Neural Networks 57 (2014) 23–28

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Comparing fixed and variable-width Gaussian networks

Věra Kůrková [a,*], Paul C. Kainen [b]

[a] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
[b] Department of Mathematics and Statistics, Georgetown University, 3700 Reservoir Rd., N.W., Washington, DC 20057, USA

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form 4 May 2014
Accepted 11 May 2014
Available online 20 May 2014

Keywords:
Gaussian radial and kernel networks
Functionally equivalent networks
Universal approximators
Stabilizers defined by Gaussian kernels
Argminima of error functionals

ABSTRACT

The role of width of Gaussians in two types of computational models is investigated: Gaussian radial-basis-functions (RBFs) where both widths and centers vary and Gaussian kernel networks which have fixed widths but varying centers. The effect of width on functional equivalence, universal approximation property, and form of norms in reproducing kernel Hilbert spaces (RKHS) is explored. It is proven that if two Gaussian RBF networks have the same input–output functions, then they must have the same numbers of units with the same centers and widths. Further, it is shown that while sets of input–output functions of Gaussian kernel networks with two different widths are disjoint, each such set is large enough to be a universal approximator. Embedding of RKHSs induced by "flatter" Gaussians into RKHSs induced by "sharper" Gaussians is described and growth of the ratios of norms on these spaces with increasing input dimension is estimated. Finally, large sets of argminima of error functionals in sets of input–output functions of Gaussian RBFs are described.

1. Introduction

Originally, artificial neural networks were built from biologically inspired computational units. These units, called perceptrons, compute functions in the form of plane waves. As an alternative, computational units in the form of spherical or elliptic waves were proposed mainly due to their good mathematical properties. Broomhead and Lowe (1988) introduced radial-basis-functions (RBFs) and Girosi and Poggio (1990) proposed more general kernel units. In particular, support vector machines (SVMs) built from units defined by symmetric positive semidefinite kernels became very popular (Cortes & Vapnik, 1995). Heaviside perceptrons cut input spaces into two halfspaces, with values of outputs equal to 0 on one half-space and 1 on the other, and so they are highly non-local. RBFs are geometrically opposite; they assign values close to 0 outside of spherical areas around their centers. Thus RBFs are localized.

Among localized computational units, a prominent position is occupied by units induced by the Gaussian function. Radial-basis-function units with the Gaussian radial function are the most common type of RBFs and Gaussian kernels with fixed widths are typical symmetric positive definite kernels. Both these computational models, the one with Gaussian RBF units having variable widths and the one with Gaussian units having fixed widths, have their advantages. Arbitrarily small widths of Gaussian RBFs were used in proofs of their universal approximation capability based on classical results on convolutions with sequences of scaled kernels (Park & Sandberg, 1991, 1993). Varying widths also play an important role in learning algorithms (see, e.g., Benoudjit, Archambeau, Lendasse, Lee, & Verleysen, 2002; Kecman, 2001; Verleysen & Hlaváčková, 1996; Wallace, Tsapatsoulis, & Kollias, 2005) and in some estimates of rates of approximation by Gaussian RBFs (see, e.g., Girosi, 1994; Girosi & Anzellotti, 1993; Kainen, Kůrková, & Sanguineti, 2009; Mhaskar, 2004). On the other hand, fixing the width allows one to fix the geometrical structure of a Hilbert space and apply the maximal margin classification algorithm (SVM) (Cortes & Vapnik, 1995). It also enables characterization of theoretically optimal solutions of learning tasks and modeling of generalization (see, e.g., Cucker & Smale, 2002; Girosi, 1998; Girosi, Jones, & Poggio, 1995; Kůrková, 2013; Poggio & Smale, 2003).

Some comparisons of capabilities of Gaussian networks with fixed and varying widths were obtained by Schmitt (2002) for the special case of input dimension equal to one. He proved that a Gaussian kernel network with a fixed width computing the same one-variable input–output function as a Gaussian RBF network with varying widths must be at least a factor of 1.5 larger.

