Neural Networks 46 (2013) 227–241

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Probabilistic neural network with homogeneity testing in recognition of discrete patterns set

A.V. Savchenko *

National Research University Higher School of Economics, 25/12 Bolshaja Pecherskaja Ulitsa, Nizhny Novgorod 603155, Russia

ARTICLE INFO

Article history:
Received 21 October 2012
Revised and accepted 6 June 2013

Keywords:
Statistical pattern recognition
Discrete patterns set
Probabilistic neural network
Homogeneity testing
Face recognition
Authorship attribution

ABSTRACT

The article is devoted to pattern recognition task with the database containing small number of samples per class. By mapping of local continuous feature vectors to a discrete range, this problem is reduced to statistical classification of a set of discrete finite patterns. It is demonstrated that the Bayesian decision under the assumption that probability distributions can be estimated using the Parzen kernel and the Gaussian window with a fixed variance for all the classes, implemented in the PNN, is not optimal in the classification of a set of patterns. We presented here the novel modification of the PNN with homogeneity testing which gives an optimal solution of the latter task under the same assumption about probability densities. By exploiting the discrete nature of patterns our modification prevents the well-known drawbacks of the memory-based approach implemented in both the PNN and the PNN with homogeneity testing, namely, low classification speed and high requirements to the memory usage. Our modification only requires the storage and processing of the histograms of input and training samples. We present the results of an experimental study in two practically important tasks: (1) the problem of Russian text authorship attribution with character *n*-grams features; and (2) face recognition with well-known datasets (AT&T, FERET and JAFFE) and comparison of color- and gradient-orientation histograms. Our results support the statement that the proposed network provides better accuracy (1%–7%) and is much more resistant to change of the smoothing parameter of Gaussian kernel function in comparison with the original PNN.

1. Introduction

Pattern recognition (Theodoridis & Koutroumbas, 2009) is a fundamental problem in artificial intelligence, data mining, computer vision, medical diagnostics and decision-support systems. This problem may usually be formulated in terms of statistical recognition (Vapnik, 1998; Webb, 2002) of a set of patterns (Borovkov, 1998): it is required to estimate the class of an input sample of random variables, with an assumption that all available information about each class is concluded in certain samples of observations. This general formulation could be applied to such crucial tasks as image recognition, voice phonemes recognition, authorship attribution, etc.

The described problem is usually reduced (Duda, Hart, & Stork, 2001) to a statistical classification of the query sample. The optimal decision is taken with a minimum Bayes risk principle (Vapnik, 1998). The unknown probability density, required in this approach, is usually estimated by means of nonparametric techniques (Györfi, Kohler, Krzyzak, & Walk, 2002; Pagan & Ullah, 1999; Rutkowski, 2004b), which can adjust themselves to the data without any explicit specification (Efromovich, 1999; Greblicki, 1978; Zhang, 2000). They include Parzen's approach (Parzen, 1962), nearest neighbor algorithms (Cover & Hart, 1968), etc., which were proved to converge to the real probability density with probability one if the training sample size is large (Rutkowski, 2004a; Wolverton & Wagner, 1969).

The widely-used parallel implementation of the nonparametric approach is a probabilistic neural network (PNN). The PNN algorithm was introduced by Specht (1988, 1990a, 1990b, 1991) and approximates the class-conditional probability distributions by finite mixtures of product components under the assumption that probability distributions can be estimated using the Parzen window. Usually, likelihood function of a given class as the sum of Gaussians with a fixed variance for all the classes is applied (Montana, 1992). This algorithm is called a "neural network" because of its natural mapping onto a feedforward network (Zhang, 2000) with two hidden layers. Moreover, the mixture components can be interpreted as probabilistic neurons in neurophysiological terms (Grim & Hora, 2008). Since 1990s, the PNNs were applied to many important real-world applications, such as regression and reinforcement learning (Heinen & Enge, 2010; Specht, 1991), texture recognition (Raghu & Yegnanarayana, 1998) face detection and

