[3] T. Ibrikci, D. Ustun, and I. E. Kaya, "Diagnosis of Several Diseases by Using Combined Kernels with Support Vector Machine," *Journal of medical systems,* vol. 36, pp. 1831-1840, 2012.

[4] R. Zhang and W. Wang, "Facilitating the applications of support vector machine by using a new kernel," *Expert Systems with Applications,* vol. 38, pp. 14225-14230, 2011.

[5] E. A. Daoud and H. Turabieh, "New empirical nonparametric kernels for support vector machine classification," *Applied Soft Computing,* 2013.

[6] N. Ye, R. Sun, Y. Liu, and L. Cao, "Support vector machine with orthogonal Chebyshev kernel," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, pp. 752-755.

[7] B. Schölkopf and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond (Adaptive computation and machine learning)," 2001.

[8] V. N. Vapnik, "Statistical learning theory," 1998.

[9] L. Zhang, W. Zhou, and L. Jiao, "Wavelet support vector machine," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 34, pp. 34-39, 2004.

[10] S. Ozer, C. H. Chen, and H. A. Cirpan, "A set of new Chebyshev kernel functions for support vector machine pattern classification," *Pattern Recognition,* vol. 44, pp. 1435-1447, 2011.

[11] K. Chan, T.-W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, and T. J. Sejnowski, "Comparison of machine learning and traditional classifiers in glaucoma diagnosis," *Biomedical Engineering, IEEE Transactions on,* vol. 49, pp. 963-974, 2002.