A New Mercer Sigmoid Kernel for Clinical Data Classification

André Carrington[1], Paul Fieguth[2], Helen Chen[3]

***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 accuracy on three clinical data sets, detecting melanoma in skin lesions better than the most popular kernels, and it ties the Gaussian RBF kernel in accuracy when three non-clinical data sets are included. It consistently classifies some points correctly that the Gaussian RBF kernel does not (and vice versa), thereby offering additional information that Multiple Kernel Learning or ensembles may be able to exploit for better classification performance.

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, 3], including applications to clinical research, diagnosis and prognosis [9, 17, 22]. 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, 3, 5].

The sigmoid kernel was once quite popular for use with SVMs [19] 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) [19]. Some literature asserts that a c.p.d. kernel is valid [4, 20], while other literature omits c.p.d. from consideration [3, 8, 21]. Therefore, as a non-Mercer kernel, the sigmoid kernel is not necessarily a trustworthy choice for clinical applications.

[1]André Carrington is in the Department of Systems Design Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada L3T 3H7 amcarrin at uwaterloo.ca

[2]Prof. Paul Fieguth is the Chair of Systems Design Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada L3T 3H7 pfieguth at uwaterloo.ca

[3]Prof. Helen Chen is in the School of Public Health and Health Systems, Applied Health Studies, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada L3T 3H7 helen.chen at uwaterloo.ca

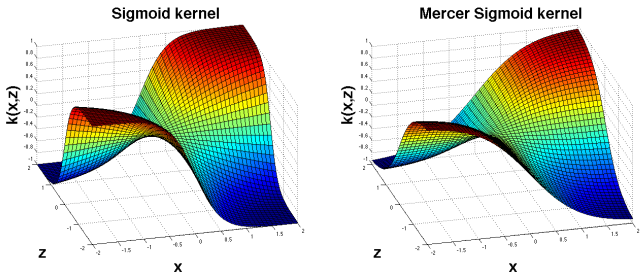

Fig. 1. The sigmoid and Mercer sigmoid kernels are similar in their output $k(\mathbf{x},\mathbf{z})$ for one-dimensional inputs $\mathbf{x}$ and $\mathbf{z}$, for a range of parameter values (r small, $b = 0$, per (1), (8)). This similarity behaviour extends to many dimensions when the former is normalized.

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 II 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 [13].

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 [21]. Kernels commonly found in SVM literature include the Gaussian RBF kernel, the linear kernel, the polynomial kernel and the sigmoid kernel [1, 6].

The sigmoid kernel [3, 19] is based on the hyperbolic tangent:

Preprint submitted to 36th Annual International IEEE EMBS Conference.
Uploaded April 12, 2014.

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

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

with a horizontal scaling parameter a, and 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):

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

$$\text{subject to} \quad 0\le\alpha_i\le\gamma, \quad i=1,...,\ell \tag{4}$$

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

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 $0<r\le\hat{r}$ for sufficiently small $\hat{r}$, dependent on the dataset [19], 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 [4, 20] — but this does not resolve the issue of determining $\hat{r}$, and the prevailing literature omits c.p.d. from consideration [3, 8, 21].

Two kernels have been created by other authors [7, 14] 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:

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

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

We can choose a hyperbolic form for ϕ,

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

leading to the proposed Mercer sigmoid (MSig) kernel

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

where there is a horizontal scaling parameter b, and a horizontal shift parameter d. The kernel is normalized by p, the dimensionality of $\mathbf{x}$ and $\mathbf{z}$, 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 that the SigN kernel has improved accuracy relative to the sigmoid kernel, on average (Table IV) and in four out of six data sets.

We then find that the MSig and SigN kernels are similar with NRMSD$\le$ 10.073% (Fig. 2) for sufficiently small r as required for the Sig and SigN kernels to be c.p.d. (we selected $-0.1\le r<0$). 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, 3, 15, 21]. It may be considered for any kernel of the form $k\left(\mathbf{x}^T\mathbf{z}\right)$, i.e. the class of zonal kernels, not just a sigmoid kernel.