In this paper, we investigate the role of widths of Gaussian functions in computational models which they generate. First, we

* Corresponding author. Tel.: +420 266053231.
E-mail addresses: vera@cs.cas.cz (V. Kůrková), kainen@georgetown.edu (P.C. Kainen).

http://dx.doi.org/10.1016/j.neunet.2014.05.005
0893-6080/

show that if input–output functions of two Gaussian RBF networks are equal, then the networks must have the same numbers of units and the same output weights, centers, and widths (up to a permutation of hidden units). This implies that possibilities of compressions of parameter spaces of Gaussian RBF networks are limited to equivalences induced by permutations. Our result holds for any input dimension d and any open domain in $\mathbb{R}^d$. Its proof takes advantage of the analyticity of the Gaussian function and properties of complex functions.

Further, we show that although sets of input–output functions of Gaussian kernel networks with two different widths are disjoint, each such set is large enough to be a universal approximator. In proving the density of Gaussian kernel networks, we use properties of Fourier transform of the Gaussian as an alternative to arguments of Mhaskar (1995), which are based on the form of derivatives of the Gaussian, and of Steinwart and Christmann (2008, p. 155), who use the Taylor series. Thus our results show that while no input–output function of a Gaussian RBF network whose units have at least two different widths can be exactly computed by a Gaussian kernel network with fixed width, each such function can be approximated with any required accuracy by Gaussian kernel networks having a given fixed width.

We also investigate how growth in the ratios of stabilizers induced by Gaussian kernels with two different widths depends on the input dimension. Finally, we describe multiple minima of empirical error functionals over sets of input–output functions computable by Gaussian RBFs. Some preliminary results appeared in the regional conference proceedings (Kůrková, 2013).

The paper is organized as follows. In Section 2, notations and basic concepts on one-hidden-layer RBF and kernel networks are introduced. In Section 3, it is shown that for two different widths, Gaussian kernel networks are not functionally equivalent. Section 4 shows that Gaussian kernel networks with fixed width are universal approximators. In Section 5, it is shown that the ratio of stabilizers with two different widths grows exponentially with increasing input dimension. Section 6 concludes the paper.

2. Dictionaries and kernels

The most widespread computational model used in neurocomputing is a *one-hidden-layer network with one linear output unit*. Such networks compute linear combinations of functions computable by a given type of computational units. The coefficients of linear combinations are called *output weights* and sets of functions computable by various types of units are called *dictionaries*. Networks with n units from a dictionary G compute functions from the set

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R}, g_i \in G \right\}.$$

The set of input–output functions of networks with any number of hidden units is denoted

$$\operatorname{span} G := \bigcup_{n=1}^{\infty} \operatorname{span}_n G = \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R},\ g_i \in G, n \in \mathbb{N}_+ \right\},$$

where $\mathbb{N}_+$ denotes the set of positive integers.

Typically, dictionaries are given as parameterized families of functions. Let $K : X \times Y \to \mathbb{R}$ be a function of two variables representing an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter vector $y \in Y \subseteq \mathbb{R}^s$. We denote by

$$G_K(X, Y) := \{K(., y) : X \to \mathbb{R} \mid y \in Y\},$$

the dictionary of computational units computing K. When Y is clear from the context, we write shortly $G_K(X)$ (for symmetric kernels, $X = Y$).

