Neurocomputing 72 (2009) 3590–3601

Contents lists available at ScienceDirect

Neurocomputing

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

Supervised data analysis and reliability estimation with exemplary application for spectral data

Frank-Michael Schleif [a,*], Thomas Villmann [b], Matthias Ongyerth [c]

[a] Computational Intelligence Group, Department of Medicine, University of Leipzig, Semmelweisstrasse 10, 04103 Leipzig, Germany
[b] University of Applied Science Mittweida, Department of Mathematics, Physics and Computer Science, Technikumplatz 17, 09648 Mittweida, Germany
[c] Computational Intelligence Group, Department of Medicine, University of Leipzig, Semmelweisstrasse 10, 04103 Leipzig, Germany

ARTICLE INFO

Available online 27 June 2009

Keywords:
Spectral analysis
Reliability estimation
Classifier optimization
Conformal prediction
Rejection region
Conformal thresholding

ABSTRACT

The analysis and classification of data is a common task in multiple fields of experimental research such as bioinformatics, medicine, satellite remote sensing or chemometrics leading to new challenges for an appropriate analysis. For this purpose different machine learning methods have been proposed. These methods usually do not provide information about the reliability of the classification. This, however, is a common requirement in, e.g. medicine and biology. In this line the present contribution offers an approach to enhance classifiers with reliability estimates in the context of prototype vector quantization. This extension can also be used to optimize precision or recall of the classifier system and to determine items which are not classifiable. This can lead to significantly improved classification results. The method is exemplarily presented on satellite remote spectral data but is applicable to a wider range of data sets.

1. Introduction

The generation of classification models is a common task in multiple fields of experimental research such as bioinformatics, medicine, satellite remote sensing or chemometrics [23,25]. Reliability estimation of the obtained classification models is frequently required. In traditional statistics this information is usually provided by significance levels, whereas for machine learning models such estimators are rare. Recently a learning theoretical approach for this problem was proposed by [33], called *conformal prediction*. We adapt this model for utilization of prototype-based classifiers like learning vector quantization (LVQ) namely supervised relevance neural gas (SRNG) [32]. This model classifies each sample prototype-based and additionally offers a level of its classification reliability.

We demonstrate the capabilities of this method for classification of satellite remote sensing spectral data. For this type of data true color images allow a visual control of classification accuracy [8]. In this specific application another aspect is given by the functional character of the data which requires an adequate handling [19,23,29]. In particular we favor the usage of functional distances for similarity determination instead of standard Euclidean (EUC) metric.

The paper is organized as follows. First we briefly introduce the main ingredients for our model. We start with a short review of the supervised relevance neural gas for prototype-based classification [32] and demonstrate how this approach can deal with different types of metrics including a functional metric. Thereafter the method of conformal prediction [33] is discussed in the light of prototype-based classifiers. It is shown how a thresholding approach can be employed in the analysis of functional spectral data combining the two measures of confidence and credibility as derived from conformal predictions. The experimental settings of our approach are defined. In the experimental section we apply our framework on data obtained from remote satellite imaging. The data are analyzed in detail and some new findings are made which have not been reported so far. The paper is closed by a summary and a discussion of open points and research directions.

2. Material and methods

2.1. Supervised neural gas for functional data

Supervised neural gas (SNG) [10] is considered as a representative for prototype-based classification approaches as introduced by Kohonen [15]. Different prototype classifiers have been proposed so far [10,15,21] as improvements of the original approach. The SNG combines the idea of neighborhood cooperativeness during learning from the unsupervised neural gas (NG) algorithm introduced in [18] with the supervised generalized

* Corresponding author. Tel.: +49 3419718955; fax: +49 3419718849.
E-mail address: schleif@informatik.uni-leipzig.de (F.-M. Schleif).
URL: http://www.uni-leipzig.de/~compint (F.-M. Schleif).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.12.040

learning vector quantizer (GLVQ) as given in [21]. Subsequently we give the basic notations and some remarks to the integration of alternative metrics into supervised neural gas. Details on SNG including convergence proofs can be found in [10].

