Pattern Recognition Letters 36 (2014) 107–116

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Geometrical and computational aspects of Spectral Support Estimation for novelty detection

Alessandro Rudi [a,b], Francesca Odone [b,*], Ernesto De Vito [c,d]

[a] Robotics, Brain and Cognitive Sciences Department, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy
[b] Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, University of Genova, Via Dodecaneso 35, 16146 Genova, Italy
[c] Dipartimento di Matematica, University of Genova, Via Dodecaneso 35, 16146 Genova, Italy
[d] INFN, Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy

ARTICLE INFO

Article history:
Received 15 May 2013
Available online 10 October 2013

Communicated by G. Moser

Keywords:
Support estimation
Kernel methods
Novelty detection

ABSTRACT

In this paper we discuss the Spectral Support Estimation algorithm (De Vito et al., 2010) by analyzing its geometrical and computational properties. The estimator is non-parametric and the model selection depends on three parameters whose role is clarified by simulations on a two-dimensional space. The performance of the algorithm for novelty detection is tested and compared with its main competitors on a collection of real benchmark datasets of different sizes and types.

1. Introduction

Support estimation emerged in the sixties in statistics with the seminal works of Rényi and Sulanke (1963) and Geffroy (1964), and in the last decades became crucial in different fields of machine learning and pattern recognition as, just to mention a few, one class estimation (Schölkopf et al., 2001), novelty and anomaly detection (Markou and Singh, 2003; Chandola et al., 2009). These problems find applications in different domains where it is difficult to gather negative examples (as it often happens in biological and biomedical problems) or when the negative class is not well defined (as in object detection problems in computer vision).

Support estimation deals with the following setting. The population data are represented by d-dimensional column vectors of features, but they live in a proper subset $C \subset \mathbb{R}^d$ distributed according to some probability distribution $p(x)d\nu(x)$, where $d\nu$ is a suitable infinitesimal *volume* element of C. For example, C could be a curve in $\mathbb{R}^d$, $d\nu$ is the arc length and $p(x)$ is the density distribution of the data on the curve. Both the set C and the distribution $p(x)d\nu(x)$ are known only through a training set $\{x_1, \dots, x_n\}$ of examples drawn independently from the population according to $p(x)d\nu(x)$. The aim of support estimation is to find a subset $C_n \subset \mathbb{R}^d$ such that C_n is similar to C, if n is large enough.

In this paper, we assume the set C is the support of the probability distribution ρ according to which the examples are drawn. Then C is defined as the smallest closed subset of $\mathbb{R}^d$ with the property that $\rho(C) = 1$.

To this purpose we review the Spectral Support Estimation algorithm introduced in De Vito et al. (2010) with an emphasis on its geometrical and computational properties and on its applicability to real novelty detection problems.

To have good estimators some geometrical a priori assumption on C is needed. For example, if C is convex, a choice for C_n is the convex hull of the training set, as first proposed in Dümbgen and Walther (1996). If C is an arbitrary set with non-zero d-dimensional Lebesgue measure, Devroye and Wise (1980) define C_n has the union of the balls of center x_i and radius ϵ with ϵ going to 0 when the number of data increases. A different point of view is taken by the so-called plug-in estimators. In such approach one first provides an estimator of the probability density and then C_n is defined as the region with high density (Cuevas and Fraiman, 1997).

However, in many applications the data approximatively live on a low dimensional submanifold, whose Lebesgue measure is clearly zero, and one may take advantage of this a priori information by using some recent ideas about dimensionality reduction, as for example, manifold learning algorithms (Donoho and Grimes, 2003; Belkin et al., 2006, and references therein) and kernel Principal Component Analysis (Schölkopf et al., 1998). Based on this idea, Hoffmann (2007) proposes a new algorithm for novelty detection, which can be seen as a support estimation problem. This point of view is further developed in De Vito et al. (2010), where a new class of consistent estimators, called Spectral Support Estimators (SSE), is proposed.

