A New Mercer Sigmoid Kernel for Clinical Data Classification

André M. Carrington, Paul W. Fieguth, Helen H. Chen

Abstract— In classification with Support Vector Machines, only Mercer kernels, i.e. valid kernels, such as the Gaussian RBF kernel, are widely accepted and thus suitable for clinical data. Practitioners would also like to use the sigmoid kernel, a non-Mercer kernel, but its range of validity is difficult to determine, and even within range its validity is in dispute. Despite these shortcomings the sigmoid kernel is used by some, and two kernels in the literature attempt to emulate and improve upon it.

We propose the first Mercer sigmoid kernel, that is therefore trustworthy for the classification of clinical data. We show the similarity between the Mercer sigmoid kernel and the sigmoid kernel and, in the process, identify a normalization technique that improves the classification accuracy of the latter.

The Mercer sigmoid kernel achieves the best mean accuracy on three clinical data sets, detecting melanoma in skin lesions better than the most popular kernels; while with non-clinical data sets it has no significant difference in median accuracy as compared with the Gaussian RBF kernel. It consistently classifies some points correctly that the Gaussian RBF kernel does not and vice versa.

Index Terms— Clinical Signal Processing, Clinical Data Classification, Support Vector Machine, Sigmoid Kernel, Mercer Kernel

I. INTRODUCTION

The strong performance of Support Vector Machines (SVM) and kernel methods make them a mainstay as one of the state-of-the-art techniques for classification [1], [2], including applications to clinical research, diagnosis and prognosis [3], [4], [5]. One of the key issues in specifying an SVM solution is choosing the right kernel for the data and task, since a wrong choice can have a detrimental and possibly profound impact on classification accuracy [1], [2], [6].

The sigmoid kernel was once quite popular for use with SVMs [7] and it continues to be used in a clinical context as indicated by ScienceDirect and Google Scholar with 30 and 1,510 hits respectively (2011 through 2014). The interest in sigmoids or S-curves stems from their success in classification with neural networks and logistic regression; their specific properties of linearity, saturation and dichotomy; and their nature as the cumulative distribution of a Gaussian. However, the sigmoid kernel is problematic because it is difficult to choose parameters that ensure that it is conditionally positive definite (c.p.d) [7]. Some literature asserts that a c.p.d. kernel is valid [8], [9], while other literature omits c.p.d. from consideration [2], [10], [11].

A. M. Carrington and P. W. Fieguth are with the Department of Systems Design Engineering and H. H. Chen is with the School of Public Health and Health Systems at the University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 {amcarrin, pfieguth, helen.chen}@uwaterloo.ca
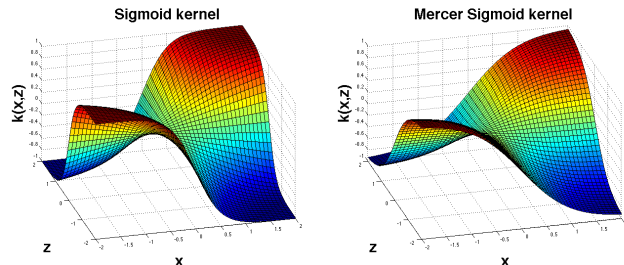

Fig. 1. The sigmoid and Mercer sigmoid kernels are similar in their output $k(\mathbf{x}, \mathbf{z})$ for one-dimensional inputs $\mathbf{x}$ and $\mathbf{z}$, for a range of parameter values (r small, $b = 0$, per (1), (8)). This similarity behaviour extends to many dimensions when the former is normalized.

Therefore, as a non-Mercer kernel, the sigmoid kernel is not necessarily a trustworthy choice for clinical applications.

As part of a clinical classification challenge related to skin lesions, we created a valid Mercer sigmoid (MSig) kernel, that is similar to a sigmoid kernel (Fig. 1) since it shares the same underlying sigmoid function.

In Section II we provide background and discuss related work. In Section III we define the MSig kernel and discuss its properties along with the sigmoid kernel and normalization. Section IV proves that the proposed kernel is a Mercer kernel, while Sections V and VI show the experiment and associated results. Finally, Section VII provides conclusions.

II. BACKGROUND AND RELATED WORK

A sigmoid (function) or S-curve is a class or family of functions that includes the logistic function, the hyperbolic tangent, the arctangent, the error function, the generalised logistic function, etc. Formally, a sigmoid function is a function that is defined for all real inputs, $x \in \mathbb{R}$; is bounded in its range or outputs, $f(x) \in (p, q)$ for finite $p, q \in \mathbb{R}$; and has a positive first derivative at all points [12].

