Neurocomputing 169 (2015) 295–305

Contents lists available at ScienceDirect

Neurocomputing

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

Median variants of learning vector quantization for learning of dissimilarity data

David Nebel [a], Barbara Hammer [b], Kathleen Frohberg [c], Thomas Villmann [a,*]

[a] University of Applied Sciences Mittweida, Faculty of Mathematics/Natural and Computer Sciences, Computational Intelligence Group, Germany
[b] CITEC Centre of Excellence, Bielefeld University, Theoretical Computer Science Group, Germany
[c] Veterinary Medicine Department, University Leipzig, Biomolekular Science Group, Germany

ARTICLE INFO

Article history:
Received 6 August 2014
Received in revised form
10 December 2014
Accepted 12 December 2014
Available online 7 May 2015

Keywords:
Learning vector quantization
Median
Dissimilarity data

ABSTRACT

Exemplar based techniques such as affinity propagation represent data in terms of typical exemplars. This has two benefits: (i) the resulting models are directly interpretable by humans since representative exemplars can be inspected in the same way as data points, (ii) the model can be applied to any dissimilarity measure including non-Euclidean or non-metric settings. Most exemplar based techniques have been proposed in the unsupervised setting only, such that their performance in supervised learning tasks can be weak depending on the given data. We address the problem of learning exemplar-based models for general dissimilarity data in a discriminative framework in this contribution. For this purpose, we consider variants of Kohonen's learning vector quantization model to handle data with only dissimilarities between available. Here the exemplars are the prototypes. The resulting classifiers represent data in terms of sparse models thereby reaching state-of-the art results in benchmarks. For a real world data set in the field of veterinary medicine we report promising results.

1. Introduction

Machine learning has revolutionized the possibility to deal with large electronic data sets. Nevertheless, rapid technological developments continue to pose challenges to the field, such as the big data challenge, or the problem of complicated non-vectorial structures, which are increasingly common. Examples of the latter include biological sequences, mass spectra, or metabolic networks, where advanced alignment techniques, background information, or general information theoretical principles, for example, drive the comparison of data points [53,46,35]. These data cannot be embedded in the Euclidean space without loss of information. The measures which are used to compare such kind of data often do not fulfill the properties of a metric. Further, for dissimilarities such as used in social network analysis even pseudo-Euclidean embedding such as proposed in [52] might fail due to asymmetric dissimilarities. These developments have caused the need for non-vectorial machine learning tools such as e.g. structure kernels, recursive models, relational models, or quotient embeddings [36,19,14].

Since learning tasks become more and more complicated, the specific objectives are often not clear a priori. This challenge requires increasingly interactive systems which allow humans to shape the problems according to human insights and expert knowledge at hand [66]. A vital property of machine learning models in this context is their interpretability by means of semantically meaningful interfaces [56]. While interpretable models enable the change of their functionality by experts, popular black box techniques such as the Support Vector Machines (SVMs, [60]) often only provide an excellent classification performance, but no insight on why this is the case. It is hardly possible to visualize its decisions to domain experts in such a way that relevant information can be inferred based thereon by a human observer. The same argument, although to a smaller degree, is valid for alternatives such as the relevance vector machine or sparse models which typically still rely on nonlinear combinations [67].

The demand of interpretability can be met with quite diverse technologies, such as sparsity, relevance learning, or enhancement by visualization [6]. One example is offered by dissimilarity based learning: this relies on pairwise comparisons of given labeled data. Hence it is usually easy to interpret the decision: a small number of closest neighbors account for the observed classification. These neighbors can directly be inspected by experts in the same way as data points. Because of this fact, similarity based techniques enjoy a large popularity in application domains such as biomedical applications, whereby the methods range from simple k-nearest neighbor classifiers and learning vector quantization up to advanced techniques such as affinity propagation which represents a clustering in terms of typical exemplars [41,17,1].

Dissimilarity based techniques can be distinguished according to different criteria: (i) The number of data points used to represent the classifier ranging from dense models such as

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.12.096
0925-2312/

k-nearest neighbor to sparse representations such as prototype based methods. To arrive at easily interpretable models, a sparse representation in terms of few data points is necessary. (ii) The degree of supervision ranging from clustering techniques such as affinity propagation to supervised learning. Here we are interested in classification techniques, i.e. supervised learning. (iii) The complexity of the dissimilarity measure the methods can deal with ranging from vectorial techniques restricted to Euclidean spaces, adaptive techniques which learn the underlying metrics, up to tools which can deal with arbitrary similarities or dissimilarities [51,61,55]. Typically, Euclidean techniques are well suited for simple classification scenarios, but they fail if complicated structures are encountered.

Learning vector quantization (LVQ) constitutes one of the few methods to infer a sparse representation in terms of prototypes from a given data set in a supervised way [41], such that it offers a good starting point as an intuitive classification technique whose decisions can directly be inspected by humans. Although original LVQ has been introduced on somewhat heuristic grounds [41], recent developments in this context provide a solid mathematical derivation of its generalization ability and learning dynamics [37]: LVQ classifiers can be substantiated by large margin generalization bounds [13,61,27]; the dynamics of LVQ type algorithms can be derived from explicit cost functions [63]. Interestingly, already the dynamics of classical LVQ provably leads to very good generalization ability in typical model situation as investigated in the framework of online learning [9].

