Neurocomputing 147 (2015) 83–95

Contents lists available at ScienceDirect

Neurocomputing

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

Kernelized vector quantization in gradient-descent learning

Thomas Villmann*, Sven Haase, Marika Kaden [1]

University of Applied Sciences Mittweida, Computational Intelligence Group, Technikumplatz 17, 09648 Mittweida, Germany

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form
1 November 2013
Accepted 3 November 2013
Available online 23 June 2014

Keywords:
Vector quantization
Online learning
Kernel distances
Support vector machines
LVQ
Self-organizing maps

ABSTRACT

Prototype based vector quantization is usually proceeded in the Euclidean data space. In the last years, also non-standard metrics became popular. For classification by support vector machines, Hilbert space representations, which are based on so-called kernel metrics, seem to be very successful. In this paper we show that gradient based learning in prototype-based vector quantization is possible by means of kernel metrics instead of the standard Euclidean distance. We will show that an appropriate handling requires *differentiable universal kernels* defining the feature space metric. This allows a prototype adaptation in the original data space but equipped with a metric determined by the kernel and, therefore, it is isomorphic to respective kernel Hilbert space. However, this approach avoids the Hilbert space *representation* as known for support vector machines. We give the mathematical justification for the isomorphism and demonstrate the abilities and the usefulness of this approach for several examples including both artificial and real world datasets.

1. Introduction

Supervised and unsupervised vector quantization for clustering and classification is an ongoing topic of research. Among other techniques, prototype based models have gained a great popularity and were successfully applied in many areas.

Famous unsupervised such models are the self-organizing map (SOM, [43]), neural gas (NG, [52]) or respective fuzzy variants like fuzzy-k-means (FCM, [8,9]). These methods are mainly applied for data clustering and compression or visualization. Although clustering is an ill-posed problem, cost functions for these models exist, which are usually based on the squared Euclidean description error. For the SOM, a cost function only exists for the Heskes-variant [35].

Supervised approaches comprise the family of learning vector quantizers (LVQ, [43]) as well as support vector machines (SVM, [72]). The intention of LVQ models is to generate class typical prototypes whereas the prototypes in SVMs define the class borders and are denoted as support vectors. Original LVQ is based on a heuristic for prototype adaptation. Variants like *robust soft* LVQ (RSLVQ, [77]) or *generalized* LVQ (GLVQ, [69]) introduce cost functions reflecting the LVQ-heuristic. Yet, the Euclidean view is kept. SVMs and GLVQ are both margin classifiers: SVMs relies on the class *separation* margin [85,80] whereas GLVQ optimizes the *hypothesis* margin [19].

Recent developments in the field address the utilization of non-Euclidean metrics in vector quantization to improve the model performance for domain-specific problems like processing of functional data, e.g. spectra, time series, etc. [38,57,88], or better interpretability of the adapted models (relevance/matrix learning, [31,73]). These metrics can be subsumed as non-standard metrics. If the underlying cost function of the vector quantizer is minimized by a gradient descent learning scheme, the employed dissimilarities or distance measures are required to be differentiable [48].

One of the most challenging ideas in classification learning is the kernel trick realized in SVMs. According to this idea, the data as well as the prototypes are implicitly mapped into a high-dimensional (may be infinite) feature mapping Hilbert space (FMHS), which is uniquely determined by the kernel [72,84]. The dissimilarities, however, are still calculated using the original data whereas model adaptation is processed in the dual space of the FMHS [20]. The feature mapping is non-linear in general. Frequently, it offers a great flexibility and good separation possibility of the mapped data in the FMHS. This mapping, however, makes it more difficult to interpret the model, because the prototypes (or support vectors for SVM) are now living in the, may be infinite-dimensional, FMHS. Moreover, the support vectors are not typical representatives of the classes, as mentioned before.