Whereas a sigmoid is a function of one input, a kernel is a function of two inputs that may be used as a measure of similarity between the inputs. In SVM classification, for example, a kernel compares an unclassified sample of data with a support vector (a weighted sample that has been classified). Formally, a kernel is a function

$$k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

that for all of its inputs $\mathbf{x}, \mathbf{z} \in X$, has a mapping ϕ from X to an inner product (feature) space F [11]. Kernels commonly found in SVM literature include the Gaussian RBF kernel, the linear kernel, the polynomial kernel and the sigmoid kernel [1], [13].

The sigmoid kernel [2], [7] is based on the hyperbolic tangent:

$$k_S(\mathbf{x}, \mathbf{z}) = \tanh\left(a \cdot \mathbf{x}^T \mathbf{z} + r\right) \quad a > 0,\ r < 0 \quad (1)$$

$$= \tanh\left(a \cdot \sum_{i=1}^{p}\{x_i z_i\} + r\right) \quad (2)$$

with vectors $\mathbf{x}$ and $\mathbf{z}$ of explanatory variables to be compared by the kernel (e.g. two patients), with a horizontal scaling parameter a, and with a central vertical bias r that changes the height of the kernel's output for inputs near the origin.

An important characteristic of kernels is whether or not they are Mercer kernels or, equivalently, positive semi-definite, since this ensures that certain assumptions hold for equations such as the SVM soft-margin objective function in dual form (3) and its associated constraints (4):

$$W(\alpha) = -\sum_{i=1}^{\ell}\alpha_i + \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (3)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad i = 1, ..., \ell \quad (4)$$

$$\sum_{i=1}^{\ell}\alpha_i y_i = 0$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are vectors of explanatory variables, y_i and y_j are the corresponding outcomes, α_i and α_j are weight parameters in constrained optimization derived from Lagrange's method (which identify the support vectors), and C is the SVM soft-margin constant [13].

The sigmoid kernel is not positive semi-definite, but it does meet a weaker condition as a conditionally positive definite (c.p.d) kernel for $a > 0$ and $0 < r \leq \hat{r}$ for sufficiently small $\hat{r}$, dependent on the dataset [7], however the actual value of $\hat{r}$ is difficult to determine. It has been argued that solving the SVM objective function (3) with a c.p.d. kernel is equivalent to solving it with an associated positive semi-definite (p.s.d.) kernel and that c.p.d. kernels are valid for use with SVMs [8], [9] — but this does not resolve the issue of determining $\hat{r}$, and the prevailing literature omits c.p.d. from consideration [2], [10], [11].

Two kernels have been created by other authors [14], [15] to emulate the sigmoid kernel and mitigate its limitations, however neither of them is Mercer.

III. DESCRIPTION AND ANALYSIS

We begin with a kernel defined in inner-product form:

$$k(\mathbf{x}, \mathbf{z}) \triangleq \langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle \quad \mathbf{x}, \mathbf{z}, \Phi \in \mathbb{R}^p,\ k \in \mathbb{R} \quad (5)$$

$$\text{where } \Phi(\mathbf{x}) = \begin{bmatrix} \phi(x_1) & \phi(x_2) & ... & \phi(x_p) \end{bmatrix}^T \quad (6)$$

We can choose a hyperbolic form for ϕ,

$$\phi(x) = \frac{1}{\sqrt{p}}\tanh\left(\frac{x-d}{b}\right) \quad \phi, x, b, d \in \mathbb{R},\ p \in \mathbb{N} \quad (7)$$

with a horizontal scaling parameter b, a horizontal shift parameter d, and a normalization parameter p that represents the dimensionality of the input vectors $\mathbf{x}$ and $\mathbf{z}$. Leading to the proposed Mercer sigmoid (MSig) kernel,

$$k_M(\mathbf{x},\ \mathbf{z}) \triangleq \frac{1}{p}\sum_{i=1}^{p}\tanh\left(\frac{x_i - d}{b}\right) \cdot \tanh\left(\frac{z_i - d}{b}\right) \quad (8)$$

whose parameters have the same meaning as in (7). Normalization by p in (8) is applied for ease of interpretation and comparison.

A. Similarity

The most fundamental question, then, is the degree of similarity between the MSig and sigmoid kernels, to gain insight into the MSig behaviour and determine whether it can replace the function of the sigmoid kernel.

