Neural Networks 97 (2018) 137–151

Contents lists available at ScienceDirect

Neural Networks

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

Parzen neural networks: Fundamentals, properties, and an application to forensic anthropology

Edmondo Trentin [a,*], Luca Lusnig [a], Fabio Cavalli [b]

[a] DIISM-Università di Siena, Via Roma 56, I-53100 Siena, Italy
[b] Research Unit of Paleoradiology and All. Sci., AOUTS Trieste, Italy

ARTICLE INFO

Article history:
Received 30 December 2016
Received in revised form 27 September 2017
Accepted 5 October 2017
Available online 18 October 2017

Keywords:
Parzen neural network
Density estimation
Unsupervised learning
Mixture of experts
Forensic anthropology

ABSTRACT

A novel, unsupervised nonparametric model of multivariate probability density functions (pdf) is introduced, namely the Parzen neural network (PNN). The PNN is intended to overcome the major limitations of traditional (either statistical or neural) pdf estimation techniques. Besides being profitably simple, the PNN turns out to have nice properties in terms of unbiased modeling capability, asymptotic convergence, and efficiency at test time. Several matters pertaining the practical application of the PNN are faced in the paper, too. Experiments are reported, involving (i) synthetic datasets, and (ii) a challenging sex determination task from 1400 scout-view CT-scan images of human crania. Incidentally, the empirical evidence entails also some conclusions of high anthropological relevance.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Estimating probability density functions (pdf) has been a major topic in pattern recognition for decades. From the statistical viewpoint, it is the potential solution to the unsupervised learning problem, aiming at a complete description of the properties underlying the distribution of the data (Duda, Hart, & Stork, 2000). Such a description may be fundamental to investigating and communicating the primary characteristics of the data, such as multimodality, skewness, presence of heavy tails, or fitness to a reference distribution of known form, e.g., Normal. Silverman (1986) reports on several real-world phenomena where estimating the pdf of the corresponding measurements collected in the field turned out to be fundamental in disproving certain prejudicious assumptions of Normality of the data distributions. Instances are the distribution of sudden infant death syndrome depending on the degranulated mast cell count, or the distribution of the height (in microns) of a steel surface above an arbitrary level. The former is crucial to medical sciences, while the latter is fundamental to engineers in order to predict the behavior of the surface with respect to contacts with other surfaces, risk of fatigue cracks, and potential gathering of lubricant. In the supervised framework pdf estimation is (explicitly, or implicitly) involved in Bayes decision rule, e.g., for estimating the class-conditional probabilities found on the right-hand side of Bayes Theorem. Other applications of pdf estimation techniques embrace a wide spectrum of domains, e.g., data compression and model selection (Liang & Barron, 2004), coding (Beirami, Sardari, & Fekri, 2016), genomic analysis (Koslicki & Thompson, 2015) and bioinformatics in general (Yang, 2010).

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be an unlabeled sample of continuous-valued random vectors in the d-dimensional space $\mathbb{R}^d$, identically distributed and independently drawn from an unknown pdf $p(\mathbf{x})$. Multivariate density estimation is the problem of constructing an estimate $\tilde{p}(\mathbf{x})$ of $p(\mathbf{x})$ relying on the information conveyed by the random sample at hand, such that $\tilde{p}(\mathbf{x})$ is as close to the true pdf $p(\mathbf{x})$ as possible. Robust estimates are sought: broadly speaking, this means that the construction of $\tilde{p}(\mathbf{x})$ shall be unbiased and asymptotically convergent to $p(\mathbf{x})$ for $n \to \infty$. Besides, it is expected to result in meaningful estimates regardless of the form of $p(\mathbf{x})$.