Several variants of LVQ were established also integrating the kernel mapping concept in those models while keeping the idea of class-typical prototypes (Kernel GLVQ, KGLVQ) [63,62]. Yet, these

* Corresponding author.
E-mail address: thomas.villmann@hs-mittweida.de (T. Villmann).
[1] M. Kaden (former name Kästner) is supported by the European Social Foundation (ESF), Saxony.

http://dx.doi.org/10.1016/j.neucom.2013.11.048
0925-2312/

models also have to deal with the possible infinite dimension of the mapping space. Usually, the Nystrøm-approximation technique is applied to obtain a finite representation [71,62], which obviously leads to an information loss for these models in general.

In this paper we offer an alternative for the integration of kernels in prototype based vector quantization. For this purpose, we consider *differentiable* universal kernels determining a new differentiable metric in the data space to be used in the vector quantization model as suggested in [91,39]. Thus, gradient based learning becomes available, whereby the topological structure of the new metric space is isomorphic to the FMHS.

The paper is structured as follows: first we briefly review gradient based learning for famous vector quantization models. Here, we focus on SOM/NG and GLVQ as prominent examples for unsupervised and supervised learning. In this framework we particularly advert to the integration of differentiable dissimilarity measures in these models. We segue from differentiable distance measures to differentiable kernel metrics. Thereafter, we present the theoretical justification that the respective data space is isomorphic to the FMHS related to this kernel. Exemplary artificial as well as real world applications and concluding remarks complete this contribution.

2. Differentiable dissimilarities in gradient based vector quantization

Vector quantization can be distinguished into unsupervised and supervised approaches as mentioned in the Introduction. Unsupervised models are usually applied for data compressing and clustering. Supervised approaches are related to classification and regression. Here we focus on Hebbian learning based models.

2.1. Self-organizing maps and neural gas as prominent unsupervised vector quantizers

Vector quantization, based on the minimization of some reconstruction error E for a given dataset $V \subseteq \mathbb{R}^n$ of vectors $\mathbf{v}$ with respect to set of prototypes $W = \{\mathbf{w}_k\}_{k \in A}$, is well-known since many years [51,50,83]. There, A is a finite index set. Prominent examples are the k-means [26,32,49,51,94], the fuzzy-k-means (FCM, [8,9,22,23]) or neural gas (NG, [52]). Self-organizing maps differ from these approaches (SOMs, [40,41,43]) in two ways: first, the index set A is equipped with a topological structure defining a distance $\|\mathbf{r}-\mathbf{r}'\|_A$ in the set A for $\mathbf{r}, \mathbf{r}' \in A$. Usually, rectangular or hexagonal lattices are preferred. However, other structures like graphs or growing structures are also under consideration [6,36,81]. This property allows the impressive visualizations of high-dimensional data spaces under the assumption of topographic mapping [87,93]. Second, the original SOM does not minimize any cost function [24]. Yet, the variant of Heskes overcomes this drawback and has to be taken in this context [34].

Usually, the reconstruction error is given in terms of the dissimilarity measure $d(\mathbf{v}, \mathbf{w}_k)$ between data and prototypes, usually assumed to be the squared Euclidean distance. During the last years, non-standard metrics have gained great popularity for faithful data processing [17,68,88]. After learning, the distribution $Q(W)$ of the prototypes in unsupervised vector quantization reflects the data density $P(V)$ [5,21,29,55,66,86,95]. Obviously, this relation also depends on the underlying dissimilarity in the data space [86,89].

