Pattern Recognition 44 (2011) 2630–2642

Contents lists available at ScienceDirect

Pattern Recognition

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

Multivariate online kernel density estimation with Gaussian kernels

Matej Kristan [a,b,*], Aleš Leonardis [a], Danijel Skočaj [a]

[a] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[b] Faculty of Electrical Engineering, University of Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 13 December 2010
Received in revised form
7 March 2011
Accepted 17 March 2011
Available online 8 April 2011

Keywords:
Online models
Probability density estimation
Kernel density estimation
Gaussian mixture models

ABSTRACT

We propose a novel approach to online estimation of probability density functions, which is based on kernel density estimation (KDE). The method maintains and updates a non-parametric model of the observed data, from which the KDE can be calculated. We propose an online bandwidth estimation approach and a compression/revitalization scheme which maintains the KDE's complexity low. We compare the proposed online KDE to the state-of-the-art approaches on examples of estimating stationary and non-stationary distributions, and on examples of classification. The results show that the online KDE outperforms or achieves a comparable performance to the state-of-the-art and produces models with a significantly lower complexity while allowing online adaptation.

1. Introduction

Many tasks in machine learning and pattern recognition require building models from observing sequences of data. In some cases all the data may be available in advance, but processing all data in a batch becomes computationally infeasible for large datasets. Furthermore, in many real-world scenarios all the data may not available in advance, or we even want to observe some process for an indefinite duration, while continually providing the best estimate of the model from the data observed so far. We therefore require online construction of models.

Traditionally, parametric models based on the Gaussian mixture models (GMM) [1] have been applied successfully to model the data in terms of their probability density functions (pdf). They typically require specifying the number of components (or an upper bound) in advance [1,2] or implementing some data-driven criteria for selection of the appropriate number of components (e.g., [3]). Improper choice of the number of components, however, may lead to models which fail to capture the complete structure of the underlying pdf. Non-parametric methods such as Parzen kernel density estimators (KDE) [4–6] alleviate this problem by treating each observation as a component in the mixture model. There have been several studies on how to efficiently estimate the bandwidth of each component (e.g., [7–12]) and to incorporate the measurement noise into the estimated bandwidths, e.g., [13]. Several researchers have recognized the drawbacks of using same bandwidth for all components. Namely, it is beneficial to apply small bandwidth to densely populated regions of the feature space, while larger bandwidths may be appropriate for sparsely populated regions. As result, non-stationary bandwidth estimators have been proposed, e.g., [11,14,15]. One drawback of the standard KDEs is that their complexity (number of components) increases linearly with the number of the observed data. To remedy this increase, methods have been proposed to reduce the number of components (compress) either to a predefined value [16,17], or to optimize some data-driven criteria [18–20]. Recently, Rubio and Lobato [17] applied the non-stationary bandwidths from [15] to the compressed distribution, and reported improved performance.

There have been several attempts to address the online estimation in the context of merging the non-parametric quality of the kernel density estimators with the Gaussian mixture models in online applications. Typically, authors constrain their models by imposing various assumptions about the estimated distributions. Arandjelović et al. [21] proposed a scheme for online adaptation of the Gaussian mixture model which can be updated by observing as little as a single data-point at a time. However, a strong restriction is made that data are temporally coherent in feature space, which prevents its use in general applications. Priebe and Marchette [22] proposed an online EM algorithm, called active mixtures, which allows adaptation from a single observation at a time, assumes the data are randomly sampled from the underlying distribution, and includes a heuristic for allocating new components, which makes it less sensitive to data ordering. Kenji et al. [23] adapted this approach to compression of data-streams by volume prototypes. Song et al. [24] aimed to alleviate the restrictions on data orderings by processing data in large blocks.

Deleclerq and Piater [25] assume each data-point is a Gaussian with a predefined covariance. All data are stored in the model and

* Corresponding author at: Faculty of Computer and Information Science, University of Ljubljana, Slovenia.
E-mail address: matej.kristan@fri.uni-lj.si (M. Kristan).
URL: http://www.vicos.uni-lj.si (M. Kristan).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.03.019

a heuristic is used to determine when a subset of the data (Gaussians) can be replaced by a single component. Han et al. [26] proposed an online approach inspired by the kernel density estimation in which each new observation is added to the model as the Gaussian kernel with a predefined bandwidth. The model's complexity is maintained through the assumption that the underlying probability density function can be approximated sufficiently well by retaining only its modes. This approach deteriorates in situations when the assumed predefined bandwidths of kernels are too restrictive, and when the distribution is locally non-Gaussian (skewed or heavy tailed distribution).

A positive side of imposing assumptions on the estimated distribution is that we can better constrain the problem of estimation and design efficient algorithms for the task at hand. A downside is that once the assumptions are violated, the algorithms will likely break down and deteriorate in performance. In this paper we therefore aim at an algorithm, which would be applicable to multivariate cases, would be minimally constrained by the assumptions and therefore efficiently tackle the difficulties of online estimation.