If we consider the sigmoid kernel (2) in one-dimension, with $a = 1,\ r = 0$, then

$$k_S(x, z) = \tanh(xz)$$

Similarly the MSig kernel (8) in one-dimension with $a = 1,\ b = 1,\ d = 0$ corresponds to

$$k_M(x, z) = \tanh(x)\tanh(z)$$

The normalized root mean squared deviation (NRMSD) between the two kernels k_S, k_M is 3.24% for $x, z \in (-1, +1)$; that is, the two kernels are arguably similar.

We can also compare the sigmoid kernel (2) with the MSig kernel (8) in general, provided that we use a similar horizontal scale $a \approx \frac{1}{b^2}$, and the same horizontal shift (i.e. let $d = 0$), and provided that the dot products in the two kernels are both normalized by dimensionality (or both not) for comparison.

The dot product in the MSig kernel is already normalized by $\frac{1}{p}$, but the dot product in the sigmoid kernel is not

$$k_S(\mathbf{x}, \mathbf{z}) = \tanh\left(a \cdot \mathbf{x}^T\mathbf{z} + r\right)$$

so we scale the inputs,

$$k_S\left(\frac{\mathbf{x}}{\sqrt{p}}, \frac{\mathbf{z}}{\sqrt{p}}\right) = \tanh\left(\frac{a}{p} \cdot \mathbf{x}^T\mathbf{z} + r\right)$$

to normalize the dot product by the same amount $\frac{1}{p}$ and call this a normalized sigmoid kernel (SigN).

This normalization does not just enable comparison, it should yield a better result because the values of input data should influence how the $\tanh$ function behaves, not the dimensionality of the input. Without normalization, the dot product as an input to the $\tanh$ function will grow as the input dimensionality of $\mathbf{x}$ and $\mathbf{z}$ grows, causing saturation in the $\tanh$ output increasingly because of dimensionality rather than the values of the input data. Our results confirm
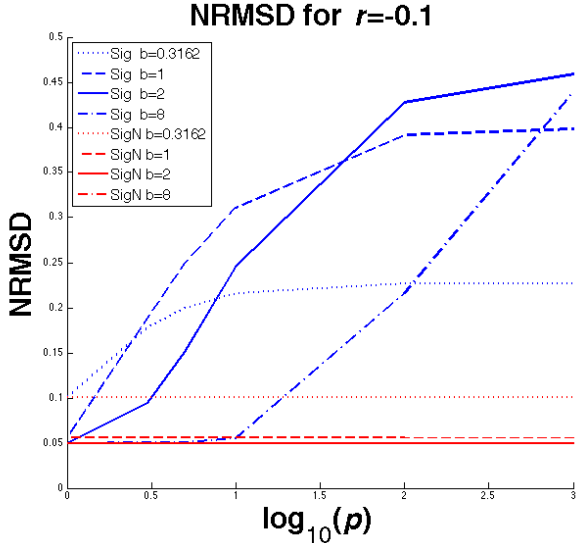
Fig. 2. The Normalized Root Mean Squared Deviation (NRMSD) between the sigmoid kernel (Sig) and the Mercer sigmoid kernel, and between the normalized sigmoid kernel (SigN) and the Mercer sigmoid kernel, for different values of b, and for $r = -0.1$, i.e. the value of r with the highest NRMSD.

that the SigN kernel has improved accuracy relative to the sigmoid kernel, on average and in four out of six data sets.

We then find that the MSig and SigN kernels are similar with NRMSD$\leq 10.073\%$ (Fig. 2) for sufficiently small r as required for the Sig and SigN kernels to be c.p.d. (we selected $-0.1 \leq r < 0$). Without normalization, the NRMSD between MSig and Sig increases with dimensionality, and it is higher than the normalized comparison for all $p > 1$ (Fig. 2).

Our normalization technique appears to be novel as we did not find it in the literature [1], [2], [16], [11]. It may be considered for any kernel of the form $k\left(\mathbf{x}^T \mathbf{z}\right)$, i.e. the class of dot-product kernels or zonal kernels, not just a sigmoid kernel.

B. Linearity and Dichotomy

A function or kernel saturates if it produces a bounded output range, for inputs that are unbounded [17]. The sigmoid and MSig kernels saturate with horizontal asymptotes $k = \pm 1$ at the outermost corner of each quadrant (or orthant), for one dimensional (or n-dimensional) inputs (Fig. 1). For inputs near the origin the sigmoid and MSig kernels act linearly, while other inputs are dichotomized to an output value of -1 or $+1$.