In mathematics, various functions of two variables are called *kernels* (from the German term "kern", introduced by Hilbert in the context of theory of integral operators (Pietsch, 1987, p. 291)). In neurocomputing and learning theory, the term kernel is often reserved for a *symmetric positive semidefinite* function. This is a kernel $K : X \times Y \to \mathbb{R}$ such that $X = Y$, $K(x, y) = K(y, x)$ for all x, $y \in X$ and for any positive integer m, any $x_1, \ldots, x_m \in X$, and any $a_1, \ldots, a_m \in \mathbb{R}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(x_i, x_j) \geq 0.$$

For symmetric positive semidefinite kernels K, the sets $\operatorname{span} G_K(X)$ of input–output functions of networks with units induced by the kernel K are contained in Hilbert spaces defined by these kernels. Such spaces are called *reproducing kernel Hilbert spaces (RKHSs)* and denoted $\mathcal{H}_K(X)$. These spaces are formed by functions from $\operatorname{span} G_K(X)$ together with limits of their Cauchy sequences with respect to the norm $\|\cdot\|_K$, so $\operatorname{span} G_K(X) \subset \mathcal{H}_K(X)$. Usually, elements of $G_K(X)$ are denoted

$$K_x(\cdot) := K(x, .).$$

The norm $\|\cdot\|_K$ is induced by the inner product $\langle \cdot, \cdot \rangle_K$, which is defined on $G_K(X) = \{K_x \mid x \in X\}$ as

$$\langle K_x, K_y \rangle_K := K(x, y).$$

In this paper, we focus on dictionaries of three types defined in terms of the Gaussian function. The first one, $G_{F_d}(X)$ is induced by the nonsymmetric function $F_d : X \times Y \to \mathbb{R}$ (where $X \subseteq \mathbb{R}^d$, $Y = \mathbb{R}_+ \times \mathbb{R}^d$, and $\mathbb{R}_+$ denotes the set of positive real numbers) defined for every $x \in X$ and $(a, c) = (a, c_1, \ldots, c_d) \in \mathbb{R}_+ \times \mathbb{R}^d$ as

$$F_d(x, (a, c)) := e^{-\|a(x-c)\|^2}.$$

So

$$G_{F_d}(X) := \left\{ F_d(., (a, c)) : X \to \mathbb{R} \mid a > 0,\ c \in \mathbb{R}^d \right\}.$$

We call networks from the set $\operatorname{span} G_{F_d}(X)$ *Gaussian RBF networks* to distinguish them from *Gaussian kernel networks* which are induced by dictionaries $G_{K_d^a}(X)$ defined for each fixed $a > 0$ corresponding to *width* $\frac{1}{a}$ as

$$G_{K_d^a}(X) := \left\{ K_d^a(., c) : X \to \mathbb{R} \mid c \in \mathbb{R}^d \right\},$$

where $K_d^a : X \times \mathbb{R}^d \to \mathbb{R}$ satisfies for every $x \in X$ and $c \in \mathbb{R}^d$

$$K_d^a(x, c) := e^{-\|a(x-c)\|^2}.$$

So $G_{K_d^a}(X)$ consists of functions on X computable by units induced by the d-variable Gaussian with a fixed width $\frac{1}{a}$. Thus we can express the dictionary $G_{F_d}(X)$ as the union of the dictionaries $G_{K_d^a}(X)$, i.e.,

$$G_{F_d}(X) := \bigcup_{a \in \mathbb{R}_+} G_{K_d^a}(X).$$

We also consider the dictionary $G_{L_d}(X)$ induced by anisotropic *elliptic Gaussian units* with widths varying in each coordinate, where the kernel $L_d : X \times \mathbb{R}_+^d \times \mathbb{R}^d \to \mathbb{R}$ is defined for each $x = (x_1, \ldots, x_d) \in X$, $a = (a_1, \ldots, a_d) \in \mathbb{R}_+^d$, and $c = (c_1, \ldots, c_d) \in \mathbb{R}^d$ as

$$L_d(x, (a, c)) := e^{-\sum_{i=1}^{d} (a_i(x_i - c_i))^2}.$$

So

$$G_{L_d}(X) := \left\{ L_d(., (a, c)) : X \to \mathbb{R} \mid a \in \mathbb{R}_+^d, c \in \mathbb{R}^d \right\}.$$