1.1. Our approach

We propose a new online kernel density estimator which is grounded in the following two key ideas. The first key idea is that unlike the related approaches, we do not attempt to build a model of the target distribution directly, but rather maintain a non-parametric model of the data itself in a form of a *sample distribution*—this model can then be used to calculate the kernel density estimate of the target distribution. The sample distribution is constructed by online clustering of the data-points. The second key idea is that we treat each new observation as a distribution in the form of a Dirac-delta function and we model the *sample distribution* by the mixture of Gaussian and Dirac-delta functions. During online operation the sample distribution is updated by each new observation in essentially the following three steps (Fig. 1a): (1) In the step 1, we update the sample model with the observed data-point. (2) In the step 2, the updated model is used to recalculate the optimal bandwidth for the KDE. (3) In the step 3, the sample distribution is refined and compressed. This step is required because, without compression, the number of components in our model would increase linearly with the observed data. However, it turns out that a valid compression at one point in time might become invalid later, when new data-points arrive. The result of these invalid compressions is that the model misses the structure of the underlying distribution and produces significantly over-smoothed estimates.

To allow the recovery from the early compression, we keep for each component in the sample distribution another model of the data that generated that component. This detailed model is in the form of a mixture model with at most two components (Fig. 1b). The rationale behind constraining the detailed model to two components is that this is the simplest detailed model that allows detection of early over-compressions. After the compression and refinement step, the KDE can be calculated as a convolution of the (compressed) sample distribution with the optimal kernel calculated at step 2.

Our main contribution is the new multivariate online kernel density estimator (oKDE), which enables construction of a multivariate probability density estimate by observing only a single sample at a time and which can automatically balance between its complexity and generalization of the observed data-points. In contrast to the standard bandwidth estimators, which require access to all observed data, we derive a method which can use a mixture model (sample distribution) instead and can be applied to multivariate problems. To enable a controlled compression of the sample distribution, we propose a compression scheme which maintains low distance between the KDE before and after compression. To this end, we propose an approximation to the multivariate Hellinger distance on mixtures of Gaussians. Since over-compressions occur during online estimation, we propose a revitalization scheme, which detects over-compressed components and refines them, thus allowing efficient adaptation.

The remainder of the paper is structured as follows. In Section 2, we define our model. In Section 3, we derive a rule for automatic bandwidth selection. We propose the compression scheme in Section 4, where we also address the problem of over-compression. The online KDE (oKDE) algorithm is presented in Section 5. In Section 6, we analyze the influence of parameters, data order, and the reconstructive and discriminative properties of the oKDE. We compare the oKDE to existing online and batch state-of-the-art algorithms on examples of estimating distributions and on classification examples. We conclude the paper in Section 7.

2. The model definition

As stated in the introduction, we aim at maintaining a (compressed) model of the observed data-points in the form of a distribution model, and use this model to calculate the KDE when required. We therefore start with the definition of the distribution of the data-points. Each separate data-point can be presented in a distribution as a single Dirac-delta function, with its probability mass concentrated at that data-point. Noting that the Dirac-delta can be generally written as a Gaussian with zero covariance, we define the model of (potentially compressed) d-dimensional data as an N-component Gaussian mixture model

$$p_s(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \phi_{\Sigma_{si}}(\mathbf{x}-\mathbf{x}_i), \tag{1}$$

where

$$\phi_{\Sigma}(\mathbf{x}-\boldsymbol{\mu}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} e^{(-1/2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))} \tag{2}$$

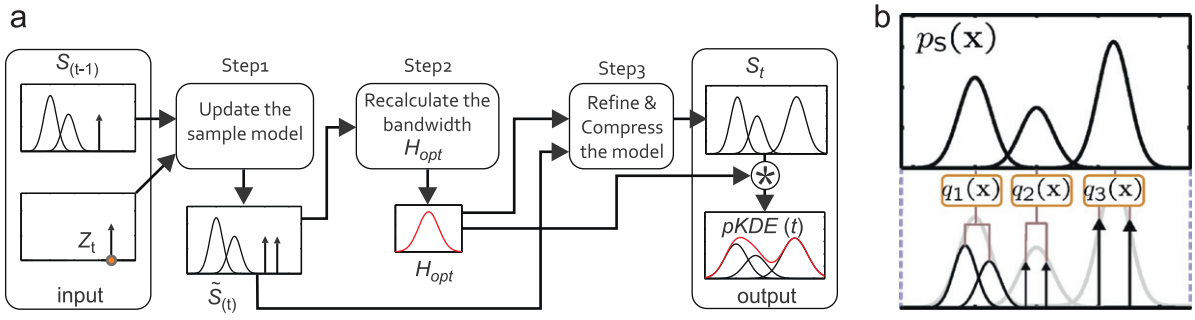

Fig. 1. A three-step summary of the online KDE iteration (a). The sample model $S_{(t-1)}$ is updated by a new observation $\mathbf{z}_t$ and compressed into a new sample model $S_{(t)}$. An illustration of the new sample model $S_{(t)}$ (sample distribution $p_s(\mathbf{x})$ along with its detailed model $\{q_i(\mathbf{x})\}_{i=1:4}$) is shown in (b).