If gradient descent learning based on the given cost function is favored, the dissimilarity measure is required to be differentiable with respect to the prototypes $\mathbf{w}_k$. For example, the cost function of the Heskes variant of SOM is

$$E_{\mathrm{SOM}} = \int P(\mathbf{v}) \sum_{\mathbf{r} \in A} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}' \in A} \frac{h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}')}{2K(\sigma)} d(\mathbf{v}, \mathbf{w}_{\mathbf{r}'})\, d\mathbf{v} \tag{2.1}$$

with the so-called neighborhood function in A

$$h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') = \exp\left(-\frac{\|\mathbf{r}-\mathbf{r}'\|_A}{2\sigma^2}\right)$$

with neighborhood range σ [34]. The neighborhood function depends on the topological structure of the index set A by the distance $\|\mathbf{r}-\mathbf{r}'\|_A$. Further, the Kronecker symbol $\delta_{\mathbf{r}}^{s(\mathbf{v})}$ assigns a data vector $\mathbf{v}$ to the winning unit by

$$s(\mathbf{v}) = \underset{\mathbf{r} \in A}{\operatorname{argmin}} \left(\sum_{\mathbf{r}' \in A} h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') \cdot d(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \right).$$

$K(\sigma)$ is a normalization constant depending on the neighborhood range σ. The stochastic gradient prototype update for the Heskes-SOM with cost function E_{SOM} is given as

$$\Delta \mathbf{w}_{\mathbf{r}} = -\varepsilon h_\sigma^{SOM}(\mathbf{r}, s(\mathbf{v})) \frac{\partial d(\mathbf{v}, \mathbf{w}_{\mathbf{r}})}{\partial \mathbf{w}_{\mathbf{r}}}. \tag{2.2}$$

depending on the derivatives of the used dissimilarity measure $d(\mathbf{v}, \mathbf{w}_{\mathbf{r}})$.

If the aspect of projective mapping can be ignored, while keeping the neighborhood cooperativeness aspect to avoid local minima in vector quantization, then the Neural Gas algorithm (NG) is an alternative to SOM proposed by Martinetz [52]. The cost function of NG to be minimized writes as

$$E_{NG} = \frac{1}{2C_\sigma} \sum_{j \in A} \int P(\mathbf{v}) h_\sigma^{NG}(\mathbf{v}, W, j)\, d(\mathbf{v}, \mathbf{w}_j)\, d\mathbf{v} \tag{2.3}$$

with the neighborhood function

$$h_\sigma^{NG}(\mathbf{v}, W, i) = \exp\left(-\frac{k_i(\mathbf{v}, W)}{\sigma}\right), \tag{2.4}$$

operating in the data space and is determined by the rank function

$$k_i(\mathbf{v}, W) = \sum_j \theta(d(\mathbf{v}, \mathbf{w}_i) - d(\mathbf{v}, \mathbf{w}_j)). \tag{2.5}$$

The function $\theta(x)$ is the Heaviside function and C_σ is a constant depending on the neighborhood range σ. The prototype update obtained from the stochastic gradient of the cost function E_{NG} results in

$$\Delta \mathbf{w}_i = -\varepsilon h_\sigma^{NG}(\mathbf{v}, W, i) \frac{\partial d(\mathbf{v}, \mathbf{w}_i)}{\partial \mathbf{w}_i} \tag{2.6}$$

and is similar to that of SOM. The winner mapping rule slightly changes to

$$s(\mathbf{v}) = \underset{j \in A}{\operatorname{argmin}} (d(\mathbf{v}, \mathbf{w}_j)) \tag{2.7}$$

compared with SOM but remains a winner-take-all rule. It turns out that NG is a robust and appropriate variant of standard k-means frequently achieving better results in shorter time due to the neighborhood cooperativeness in learning [18,52].

We remark that both gradient descent learning updates (2.2) and (2.6) contain the derivatives $\partial d(\mathbf{v}, \mathbf{w}_i)/\partial \mathbf{w}_i$ of the dissimilarity measure. According to the winner determination together with the neighborhood cooperativeness in learning, SOM as well as NG realizes an Hebbian learning paradigm [67].

2.2. Learning vector quantization for supervised vector quantization

Prototype based classification learning in the context of vector quantization is mainly influenced by the pioneering work of Kohonen establishing the family of learning vector quantizers (LVQ, [44]) for supervised learning. It is based on a heuristic to approximate Bayes