Let us first clarify some notations: Let $c_{\mathbf{v}} \in \mathscr{L}$ be the label of input $\mathbf{v}$, $\mathscr{L}$ a set of labels (classes) with $\mathscr{L} = N_{\mathscr{L}}$. Let $V \subseteq \mathbb{R}^{D_V}$ be a finite set of inputs $\mathbf{v}$. LVQ uses a fixed number of prototypes (weight vectors, codebook vectors) for each class. Let $\mathbf{W} = \{\mathbf{w}_{\mathbf{r}}\}$ be the set of all codebook vectors and $c_{\mathbf{r}}$ be the class label of $\mathbf{w}_{\mathbf{r}}$. Furthermore, let $\mathbf{W}_c = \{\mathbf{w}_{\mathbf{r}} | c_{\mathbf{r}} = c\}$ be the subset of prototypes assigned to class $c \in \mathscr{L}$ and $\mathbf{W}_c$ is the cardinality of $\mathbf{W}_c$.

In vector quantization a stimulus vector $\mathbf{v} \in V$ is mapped onto that neuron $\mathbf{s} \in A$ the pointer $\mathbf{w}_s$ of which is closest to the presented stimulus vector $\mathbf{v}$,

$$\Psi^{\lambda}_{V \to \mathscr{A}} : \mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \underset{\mathbf{r} \in A}{\operatorname{argmin}} d^{\lambda}(\mathbf{v}, \mathbf{w}_{\mathbf{r}}) \tag{1}$$

$d^{\lambda}(\mathbf{v}, \mathbf{w})$ is an arbitrary differentiable similarity[1] measure, which may depend on a parameter vector $\boldsymbol{\lambda}$. For the moment we take $\boldsymbol{\lambda}$ as fixed. The neuron $\mathbf{s}(\mathbf{v})$ is called winner or best matching unit. The subset of the input space

$$\Omega^{\lambda}_{\mathbf{r}} = \{\mathbf{v} \in V : \mathbf{r} = \Psi_{V \to A}(\mathbf{v})\} \tag{2}$$

which is mapped to a particular neuron $\mathbf{r}$ according to (1), forms the (masked) receptive field of that neuron forming a Voronoi tessellation. If the class information of the weight vector is used, the boundaries $\partial\Omega^{\lambda}_{\mathbf{r}}$ generate the decision boundaries for classes. A training algorithm should adapt the prototypes such that for each class $c \in \mathscr{L}$, the corresponding codebook vectors $\mathbf{W}_c$ represent the class as accurately as possible. This means that the set of points in any given class $V_c = \{\mathbf{v} \in V | c_{\mathbf{v}} = c\}$, and the union $\mathscr{U}_c = \bigcup_{\mathbf{r}|_{\mathbf{w}_{\mathbf{r}} \in \mathbf{W}_c}} \Omega_{\mathbf{r}}$ of receptive fields of the corresponding prototypes should differ as little as possible.

We suppose to have m data vectors $\mathbf{v}_i$. As pointed out in [10], the neighborhood learning for a given input $\mathbf{v}_i$ with label c is applied to the subset $\mathbf{W}_c$. The respective cost function is

$$Cost_{SNG}(\gamma) = \sum_{i=1}^{m} \sum_{\mathbf{r}|\mathbf{w}_{\mathbf{r}} \in \mathbf{W}_{c_i}} \frac{h_{\gamma}(\mathbf{r}, \mathbf{v}_i, \mathbf{W}_{c_i}) \cdot f(\mu_{\lambda}(\mathbf{r}, \mathbf{v}))}{C(\gamma, K_{c_i})} \tag{3}$$

with $f(x) = (1 + \exp(-x))^{-1}, h_{\gamma}(\mathbf{r}, \mathbf{v}, \mathbf{W}) = \exp(-k_{\mathbf{r}}(\mathbf{v}, \mathbf{W})/\gamma)$ and $\mu_{\lambda}(\mathbf{r}, \mathbf{v}) = (d^{\lambda}_{\mathbf{r}} - d^{\lambda}_{\mathbf{r}_-})/(d^{\lambda}_{\mathbf{r}} + d^{\lambda}_{\mathbf{r}_-})$ whereby $d^{\lambda}_{\mathbf{r}_-}$ is defined as the squared distance to the best matching prototype but labeled with $c_{\mathbf{r}_-} \neq c_{\mathbf{v}}$, say $\mathbf{w}_{\mathbf{r}_-}$ and $d^{\lambda}_{\mathbf{r}} = d^{\lambda}(\mathbf{v}, \mathbf{w}_{\mathbf{r}})$. For a detailed formal analysis of SNG we refer to [10].