If the sigmoid or MSig kernel fits the data such that the region of saturation mitigates the effect of outliers or large values in SVM classification/optimization then the signal-to-noise ratio (SNR) of true data is improved. For this purpose, applying dichotomization (tanh) within each dimension, as in the MSig kernel (8), is preferred to applying it once overall, as in the sigmoid kernel (2).

Dichotomization within each dimension suits binary data and nominal data that are converted to binary data; and our

clinical data sets have heterogeneous data types that include binary and nominal data.

C. Covariance

Genton analyzed machine learning kernels from a statistics perspective and remarked that kernels are covariances [18], presumably because Mercer kernels must be a dot product (implicitly or explicitly) of basis functions in $\mathbf{x}$ and $\mathbf{z}$. We examine the sigmoid and MSig kernels from this perspective.

The dot product in the sigmoid kernel (2), but not the kernel itself, is a sum of covariances $x_i z_i$ for each dimension i of the *input space*; whereas for the MSig kernel, if we let

$$x_i' = \phi\left(x_i\right) \quad \text{from (7)}$$

then the kernel can be re-written as

$$k_M\left(\mathbf{x}, \mathbf{z}\right) = \sum_{i=1}^{p} x_i' z_i' \quad \text{from (7), (8)}$$

which is a sum of covariances $x_i' z_i'$, where x_i' and z_i' are the axes for each dimension i of the *feature space*. For every Mercer kernel there exist such feature space axes x_i' and z_i', implicitly or explicitly. Finally, we note that the two sums of covariances, are traces of the cross-covariance of $\mathbf{x}$ with $\mathbf{z}$, and $\mathbf{x}'$ with $\mathbf{z}'$, respectively.

IV. MERCER COMPLIANCE

From [19] and [18] a valid kernel, a kernel that is positive semidefinite and symmetric, and a Mercer kernel, are equivalent. To prove Mercer compliance, we therefore prove that the kernel is valid.

From [11]:
- $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$ is a valid kernel for real-valued $f\left(\cdot\right)$ on X, $X \subseteq \mathbb{R}^p$ (9), and
- $k\left(\mathbf{x}, \mathbf{z}\right) = k_1\left(\mathbf{x}, \mathbf{z}\right) + k_2\left(\mathbf{x}, \mathbf{z}\right)$ is a valid kernel if k_1 and k_2 are valid kernels (10).

Let $f\left(x_i\right) = \frac{1}{\sqrt{p}} \tanh\left(\frac{x_i - d}{b}\right)$, where all of the variables and parameters are real-valued except for $p \in \mathbb{N}$ (11).

From (9) and (11), $k\left(x_i, z_i\right) = \frac{1}{\sqrt{p}} \tanh\left(\frac{x_i - d}{b}\right) \cdot \frac{1}{\sqrt{p}} \tanh\left(\frac{z_i - d}{b}\right)$ is a valid kernel (12).

From (10) and (12), $k_M\left(\mathbf{x}, \mathbf{z}\right) = \sum_{i=1}^{p} k\left(x_i, z_i\right) = \sum_{i=1}^{p} \frac{1}{\sqrt{p}} \tanh\left(\frac{x_i - d}{b}\right) \cdot \frac{1}{\sqrt{p}} \tanh\left(\frac{z_i - d}{b}\right)$ is a valid kernel; which is our proposed kernel.

V. EXPERIMENTAL DATA AND METHOD

This paper is written in the context of melanoma research using a skin lesion data set that consists of sixty sequential cases from Dr. Eric Ehrsam's dermatology blog [20]. We also tested our proposed kernels with two other clinical data sets from the machine learning repository at the University of California at Irvine [21], the Statlog Heart data set and the Pima Indians Diabetes data set; and with three non-clinical data sets, the Mushrooms data set (using a subset

	Kernel						SVM	
	Poly	RBF	Sig		MSig			
Limit	d	$\log \sigma$	a	r	b	d	$\log C$	kkt
Lower	2	-1	ε	-5	$\frac{1}{\sqrt{a}}$	-2	-1	0
Upper	7	3	10	$-\varepsilon$		+2	3	1

TABLE II

DATA SET SUMMARY

	Instances		10-fold	Separate	Numeric	
Data Set	Training	Test	CV	Test	Only	Source
Skin Lesion	57	57	yes	no	no	[20]
Heart	270	270	yes	no	no	UCI
Diabetes	512	256	no	yes	no	UCI
Mushrooms	200*	200*	no	yes	no	UCI
Ionosphere	176	175	no	yes	yes	UCI
Sediment	1413	471	no	yes	yes	UCI

*200 instances were used for training (101..200 and 4101..4200) and testing (201..300 and 4201..4300). These two ranges were arbitrarily selected.