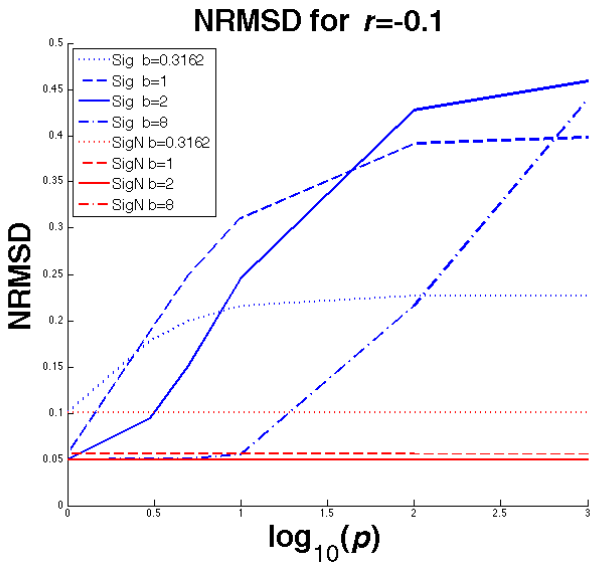

Fig. 2. The Normalized Root Mean Squared Deviation (NRMSD) between the sigmoid kernel (Sig) and the Mercer sigmoid kernel, and between the normalized sigmoid kernel (SigN) and the Mercer sigmoid kernel, for different values of b, and for $r = -0.1$, i.e. the value of r with the highest NRMSD.

B. Linearity and Dichotomy

A function or kernel saturates if it produces a bounded output range, for inputs that are unbounded [10]. 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 [12], presumably because Mercer kernels must be a dot product (implictly 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

$$x'_i = \phi(x_i) \quad \text{from (7)}$$

TABLE I
HYPERPARAMETERS FOR THE KERNELS (2) (8) AND SVM WERE GENERATED FROM A UNIFORM DISTRIBUTION WITH LOWER AND UPPER LIMITS DERIVED FROM LITERATURE [1, 19] AND EXPERIENCE. WE DENOTE $\varepsilon = 10^{-15}$ AND log AS THE BASE 10 LOGARITHM.

	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

then the kernel can be re-written as

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

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

Per Lanckriet et al, a "kernel *is a function* k, *such that* $k(\mathbf{x}, \mathbf{z}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle$ *for all* x, z $\in \mathscr{X}$, *where* Φ is a mapping from $\mathscr{X}$ to an (inner product) feature space $\mathscr{F}$. A kernel matrix *is a square matrix* $K \in \mathbb{R}^{n \times n}$ *such that* $K_{ij} = k(x_i, x_j)$ *for some* $x_1, \ldots, x_n \in \mathscr{X}$ *and some kernel function* k." [18, 21].

A. The Mercer sigmoid kernel (8) satisfies $k(\mathbf{x}, \mathbf{z}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle$, per (5-8), with a kernel matrix $K \in \mathbb{R}^{p \times p}$ for p-dimensional inputs.

B. The Mercer sigmoid kernel (8) has ϕ as a mapping from the input data space $\mathbb{R}^p$ to an (inner product) feature space $\mathbb{R}^p$, per (5-8).