A severe drawback of standard LVQ type classifiers is their dependency on the Euclidean metric. This problem can partially be avoided by appropriate metric learning, see e.g. [29,61,43], or by kernel variants, see e.g. [55,59,69], which turn LVQ classifiers into state-of-the-art techniques e.g. in connection to humanoid robotics or computer vision [16,39]. However, if data are inherently non-Euclidean, these techniques cannot be applied. Beginning with the pioneering work by Bezdek [7,31,30], median and relational variants of vector quantization are applied in unsupervised learning for clustering and data compression. Recently, these approaches were reconsidered for clustering relational and ordered dissimilarity data [32,38]. Other newer approaches also use semi-supervised techniques [21,23]. A more theoretically motivated work on learning dissimilarity data is [70]. Here the data sample relations are represented in terms of dissimilarity functions and respective conditions are studied.

An extension of LVQ type learning by means of an implicit embedding in pseudo-Euclidean space has been proposed supposing the prototypes to be linear combinations of the data points [24]. Even though yielding state-of-the-art results, this technique faces two problems: it cannot be used for asymmetric dissimilarities where no pseudo-Euclidean embedding exists; by representing prototypes in terms of distributed coefficient vectors, interpretability, one of the LVQ's main benefits is lost.

In this contribution, we address these problems by taking an alternative point of view: we consider LVQ-algorithms as exemplar based learners suitable for arbitrary dissimilarities. In particular, we start from three different variants of LVQ, where the exemplars are also denoted as prototypes: the first LVQ-variant with an underlying cost function denoted as *generalized* LVQ (GLVQ, [57]), the *robust soft* LVQ (RSLVQ, [64]) and the Soft Nearest Prototype Classifier (SNPC, [64]). The GLVQ cost function is based on the approximation of the classification error generating a classification model maximizing the hypothesis margin but not being generative in the above sense. Otherwise, RSLVQ and SNPC are probabilistic supervised generative models with the additional discriminative power depending on their ability to represent a suitably nonlinear and possibly complicated decision boundary. Often, these two features, the generative and discriminative ability of the classifier, are taken into account in separate steps only, e.g. training a generative model individually on each given class and, afterwards, incorporating the supervised label information as side information for the adaptation of only a few model parameters for better classification performance, such as e.g. metric parameters [10,18]. The previously mentioned approaches try to incorporate both aspects [49,25]. Unfortunately, these methods have not been extended to general dissimilarity data, where we do not assume an underlying Euclidean metric. Thus, the focus of this paper is to close this gap.

The paper is structured as follows: First we briefly introduce the algorithms GLVQ, RSLVQ, and SNPC as cost function based variants of the original heuristic LVQ assuming vectorial data. Thereafter we propose the respective median variants requiring only dissimilarities between the data to be known. After presenting the mathematical theory, we compare the performance of these approaches on benchmark data. A bio-medical research application in the area of veterinary medicine demonstrates the abilities of these classifier models for real world problems.

2. Variants of Kohonen's learning vector quantization

Learning vector quantization comprises a family of prototype based vector quantizers for classification of vectorial data, which are trained to minimize the classification error [40,41]. For this purpose each data class is represented by at least one prototype vector. After learning, unknown data points are classified according to the class of the closest prototype. The closest prototype is determined according to the used dissimilarity measure. In the original standard variants LVQ1…LVQ3 this is the Euclidean distance. These variants, however, are trained according to heuristic optimization schemes based on stochastic data presentation. During training the prototypes are updated depending on vector shifts.

2.1. Generalized learning vector quantization – GLVQ

The GLVQ algorithm is a generalization of the heuristic learning schemes [13,26,62]. It approximates the classification error by a *differentiable* cost function such that gradient descent learning becomes available [57].

Let $x_i \in \mathbb{X}$, $i = 1, .., N$, be the data points to be learned and $\theta_j \in \Theta$, $j = 1, .., M$, be the prototypes. Further, let $c(\cdot)$ be the formal class label function, which assigns to each data point the class label $y_i = c(x_i)$. Analogously, $c_j = c(\theta_j)$ returns the predefined class label of the prototype. We introduce the distances $d^+(x_i)$ and $d^-(x_i)$ as

$$d^+(x_i) = \min_{\{\theta_j : y_i = c_j\}} d(x_i, \theta_j) \tag{2.1}$$

$$d^-(x_i) = \min_{\{\theta_j : y_i \neq c_j\}} d(x_i, \theta_j) \tag{2.2}$$

describing the minimal distances from x_i to the closest prototype of the same class (correct) and to closest prototype of any other class (incorrect), respectively. These quantities determine the classifier function

$$\mu_\alpha(x_i) = \frac{d^-(x_i) - d^+(x_i)}{d^+(x_i) + d^-(x_i)} + \alpha \tag{2.3}$$

with $\mu_\alpha(x_i) \in I_\alpha = [-1+\alpha, 1+\alpha]$. For $\alpha = 0$, the classifier function $\mu_\alpha(x_i)$ becomes negative if $d^+(x_i) > d^-(x_i)$ is valid, i.e. data point would be incorrectly classified. A value $\alpha \neq 0$ would shift this decision boundary as well as the interval I_α. Sato and Yamada defined the cost function of GLVQ to be minimized by *stochastic gradient descent learning* as