* Tel.: +7 9506243285.
E-mail address: avsavchenko@hse.ru.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.06.003

recognition (Lin, Kung, & Lin, 1997), 3D objects and handwritten digits recognition (Polat & Yıldırım, 2006), optical character recognition (Romero, Touretzky, & Thibadeau, 1997), video image recognition (Aibe, Mizuno, Nakamura, Yasunaga, & Yoshihara, 2004), phoneme recognition (Maheswari, Kabilan, & Venkatesh, 2009), sentence alignment (Fattah, Ren, & Kuroiwa, 2006), medical disease estimation (Mantzarisa, Anastassopoulosb, & Adamopoulosc, 2011), analysis of electroencephalogram signals by Übeyli (2009), earthquake magnitude prediction (Adelia & Panakkat, 2009) and partial volume segmentation in brain MR image (Song, Jamshidi, Lee, & Huang, 2007).

In practice, the PNN is characterized by extremely fast training procedure and convergence to the Bayes-optimal decision surface (Specht, 1988). Though it is well known that the PNN is not a commonly used classifier and its performance is often worse than other modern classifiers as SVMs (Cortes & Vapnik, 1995), the PNN is an excellent classifier of a *set* of patterns, outperforming such classifiers as MLP (multilayer perceptron) trained by back propagation (Savchenko, 2012a). Really, this task model sets for different classes may contain equal patterns which cause the learning by single pattern to be inefficient. As a matter of fact, such neural network classifiers as SVM or MLP should be used to compare the patterns extracted from the model sets, e.g. an estimation of probability density. This approach shows good recognition accuracy only if the database contains a lot of samples per class. Unfortunately, in many practical cases we do not have enough samples (in the worst case, one sample per class, see Tan, Chen, Zhou, & Zhang, 2006). On the contrary, conventional PNN was successfully applied in such task (Savchenko, 2012a).

The PNN was proved to be an asymptotically-optimal rule in the classification task (Specht, 1990; (Musavi, Chan, Hummels, & Kalantri, 1994; Rutkowski, 2004a)) if a query object is a single feature vector. Unfortunately, conventional PNN does not provide an optimal solution (Borovkov, 1998; Savchenko, 2012a) if the query object is represented by a set of features with the size approximately equal to the training set size. Really, in this case the task should be reduced to a homogeneity testing of query and training samples (Borovkov, 1998; Kullback, 1997). Hence, we have recently introduced the PNN with homogeneity testing (Savchenko, 2012a), which saves all advantages of the conventional PNN but yields an optimal decision boundary in statistical recognition of a set of patterns. We have shown in the experiment with Russian text authorship attribution (Kukushkina, Polikarpov, & Khmelev, 2001) with simple features (frequency of punctuation marks in the sentence) that our PNN achieves better accuracy and is much more resistant to change the smoothing parameter of Gaussian kernel function (Jones, Marron, & Sheather, 1996).

Unfortunately, our network (Savchenko, 2012a) possesses the same shortcoming as the conventional PNN (Specht, 1990a). First of all, it requires large memory to store all training samples (so called memory-based approach to classification). Second, the classification speed is low as the network is based on an exhaustive search through all training samples. Hence, in this paper we explore the possibility to increase the classification speed (Savchenko, 2012b) and decrease the necessary amount of memory for our network with preservation of the optimality property. It is known that if the recognized features are *discrete* and *finite*, then the whole information about the sample is contained in its histogram (Webb, 2002). For this case the novel recognition criterion is presented here on the basis of our PNN with homogeneity testing. The thorough experimental study in two practically important recognition tasks (authorship attribution and face recognition) is discussed.

