Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 19 (2006) 610–622
Neural Networks
www.elsevier.com/locate/neunet

Comparison of relevance learning vector quantization with other metric adaptive classification methods

Th. Villmann [a,*], F. Schleif [b], B. Hammer [c]

[a] Clinic for Psychotherapy, University Leipzig, Karl-Tauchnitz-Str. 25, 04107 Leipzig, Germany
[b] Department of Computer Science, Bruker Daltonik GmbH Leipzig and University Leipzig, Karl-Tauchnitz-Str. 25, 04107 Leipzig, Germany
[c] Department of Mathematics and Computer Science, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

Received 18 July 2005; accepted 18 July 2005

Abstract

The paper deals with the concept of relevance learning in learning vector quantization and classification. Recent machine learning approaches with the ability of metric adaptation but based on different concepts are considered in comparison to variants of relevance learning vector quantization. We compare these methods with respect to their theoretical motivation and we demonstrate the differences of their behavior for several real world data sets.

Keywords: Learning vector quantization; Relevance learning; Metric adaptation; Classification

1. Introduction

Data in interesting domains such as language processing, logic, chemistry, and bioinformatics often possess an inherent structure. Typical difficulties arise for machine learning within these domains: a very high or varying data dimensionality, correlations of the data elements, or a sparsely covered data space, to name just a few. Because of this fact, standard vector processing by means of Euclidean vectors faces severe problems in these cases and several approaches to deal in a more adequate way with these data structures have been developed. A very successful and interesting possibility for recursive data structures is given by the dynamics of recurrent and recursive neural networks (Gori, Frasconi, & Sperduti, 1998; Sperduti & Starita, 1997). Recently, this idea has been extended to more general graph structures in several ways (Gori, Bianchini, & Scarselli, 2001; Micheli, Hammer, & Sperduti, in press; Passerini, Ceroni, Frasconi, & Vullo, 2003; Sona, Micheli, & Sperduti, 2004); however, the data structures, which can be tackled in this way, are still restricted. Another very general alternative to deal with structured data is offered within similarity-based machine learning approaches. Here, fairly general data structures can be dealt with as soon as a similarity measure for these structures has been defined or data are embedded into a metric space (Hammer & Jain, 2004). A popular application of this idea can be found in connection to support vector machines (SVM) and other kernel methods, where a variety of different kernels such as string kernels, graph kernels, or kernels derived from a probabilistic model have been defined (Cristianini, Lodhi, Shawe-Taylor, & Watkins, 2002; Diekhans, Jaakkola, & Haussler, 2000; Gärtner, 2003). The use of specific kernels is not restricted to the SVM, but it can readily be transferred to general metric based approaches such as the median self-organizing map or nearest neighbor classification (Kohonen & Somervuo, 2002). Naturally, the similarity measure plays a crucial role in these approaches and an appropriate choice of the distance might face severe difficulties. The design of general similarity measures which can be used for any learning tasks, i.e. which guarantee a universal approximation ability and distinguishability of arbitrary structures, is one possible line of research; however, the resulting representation of data is often too complex for the concrete task and, moreover, a universal design might not be efficient for complex data structures due to principled problems (Flach, Gärtner, & Wrobel, 2003). Therefore, similarity measures which are constructed for the concrete problem based on the given data are particularly interesting since they offer an automated design of problem specific representations. A prime example of this idea is the Fisher kernel, which derives a similarity measure

* Corresponding author. Tel.: +49 341 97 18868; fax: +49 341 21 31257.
E-mail address: villmann@informatik.uni-leipzig.de (T. Villmann).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.07.013

based on a statistical model of the data (Diekhans et al., 2000). Still, the resulting kernel is fairly general since it mirrors general statistical properties of the given data set. In case of supervised learning tasks, such as classification, only those properties are relevant which are related to the class labels whereas, statistical information, which is independent of the class distribution can be abandoned. The focus of this article is a presentation and comparison of approaches, which adapt a similarity measure based on given class information for supervised learning tasks.