2.1.1. Incorporation of a functional metric to SNG

As pointed out before, the similarity measure $d^{\lambda}(\mathbf{v}, \mathbf{w})$ is only required to be differentiable with respect to $\boldsymbol{\lambda}$ and $\mathbf{w}$. The triangle inequality has not to be fulfilled necessarily. This leads to a great freedom in the choice of suitable measures and allows the usage of non-standard metrics in a natural way. We now review a functional metric as given in [16]. This type of metric is especially suited in case of functional data because it takes consecutive points into account which is a natural property in case of functional data. In [16] a successful application of this type of metric was shown using the well-known *tecator* data provided in [2].

The corresponding derivations can be plugged into the above equations leading to SNG with a functional metric, whereby the data are functions represented by vectors and, hence, the vector dimensions are spatially correlated. A similar situation can be observed for satellite spectra as demonstrated in [26].

Common vector processing does not take the spatial order of the coordinates into account. As a consequence, the functional aspect of spectral data is lost. For proteom spectra the order of signal features (peaks) is due to the nature of the underlying biological samples and the measurement procedure. The masses of measured chemical compounds are given ascending and peaks encoding chemical structures with a higher mass follows chemical structures with lower masses. In addition, multiple peaks with different masses may encode parts of the same chemical structure and hence are correlated.

Lee proposed a distance measure taking the functional structure into account by involving the previous and next values of x_i in the i-th term of the sum, instead of x_i alone. Assuming a constant sampling period τ, the proposed norm is

$$\mathscr{L}^{FCC}_p(\mathbf{v}) = \left(\sum_{k=1}^{D} (A_k(\mathbf{v}) + B_k(\mathbf{v}))^p \right)^{1/p} \tag{4}$$

with

$$A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leqslant v_k v_{k-1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k| + |v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \tag{5}$$

$$B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leqslant v_k v_{k+1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k| + |v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \tag{6}$$

are, respectively, of the triangles on the left and right sides of x_i. Just as for L_p, the value of p is assumed to be a positive integer. At the left and right ends of the sequence, x_0 and x_D are assumed to be equal to zero. The derivatives for the functional metric taking $p = 2$ are given in [16]. Now we consider the scaled functional norm where each dimension v_i is scaled by a parameter $\lambda_i > 0$ $\lambda_i \in (0, 1]$ and $\sum_i \lambda_i = 1$. Then the scaled functional norm is

$$\mathscr{L}^{FCC}_p(\lambda\mathbf{v}) = \left(\sum_{k=1}^{D} (A_k(\lambda\mathbf{v}) + B_k(\lambda\mathbf{v}))^p \right)^{1/p} \tag{7}$$

with

$$A_k(\lambda\mathbf{v}) = \begin{cases} \frac{\tau}{2}\lambda_k|v_k| & \text{if} 0 \leqslant v_k v_{k-1} \\ \frac{\tau}{2} \frac{\lambda_k^2 v_k^2}{\lambda_k|v_k| + \lambda_{k-1}|v_{k-1}|} & \text{else} \end{cases} \tag{8}$$

$$B_k(\lambda\mathbf{v}) = \begin{cases} \frac{\tau}{2}\lambda_k|v_k| & \text{if} 0 \leqslant v_k v_{k+1} \\ \frac{\tau}{2} \frac{\lambda_k^2 v_k^2}{\lambda_k|v_k| + \lambda_{k+1}|v_{k+1}|} & \text{else} \end{cases} \tag{9}$$

The corresponding derivations can be found in [26]. Using this parametrization one can emphasize/neglect different parts of the function for classification. This distance measure can be put into SNG as shown above and has been applied subsequently in the analysis of the spectra. SNG with a parametrized metric is subsequently referred as SRNG. The functional metric will be just referred as FUNC and will be always used with metric adaptation if not stated otherwise.

2.2. Conformal prediction—reliability estimation

In the analysis of spectral data the determination of a classifier is a difficult task. The data are functional and in general high

[1] A similarity measure is a non-negative real-valued function, which, in contrast to a distance measure, does not necessarily fulfill the triangle inequality and the symmetry property.