The key concept of this study is the application of the math model of the recognized object represented by independent identically distributed (i.i.d.) random variables (local feature vectors) to pattern recognition. As a matter of fact, most widely-used simple local features are continuous (e.g., gradient orientations (Dalal & Triggs, 2005; Savchenko, 2012c)) or have too wide range of definition (e.g., 256 intensity levels of gray-scale pixel). We experimentally show that their application with the PNN or the PNN with homogeneity testing leads to a significant decrease of accuracy in comparison with conventional nearest-neighbor methods (Lowe, 2004). It is demonstrated that the mapping of such features to a discrete range (e.g, 8 angle intervals in the Lowe's (2004) SIFT method of object recognition), estimation of their histograms (e.g., HOG, histograms of oriented gradients proposed by Dalal and Triggs (2005)) and comparison of these histograms with our criterion cause either better computing efficiency or much lower error rate.

The rest of the paper is organized as follows: Section 2 presents statistical recognition of a set of patterns using the conventional PNN (Rutkowski, 2004a; Specht, 1990a, 1990b (Section 2.1) and the PNN with homogeneity testing (Savchenko, 2012a) (Section 2.2). In Section 3, we introduce the modification of our PNN for recognition of discrete patterns sets. In Section 4, we present the experimental results in the author identification task (Juola, 2006) with well-known texts from the Russian literature Moshkov e-library (2013) and *n*-grams frequencies features (Kukushkina et al., 2001; Stamatatos, 2009). Section 5 demonstrate an application of our approach to the face recognition task (Zhao & Chellappa, 2005). In Section 5.1 we briefly discuss the color histograms (Rui, Huang, & Chang, 1999) and the histograms of oriented gradients (Dalal & Triggs, 2005) as image features widely used in face recognition. Section 5.2 presents the comparison of our approach with other methods (Euclidean distance, conventional PNN, Kullback–Leibler minimum information discrimination principle, Jensen–Shannon divergence) with well-known facial datasets (AT&T, 2013; FERET, 2013; JAFFE, 2013). Finally, concluding comments are given in Section 6.

2. Statistical recognition of a set of patterns

2.1. Probabilistic neural network

Let a set $\mathbf{X} = \{\mathbf{x}_j\}, j = \overline{1,n}$ of i.i.d. random variables with unknown probability distribution $\mathbf{P}$ be specified. Here n is a sample size, $\mathbf{x}_j = \{x_{j;1}, \ldots, x_{j;M}\}$—is a vector of features with a fixed number of dimensions $M = const$. The recognition problem is to estimate the class of $\mathbf{X}$. It is assumed that each class $r \in \{1., , , .R\}$ is defined by a training set of i.i.d. random variables $\mathbf{X}_r = \left\{\mathbf{x}_j^{(r)}\right\}, j = \overline{1,n_r}$ with unknown probability distribution $\mathbf{P}_r$. Here n_r is a training sample size, $\mathbf{x}_j^{(r)}$ is a feature vector with M dimensions.

Following machine learning techniques, the task is reduced to the classification problem in which training set consists of pairs $\left\{\left(\mathbf{x}_j^{(r)}, r\right)\right\}, r = \overline{1,R}, j = \overline{1,n_r}$. Then all vectors $\mathbf{x}_j$ of query sample are classified by, e.g., SVM or MLP (Duda et al., 2001). At last, the classifier outputs for each query vector are combined to make a final decision (Hsu & Lin, 2002). Unfortunately, such algorithm does not lead to a high recognition accuracy (Savchenko, 2012a). Really, though sets $\mathbf{X}$ and $\mathbf{X}_r$ contain i.i.d. variables, it does not mean that each feature vector $\mathbf{x}_j^{(r)}$ may be used to identify class r. Moreover, training sets for distinct classes may contain identical feature vectors. For instance, in image recognition each training sample is a set of pixel intensities, and different images usually contain pixels with the same intensity. Only the whole sample $\mathbf{X}_r$ should be used to uniquely identify class r. In general, every training sample contains different number of features, hence, they cannot be united into one feature vector of a fixed size.

Thus, it is more common to apply the statistical approach (Borovkov, 1998; Vapnik, 1998), in which each class is assumed to be fully determined by the distribution $\mathbf{P}_r$, $r = \overline{1,R}$ of its feature