Pattern classification plays an important role in data processing. It is used for discrimination of patterns according to certain criteria, which may be of statistical type, structural differences, feature discrimination, etc. Thereby, the representation of the objects significantly influences the ability for discrimination. An improper representation of the object may lead to vanishing differences whereas, a suitable representation offers a clear separation of object classes. In this sense, classical statistical discriminant analysis techniques like Fisher-discriminant analysis project data onto a one-dimensional representation which should deliver the best separation of classes. Obviously, the optimal representation depends on the classification task. It is closely related to the definition of the similarity or metric between the objects used for classification (Hammer & Villmann, 2003). Hence, the similarity should be chosen adequately to the given classification task. An appropriate choice of the metric can substitute adaptations of the representation and vice versa. Also quite common, the standard Euclidean metric may not be the best choice.

One family of intuitive metric based classification algorithms is learning vector quantization (LVQ). This family comprises prototype based algorithms which try to represent the objects by typical representatives (prototypes, weight vectors). The discriminant property is realized by labeling of the prototypes such that they adapt specifically according to the given classes. As it is described in more detail later, non-standard metrics as well as metric adaptation can easily be included in advanced modifications of LVQ. A further famous approach for classification are support vector machines (SVMs). Here, the key idea is to map the data into a possibly high-dimensional representation space, which allows a linear separation of the classes. The choice of the mapping (kernel mapping) is crucial and the incorporation of metric adaptation is thus quite interesting.

In the present paper, we compare several machine learning approaches for classification in the light of metric adaptation and usage of non-standard metrics according to the given classification task. In particular, we compare the recent developments of relevance learning in LVQ, distance metric learning in SVM and relevance learning in information theoretic based LVQ. These approaches represent different paradigms as prototype-based classification, kernel regression classification, kernel mapping-based classification and mutual information maximization-based classification, respectively. We demonstrate the consequences in several real life experiments.

Since the approaches have been designed for standard vectorial data, we also present results for vectorial data sets to allow a fair comparison of the methods. However, it turns out that even for this comparably simple vector representation of data a problem-adapted metric, which integrates some structure into the problem in form of relevance information is superior to a simple Euclidean metric. To demonstrate the principled applicability of the methods to more complex data, we also integrate an example from bioinformatics with more complex input signals, the classification of spectra. Since these data are obtained as the function values of a spectrum at different wave lengths, typical characteristic of structures can be observed: a very high dimensionality, a close correlation of subsequent entries, and an only sparsely covered data space. Here, the design of metrics which also take the generalization ability of the classifier into account and which go beyond the standard Euclidean metric is of crucial importance.

The paper is structured as follows: in Section 2, the investigated methods are described, including the respective extensions to non-standard metrics and metric adaptation. They are compared for several classification tasks in Section 3 followed by a summary.

2. Methods for classification

We concentrate on four methods for classification in machine learning according to the above outlined four directions and consider them in the context of non-standard metrics and metric adaptation. These chosen methods are popular representants of their methodologies behind:

- prototype-based classification: Generalized Learning Vector Quantization (GLVQ) as an extension of the basic LVQ-algorithms introduced by Sato and Yamada (1995) combined with relevance learning and neighborhood cooperativeness (Supervised Relevance Neural Gas, SRNG) (Hammer, Strickert, & Villmann, 2005a).
- mutual information maximization: Information Theoretic LVQ according to information theoretic measures proposed by Torkkola (Torkkola, 2003; Torkkola & Campbell, 2000).
- kernel regression classification: Regression Parametric Distance Metric Learning (RPDML) invented by Zhang, Kwok, and Yeung (2003a).
- kernel based classification: Kernel Based Distance Learning (KBDL) suggested by Tsang and Kwok (Tsang & Kwok, 2003; Zhang et al., 2003a).

In the following, we briefly explain the basic ideas of the algorithms. It is followed by numerical considerations.

2.1. Supervised neural gas for generalized learning vector quantization

Supervised neural gas is considered as a representative for prototype-based classification approaches. It can be combined with the demanded feature of relevance learning. Moreover, it is a stochastic gradient descent algorithm, which is a margin