The techniques for pdf estimation can be categorized into two broad families, namely parametric and nonparametric. Parametric density estimators assume that $\tilde{p}(\mathbf{x})$ has a fixed, known pdf form whose adaptivity to the specific data at hand lies in a set of parameters that characterize uniquely the estimated solution, such that the overall density estimation problem reduces to the task of estimating suitable values of the parameters from the observed data. The most popular instance of the parametric family is the Gaussian mixture model with maximum likelihood (ML) estimation of the mean vectors, covariance matrices, and mixing parameters of the mixture (Duda et al., 2000). Albeit popular, parametric estimation techniques require a somewhat arbitrary assumption on the form of the underlying, unknown distribution. Therefore, they may

* Corresponding author.
E-mail address: trentin@dii.unisi.it (E. Trentin).

https://doi.org/10.1016/j.neunet.2017.10.002
0893-6080/© 2017 Elsevier Ltd. All rights reserved.

not turn out to be viable under severe, real-world circumstances. On the other hand, nonparametric techniques, for instance the k_n-nearest neighbor (k_n-NN) (Duda et al., 2000; Fukunaga, 1990), drop the assumption and attempt a direct estimation of the pdf form from the data sample. Although nonparametric estimators involve some sort of parameters, these parameters do not constitute a fixed pdf-specific set, and their role (and, possibly even their number) rather reflects the nature and the size of the random sample. The Parzen Window (PW) is one of the most popular nonparametric approaches to pdf estimation, relying on a combination of local kernel functions centered in the patterns of the training sample (Duda et al., 2000). Although effective, PW suffers from several limitations, including: (i) the estimate is not expressed in a compact functional form, i.e. a probability law, but it is a sum of as many local kernels as the size of the sample; (ii) the local nature of the kernels tends to yield a fragmented model, which is basically "memory based" and, by definition, is prone to overfitting the "training" data (actually, there is no training at all); (iii) the whole training sample has to be kept always in memory in order to compute the estimate of the pdf over any new (say, test) pattern, resulting in a high complexity of the technique in space and time; (iv) the form of the kernel chosen has a deep influence on the eventual form of the estimated model; (v) the PW model heavily depends on the choice of an "initial" width of the kernels. Similar limitations are found in nonparametric methods in general.

In short, robust multivariate density estimation is still an open problem to date. Moreover, Vapnik stated that pdf estimation is inherently a "hard (...) computational problem", and "one cannot guarantee a good approximation using a limited number of observations" (Vapnik, 1995). Therefore, the scientific community has long been expecting improved pdf estimation techniques, possibly relying on more flexible paradigms rooted in the artificial neural network (ANN) area. ANNs can actually realize quite general input–output mappings. In principle, some relevant families of feed-forward ANNs, for instance multilayer perceptrons (MLP) (Bishop, 1995), can realize alternative nonparametric models (Trentin & Freno, 2009). Unlike parametric statistical estimators, they do not assume any specific, fixed pdf form. For instance, given suitable values for its connection weights, an MLP could as well approximate a Laplacian or a mixture of Gaussians. Actually, given their "universal approximation" property (Bishop, 1995; Cybenko, 1989), MLPs could be a suitable model for any given continuous pdf. This additional flexibility stems from their being capable of realizing a much broader family of functions than the sole class of pdfs. Unfortunately, this flexibility is also the reason why it may be hard to constrain them to satisfy the axioms of probability. Indeed, while ANNs are intensively used for estimating posterior probabilities in classification tasks (Bishop, 1995), thus far their practical exploitation for density estimation has been quite limited. It turns out that modeling probabilities via ANNs is readily achieved by standard supervised backpropagation (BP), once 0/1 target outputs are defined for the training data (Bishop, 1995), along the line of the Widrow–Hoff algorithm for linear discriminants (Duda et al., 2000). Moreover, it is simple to introduce constraints that ensure the ANN outputs may be interpreted as posterior probabilities, for instance by using logistic sigmoids in the output layer and applying any normalization mechanism such that all the outputs sum to one. Learning a pdf, to the contrary, is an unsupervised and far less obvious task. Since regular, supervised BP cannot be applied, variations on the theme of MLP training in the unsupervised framework are sought.