C. $\mathbb{R}^p$ is a Euclidean space and therefore an inner product space.

From A, B and C we conclude that the Mercer sigmoid kernel is a valid kernel. We then use the assertion from Lanckriet et al, that all valid kernel matrices are positive semi-definite [18], to conclude that the Mercer sigmoid kernel is positive semi-definite. Finally, Genton explains that a kernel being positive semi-definite, is a necessary and sufficient condition for it to be a Mercer kernel [12]. Therefore, the Mercer sigmoid kernel is a Mercer 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 [11]. We also tested our proposed kernels with two other clinical data sets from the machine learning repository at the University of California at Irvine [16], 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 of 400 points), the Ionosphere data set and the Sediment data

TABLE II
CLINICAL DATA CLASSIFICATION ACCURACY 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	81.0	-	88.7	86.7	87.7	**89.3**
Heart	84.4	77.6	84.6	84.1	84.6	**85.3**
Diabetes	79.3	80.9	80.9	**82.4**	82.0	81.6
Average	81.6	-	84.7	84.4	84.8	**85.4**
Difference	-3.8	-4.2	-0.7	-1.0	-0.6	0

TABLE III
NON-CLINICAL DATA CLASSIFICATION ACCURACY 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	98.5	98.0	99.5	**100**
Ionosphere	86.3	90.3	**94.9**	88.0	91.4	92.0
Sediment	-	-	**84.7**	83.7	83.2	84.1
Average	-	-	**92.7**	89.9	91.4	92.0
Difference	-3.8	-2.3	0	-2.8	-1.3	-0.7

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 in a 1:2 or 2:3 ratio. 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 [2] 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 [6]. 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 II) and non-clinical data (Table III) and together (Table II), in terms of the highest classification accuracy with at least 50% sensitivity and specificity. While the experiment has many iterations (60 or 300); and the overall experiment was run several times with consistent results, further runs are required to evaluate the statistical significance the best results from one experimental run to another.

TABLE IV
A COMBINED SUMMARY OF CLINICAL AND NON-CLINICAL RESULTS

	Accuracy by Kernel (*Non-Mercer)					
Data Set	Lin	Pol	RBF	Sig*	SigN*	MSig
Average	-	-	**88.7**	87.2	88.1	**88.7**
Rank	5	4	1	3	2	1

The Mercer sigmoid kernel, on average, performs better than the sigmoid kernel and the Gaussian RBF kernel. It also consistently classifies some points correctly that the Gaussian RBF kernel does not (3.8% of points, on average).

We note that the Mercer sigmoid kernel's better performance on clinical data versus non-clinical data appears is 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.

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 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, although it has not been investigated in other contexts such as with genomic data or with big data.

The Mercer sigmoid kernel outperforms other kernels tested in our SVM classification experiments with three clinical data sets and it has the best performance tied with Gaussian RBF kernel across all six clinical and non-clinical data sets. While it achieves the same best accuracy overall it consistently classifies some points correctly that the Gaussian RBF kernel does not thereby providing additional information that Multiple Kernel Learning or ensembles may be able to exploit for better classification accuracy.

ACKNOWLEDGMENT

This work was kindly supported by AGFA and the Ontario Centres for Excellence (OCE). The authors would like to thank Prof. Ali Ghodsi for his constructive comments; and Prof. Aparna Mishra Tarc and Mrs. Krissy Carrington for editorial assistance.

REFERENCES

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

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

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

[4] Sabri Boughorbel, J-P Tarel, and Nozha Boujemaa. Conditionally positive definite kernels for svm based image recognition. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 113–116. IEEE, 2005.

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

[6] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

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

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

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

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

[11] Dr. Eric Ehrsam. Dermoscopy. `http://dermoscopic.blogspot.com`, 2007. URL `http://dermoscopic.blogspot.com`.

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

[13] 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*, pages 195–201. Springer, 1995.

[14] Liu Han, Liu Ding, and Deng Ling-Feng. Chaotic time series prediction using fuzzy sigmoid kernel-based support vector machines. *Chinese Physics*, 15(6):1196, 2006.

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

[16] Bache K and M Lichman. Uci machine learning repository. `http://archive.ics.uci.edu/ml`, 2013.

[17] 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*, pages 183–192. ACM, 2010.

[18] 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*, 5:27–72, 2004.

[19] 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*, pages 1–32, 2003.

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

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

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