The contribution of this paper is threefold. First, we review the SSE algorithm emphasizing its geometrical and computational

* Corresponding author.
E-mail addresses: alessandro.rudi@iit.it (A. Rudi), francesca.odone@unige.it (F. Odone), devito@dima.unige.it (E. De Vito).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.09.025

aspects (while we refer the reader interested in its statistical properties to De Vito et al. (2010)). Second, we discuss the dependence of the algorithms on its hyper-parameters with the help of a thorough qualitative analysis on synthetic data. This analysis also allows us to show the quality of the estimated support, which adapt nicely and smoothly to the training data, similarly to kernel PCA (Hoffmann, 2007). Third, we show the appropriateness of the algorithm on a large choice of real data and compare its performances against well known competitors, namely K-Nearest Neighbours, Parzen windows (Tarassenko et al., 1995), one class Support Vector Machines (Schölkopf et al., 2001), and kernel PCA for novelty detection (Hoffmann, 2007). To make the match fair, for each algorithm we select the optimal choice for the hyper-parameters following a procedure developed in Rudi et al. (2012).

To have an intuition of the SSE algorithm, suppose C is a r-dimensional linear subspace of $\mathbb{R}^d$. Consider the $d \times d$-matrix

$$T = \int_C xx' p(x)\, dx,$$

here the volume element dv of $C$is simply the r-dimensional Lebesgue measure dx. It is easy to check that the null space of T is the orthogonal complement of C in $\mathbb{R}^d$, that is, C is the linear span of all the eigenvectors of T with non-zero eigenvalues. Since a consistent estimator of T is the empirical matrix $T_n = \frac{1}{n}\sum_{i=1}^n x_i x_i'$, one can define C_n as the linear span of the eigenvectors of T_n whose eigenvalue is bigger than a threshold λ. As in supervised learning, the thresholding ensures a stable solution with respect to the noise. Now, if λ goes to zero when n increases, C_n becomes closer and closer to C, providing us with a consistent estimator. Furthermore, to test if a new point x of $\mathbb{R}^d$ belongs to C or not, a simple decision rule is given by $F_n(x) = \sum_{\ell=1}^r (u_\ell' x)^2$, where $u_1, \ldots, u_r$ are the eigenvectors spanning C_n. Indeed, $0 \leqslant F_n(x) \leqslant x'x$ for all $x \in \mathbb{R}^d$, but it is close to $x'x$ (that is, the norm of x is near to the projection of x over C_n) if and only if x is near to C. Note that if T_n is replaced by the covariance matrix, then C_n is nothing else than the principal component analysis.

More in general, if C is not a linear subspace the above algorithm does not work, as it happens in binary classification problems with linear Support Vector Machines if the two classes are not linearly separated. This suggests the use the kernel trick which requires a feature map Φ, mapping the input space $\mathbb{R}^d$ into the feature space $\mathcal{H}$, with $\Phi(C)$ a linear subspace of $\mathcal{H}$. This strong condition is satisfied by the *separating reproducing kernels* introduced in De Vito et al. (2010).

The paper is organized as follows. Section 2 introduces the separating kernels by emphasizing their geometrical properties. In Section 3 we review the SSE algorithm and in Section 4 we discuss how the algorithm is influenced by the choice of the parameters, supporting our theoretical analysis with simulations on synthetic data. In Section 5 we compare SSE with other methods from the literature on a vast selection of real datasets. Section 6 is left to a final discussion.

2. Separating kernels

We now set the mathematical framework, we discuss the role of the separating kernels and we give some examples of separating kernels.