TABLE III

MEAN ACCURACY FOR CLINICAL DATA CLASSIFICATION RESULTS WITH
AT LEAST 50% SENSITIVITY AND 50% SPECIFICITY. THE TOP RESULT
PER ROW IS HIGHLIGHTED IN BOLD FONT.

	Accuracy by Kernel (*Non-Mercer)					
Data Set	Lin	Pol	RBF	Sig*	SigN*	MSig
Skin Lesion	84.1	-	87.6	86.8	87.1	**88.4**
Heart	84.9	80.1	84.8	84.4	**85.0**	**85.0**
Diabetes	80.1	81.1	81.0	80.0	81.6	**81.9**
Average	83.0	-	84.5	83.7	84.6	**85.1**
Difference	-2.1	-2.9	-0.6	-1.4	-0.5	0

TABLE IV

MEAN ACCURACY FOR NON-CLINICAL DATA CLASSIFICATION RESULTS
WITH AT LEAST 50% SENSITIVITY AND 50% SPECIFICITY. THE TOP
RESULT PER ROW IS HIGHLIGHTED IN BOLD FONT.

	Accuracy by Kernel (*Non-Mercer)					
Data Set	Lin	Pol	RBF	Sig*	SigN*	MSig
Mushrooms	99.5	98.5	99.5	98.5	99.5	**100**
Ionosphere	88.2	90.8	**94.4**	87.7	91.4	93.4
Sediment	**85.5**	84.5	84.9	79.1	84.5	84.5
Average	91.1	91.3	**92.7**	88.3	91.8	92.6
Difference	-1.6	-1.4	0	-4.4	-0.9	-0.1

of 400 points), the Ionosphere data set and the Sediment data set. These data sets range from a few features (< 10) to many features (> 100) and include both heterogeneous and homogeneous data types.

We use 10-fold cross-validation for the skin lesion and Statlog Heart data sets, while other data sets are split into disjoint training and validation sets as summarized in Table II. The data sets are centered and normalized such that the two-sided third standard deviation becomes ± 1 following guidance in the literature [1]. There are eight hyperparameters which are generated as random variables [22] with a uniform distribution (Table I) as opposed to grid search. In all iterations or folds we test with sixty sets of hyperparameters. Our implementation also calculates class-specific soft-margin parameters C_+ and C_- from C to achieve a balanced success rate with imbalanced data [1].

Popular kernels are selected for comparison with Mercer sigmoid (MSig) kernel: the linear (Lin), polynomial (Pol), Gaussian RBF (RBF) and sigmoid (Sig) kernels [13]. We also produce results for the normalized sigmoid (SigN) kernel. Our implementation solves the SVM using Quadratic Programming (QP) unless it takes too many iterations to solve, in which case it switches to Sequential Minimal Optimization (SMO). SMO is used as the default for the MSig kernel.

VI. RESULTS

We report the results of our classification experiments on clinical data (Table III) and non-clinical data (Table IV) in terms of the highest classification accuracy with at least 50% sensitivity and specificity. The experiments have 60 iterations to cover 60 sets of hyperparameter; and for cross-validated heart and skin data, these iterations are multiplied by five, since five different sets of ten-folds were chosen. These experiments are repeated 25 times (i.e. $n = 25$ or $n = 30$) [23] so that the statistical significance of the best result can be evaluated.

With clinical data the median accuracy of the Mercer sigmoid kernel is significantly better ($\alpha = 5\%$) than all of the other kernels for results with at least 50% sensitivity and specificity. With non-clinical data there is no significance difference between the median accuracy of the Mercer sigmoid kernel and the Gaussian RBF kernel. Over all of the data sets, the Mercer sigmoid kernel consistently classifies some points correctly that the Gaussian RBF kernel does not (e.g. 3.6% of test instances, on average).

We note that the Mercer sigmoid kernel's better performance on clinical data versus non-clinical data appears to be correlated with the heterogeneity of the data. That is, the Mercer sigmoid kernel outperforms the Gaussian RBF kernel on all three clinical data sets and one non-clinical data set (Mushrooms) where multiple data types are present: real numbers, counts, binary values and categorical/nominal values. Whereas the other two data sets consist only of real numbers (Table II).