The most straightforward idea is to apply ML gradient-ascent training to the MLP. If $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the sample of the independent and identically distributed (i.i.d.) data drawn from the unknown pdf $p(\cdot)$, and if we write Ω to denote the ordered set of the MLP parameters (weights and biases), then the likelihood of the model given the sample is $p(\mathcal{T} \mid \Omega) = \prod_{i=1}^{n} p(\mathbf{x}_i \mid \Omega)$, where $p(\mathbf{x}_i \mid \Omega)$ is the MLP output for input $\mathbf{x}_i$. Gradient ascent prescribes the modification Δw of a generic parameter $w \in \Omega$ according to $\Delta w = \eta \partial \prod_{i=1}^{n} p(\mathbf{x}_i \mid \Omega)/\partial w$, where $\eta \in \mathbb{R}^+$ is the learning rate. Unfortunately, since an unconstrained mixture of sigmoids is not a pdf, training the MLP this way will only lead to an unbounded increase in the value of its weights (hence, of the integral of the function the MLP realizes), yielding a degenerate solution. This phenomenon is known as the "divergence problem" (Trentin & Gori, 2003). A few alternative approaches to MLP training for pdf estimation were thus proposed in the literature. A survey of ANNs for density estimation is given in Section 1.1.

The main contribution of this paper is a novel algorithm for the non-parametric estimation of multivariate pdfs relying on ANNs. The resulting model, called Parzen neural network (PNN), turns out to be simple yet general and robust, to the point that it improves significantly over the traditional approaches in the field. Moreover, a theoretical analysis of the proposed technique is contributed, showing that under mild assumptions the algorithm actually converges asymptotically to the true pdf, as long as the latter belongs to a broad family of "interesting" pdfs. Applicatively, another contribution of the paper is to exploit the novel pdf estimator in order to provide forensic anthropologists with an adaptive, automatic, and efficacious solution to the problem of sex determination from human remains. The basic idea behind the very notion of PNN is outlined in Section 1.2. Section 2 presents the PNN training algorithm, and the major practical issues that pertain its concrete utilization (e.g., model selection) are faced in Section 2.1. The latter includes also a concise discussion of the application of PNNs to classification tasks, where the PNN is either proposed as (1) an estimator of the class-conditional densities of individual classes within a Bayesian framework, resulting in a stand-alone classifier itself via Bayes theorem, or (2) as a complementary model that can be combined with other ANN-based maximum-a-posteriori discriminant functions. An illustrative demonstration of the PNN behavior is given, as well, in Section 2.2. Section 3 hands out a thorough theoretical investigation of the PNN properties in terms of computational complexity (Section 3.1), unbiased modeling capabilities (Section 3.2), and convergence (Section 3.3). It turns out that the approach overcomes the limitations of traditional statistical/neural density estimation techniques. Experiments are presented in Section 4. Some empirical demonstrations of the PNN behavior on synthetic data are given first, in Sections 4.1–4.3. Then, the PNN (in combination with standard paradigms) is successfully used in solving a hard, real-world forensic anthropology task (Section 4.4). The task is sex classification from multi-detector computerized tomography (MDCT) lateral scout-view images of human skulls, relying on 1400 images collected in the field. This is a fundamental step for identification in forensic cases and for paleodemographic studies on ancient populations (Brothwell, 1981). Conclusions are drawn in Section 5, from both the connectionist and the anthropological standpoints. Finally, Appendix A presents the pseudocode of the model selection algorithm introduced in Section 2.1, while Appendix B hands out the proof of the convergence theorem stated earlier in Section 3.3.

1.1. State-of-the-art of ANNs for pdf estimation

A solution to the divergence problem may lie in controlling the growth of the integral of the function $\phi(\cdot)$ realized by the MLP during training. Roughly speaking, the MLP output might be normalized, at the end of each training iteration, by the numerical integral of $\phi(\cdot)$. The general mathematical formalization of this idea was proposed by Modha and Fainman (1994). Modha and Fainman assume a ML criterion where the output activation function of the MLP is the exponential, ensuring the correct range for a pdf,

Download English Version:

https://daneshyari.com/en/article/6863156

Download Persian Version:

https://daneshyari.com/article/6863156

Daneshyari.com