2.1. The framework

We assume that the input space is $\mathbb{R}^d$ with the euclidean scalar product $x't$ between two column vectors x and t. The population lives on a closed subset $C \subset \mathbb{R}^d$ and is distributed according to some probability density p only defined on C, namely

$$p(x) > 0 \quad \forall x \in C \qquad \text{and} \qquad \int_C p(x)\, dv(x) = 1,$$

where again dv is the infinitesimal *volume* element of C. For any measurable subset E of $\mathbb{R}^d$, we set

$$\rho(E) = \int_{C\cap E} p(x)\, dv(x),$$

then ρ is a probability measure on $\mathbb{R}^d$ and C is the smallest closed subset of $\mathbb{R}^d$ such that $\rho(C) = 1$, namely C is the support of the measure ρ. In general, ρ does not have density with respect to the Lebesgue measure of $\mathbb{R}^d$, as it always happens if C is an r-dimensional sub-manifold with $r < d$. Further, we assume the measure ρ is unknown, but we have a training set $\{x_1, \ldots, x_n\}$ sampled independently and identically distributed according to ρ.

Our aim is to find a closed subset $C_n \subset \mathbb{R}^d$ such that C_n is statistically consistent, *i.e.,* it becomes similar to C with high probability when the number of examples n goes to infinity.

Since in the general case C is not a linear subspace, we consider a suitable feature map Φ from the input space $\mathbb{R}^d$ into a Hilbert space $\mathcal{H}$, whose scalar product will be denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. As a common practice for kernel machines, we state the condition on the feature map in terms of its reproducing kernel $K(x,t) = \langle \Phi(x)\Phi(t) \rangle_{\mathcal{H}}$. As usual, we identify $\mathcal{H}$ with the reproducing kernel Hilbert space associated with K, so that the elements of $\mathcal{H}$ are functions on $\mathbb{R}^d$, the feature map is given by $\Phi(x) = K(\cdot, x) \in \mathcal{H}$, and for any $f \in \mathcal{H}$ and $x \in \mathbb{R}^d, f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$ (Steinwart and Christmann, 2008).

In the case of SSE we need to assume K satisfies the following properties:

(i) *Mercer*: the map $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is continuous, *i.e., K* is a Mercer kernel.
(ii) *Normalization*: for all $x \in X$ it holds that $K(x,x) = 1$.
(iii) *Separability*: for any closed subset $T \subset \mathbb{R}^d$ and any point $x \notin T$ there exists $f \in \mathcal{H}$ such that $\langle f, \Phi(x) \rangle_{\mathcal{H}} \neq 0$ and $\langle f, \Phi(t) \rangle_{\mathcal{H}} = 0$ for all $t \in T$.

As shown in De Vito et al. (2010) this assumption is crucial to prove the statistical consistency of the SSE algorithm.

The requirement that K is a Mercer kernel is very natural for kernel machines, whereas the normalization assumption simply makes the computation easy and, as shown in De Vito et al. (2010), the separating property is preserved after normalization. The crucial requirement is the separability condition. Indeed, it implies that

$$\Phi(C) = \overline{span}\{\Phi(x) | x \in C\} \cap \Phi(\mathbb{R}^d),$$

which means that $\Phi(C)$ is the intersection of a linear space of $\mathcal{H}$ and $\Phi(\mathbb{R}^d)$, here $\overline{span}\{\Phi(x)\,|\, x \in C\}$ is the closed subspace generated by the family $\{\Phi(x)_{x\in C}\}$.

Examples of separating kernels are given in De Vito et al. (2010), here we list two general purpose kernels that can be applied on a large class of problems:

(a) Laplacian (Abel) kernel:

$$K(x,t) = \exp\left(-\frac{|\, x - t\,|}{\gamma}\right) \tag{1}$$

where $\gamma > 0$ and $|x - t|$ is the euclidean norm in $\mathbb{R}^d$;
(b) ℓ_1-kernel:

$$K(x,t) = \exp\left(-\sum_{j=1}^d |\, x^j - t^j\,| / \gamma\right) \tag{2}$$

where $\gamma > 0$ and x^j and t^j are the j-th component of the vectors x and t, respectively.