The Mercer sigmoid kernel uses less support vectors (SV) than the Gaussian RBF kernel with the six data sets: 197 versus 229 SV for the best results; and 200 versus 236 SV on average; while the sigmoid kernel uses 138 SV on average. The Mercer sigmoid also had the smallest average execution time of 291ms (with a 2.3GHz Intel quad core i7 computer) which is not surprising given that we use SMO to solve the SVM, whereas the Gaussian RBF kernel took 551ms on average using QP.

VII. CONCLUSIONS

A Mercer sigmoid kernel that is similar to the (normalized) sigmoid kernel (when the shift parameter $d = 0$), is now

available for classification in clinical applications, free of the limitations and concerns that encumber the sigmoid kernel and thereby fulfilling interests expressed in the literature. While it has not been investigated in other contexts such as with genomic data or with big data, it is a separable kernel with lower storage complexity which is crucial for big data[18].

Detecting melanoma with heterogeneous skin lesion data is one of our primary interests and importantly the Mercer sigmoid kernel has better median accuracy than other kernels with that data set and with two other clinical data sets in our experiments, for results with at least 50% sensitivity and specificity. With the three non-clinical data sets in our experiments there was no significant difference between the median accuracy of the Mercer sigmoid kernel and the Gaussian RBF kernel.

ACKNOWLEDGMENT

REFERENCES

[1] Asa Ben-Hur and Jason Weston, "A user's guide to support vector machines," in *Data mining techniques for the life sciences*, pp. 223–239. Springer, 2010.

[2] Christopher M Bishop, *Pattern recognition and machine learning*, vol. 4, Springer New York, 2006.

[3] Joseph A Cruz and David S Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, pp. 59, 2006.

[4] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee, "An integrated machine learning approach to stroke prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 183–192.

[5] Tatiana Tommasi, Elisabetta La Torre, and Barbara Caputo, "Melanoma recognition using representative and discriminative kernel classifiers," in *Computer Vision Approaches to Medical Image Analysis*, pp. 1–12. Springer, 2006.

[6] Pavel Brazdil, Christophe Giraud Carrier, Carlos Soares, and Ricardo Vilalta, *Metalearning: Applications to data mining*, Springer, 2008.

[7] Hsuan-Tien Lin and Chih-Jen Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," *submitted to Neural Computation*, pp. 1–32, 2003.

[8] Sabri Boughorbel, J-P Tarel, and Nozha Boujemaa, "Conditionally positive definite kernels for SVM based image recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2005, pp. 113–116.

[9] Bernhard Scholkopf, "The kernel trick for distances," *Advances in neural information processing systems*, pp. 301–307, 2001.

[10] Nello Cristianini and John Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press, 2000.

[11] John Shawe-Taylor and Nello Cristianini, *Kernel methods for pattern analysis*, Cambridge university press, 2004.

[12] Jun Han and Claudio Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *From Natural to Artificial Neural Computation*, pp. 195–201. Springer, 1995.

[13] Christopher JC Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[14] Gustavo Camps-Valls, José David Martín-Guerrero, José Luis Rojo-Álvarez, and Emilio Soria-Olivas, "Fuzzy sigmoid kernel for support vector classifiers," *Neurocomputing*, vol. 62, pp. 501–506, 2004.

[15] Liu Han, Liu Ding, and Deng Ling-Feng, "Chaotic time series prediction using fuzzy sigmoid kernel-based support vector machines," *Chinese Physics*, vol. 15, no. 6, pp. 1196, 2006.

[16] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al., "A practical guide to support vector classification," 2003.

[17] Richard O Duda, Peter E Hart, and David G Stork, *Pattern Classification and Scene Analysis 2nd ed.*, John Wiley and Sons, 1995.

[18] Marc G Genton, "Classes of kernels for machine learning: a statistics perspective," *The Journal of Machine Learning Research*, vol. 2, pp. 299–312, 2002.

[19] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[20] Dr. Eric Ehrsam, "Dermoscopy," http://dermoscopic.blogspot.com, 2007.

[21] Bache K and M Lichman, "UCI machine learning repository," http://archive.ics.uci.edu/ml, 2013.

[22] James Bergstra and Yoshua Bengio, "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[23] Robert V Hogg and Elliot A Tanis, *Probability and statistical inference*, Prentice Hall Inc., 7 edition, 2005.