Available online at www.sciencedirect.com
SCIENCE DIRECT
Neurocomputing 69 (2006) 651–659
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression analysis

Marc Strickert[a,*], Udo Seiffert[a], Nese Sreenivasulu[b], Winfriede Weschke[b], Thomas Villmann[c], Barbara Hammer[d]

[a]*Pattern Recognition Group, Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Germany*
[b]*Gene Expression Group, IPK Gatersleben, Germany*
[c]*Clinic for Psychotherapy, University Leipzig, Germany*
[d]*Institute of Computer Science, Technical University of Clausthal, Germany*

Available online 10 January 2006

Abstract

A correlation-based similarity measure is derived for generalized relevance learning vector quantization (GRLVQ). The resulting GRLVQ-C classifier makes Pearson correlation available in a classification cost framework where data prototypes and global attribute weighting terms are adapted into directions of minimum cost function values. In contrast to the Euclidean metric, the Pearson correlation measure makes input vector processing invariant to shifting and scaling transforms, which is a valuable feature for dealing with functional data and with intensity observations like gene expression patterns. Two types of data measures are derived from Pearson correlation in order to make its benefits for data processing available in compact prototype classification models. Fast convergence and high accuracies are demonstrated for cDNA-array gene expression data. Furthermore, the automatic attribute weighting of GRLVQ-C is successfully used to rate the functional relevance of analyzed genes.

Keywords: Prototype-based learning; Adaptive metrics; Correlation measure; Learning vector quantization; GRLVQ; Gene expression analysis

1. Introduction

Pattern classification is the key technology for solving tasks in diagnostics, automation, information fusion, and forecasting. The backbone of pattern classification is the underlying distance metric: it defines how data items are compared, and it controls the grouping of data. Thus, depending on the definition of the distance, a data set can be viewed and processed from different perspectives. Unsupervised clustering with a specific similarity measure, for example, visualized as the result of a self-organizing map (SOM), provides first hints about the appropriateness of the chosen metric for meaningful data grouping [5]. In prototype-based models like the SOM, a data item can be compared with an 'average' data prototype in various ways, for example, according to the Euclidean distance or the Manhattan block distance. Different physical and geometric interpretations are obtained then, because the former measures diagonally across the vector space, while the latter sums up distances along each dimension axis. In any case, the specific structure of the data space can and should be accounted for by selecting an appropriate metric. Once a suitable metric is identified, it can be further utilized for the design of good classifiers. In supervised scenarios, auxiliary class information can be used for adapting parameters improving the specificity of data metrics during data processing, as proposed by Kaski for (semi-)supervised extensions of the SOM [4]. Another metric-adapting classification architecture is the generalized relevance learning vector quantization (GRLVQ) developed by Hammer and Villmann [3].

Data metrics in mathematical sense, however, might be too restrictive for some applications in which a relaxation

*Corresponding author. Fax: +49 39482 5 137.
E-mail addresses: stricker@ipk-gatersleben.de (M. Strickert), seiffert@ipk-gatersleben.de (U. Seiffert), srinivas@ipk-gatersleben.de (N. Sreenivasulu), weschke@ipk-gatersleben.de (W. Weschke), villmann@informatik.uni-leipzig.de (T. Villmann), hammer@in.tu-clausthal.de (B. Hammer).

0925-2312/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2005.12.004

to more general similarity measures would be useful. For example, in biological sciences often functional aspects of collected data play an important role: general spatio-temporal patterns in time series, intensity fields, or observation sequences might be more inter-related than patterns that are just spatially close in Euclidean sense. This applies to the aim of the present work, the analysis of gene expression patterns, for which the Pearson correlation is commonly used. Since recent technological achievements allow probing of thousands of gene expression levels in parallel, fast and accurate methods are required to deal with the resulting large data sets. Thereby, the definition of genetic similarity in terms of Pearson correlation should be possible, and the curse of dimensionality, related to only few available experiments in high-dimensional gene expression space, should be reduced to a minimum. Many commercial and freely available bioinformatics tools, such as ArrayMiner, GeneSpring, J-Express Pro, and Eisen's Gene Cluster use Pearson correlation for analysis. The common goal of these programs is the identification of key regulators and clusters of coexpressed genes that determine metabolic functions in developing organisms. Usually, only the metric of algorithms, which have been initially designed for processing Euclidean data, is exchanged by a 1 minus correlation term. Here, GRLVQ-C is proposed, a classifier that is mathematically derived 'from the scratch' for correlation-based classification. Its foundations are the generic update rules of generalized relevance learning vector quantization (GRLVQ, [2,3]). This allows incorporation of auxiliary information for genetic distinction, such as the developmental stage of the probed tissues, or the stress factors applied to the growing organisms. Using the GRLVQ approach with its rigid classification cost function, a fast prototype-based and intuitive classification model with very good generalization properties is derived. Both, data attribute relevances and prototype locations are obtained as a result of optimizing Pearson correlationships. The specific requirements of gene expression analysis are met in two ways: firstly, the implemented correlation measure accounts for the nature of gene expression experiments which, due to physico-chemical reasons, tend to differ in their overall intensities and in their dynamic ranges, but not in their general structure of expressed patterns. Secondly, automatic relevance weighting attenuates the curse of high dimensionality. The properties and benefits of the proposed GRLVQ-C classifier are demonstrated for real-life data sets.

2. Generalized relevance LVQ (GRLVQ) and extensions

Let $\mathbf{X} = \{(\mathbf{x}^i, y^i) \in \mathbb{R}^d \times \{1,\ldots,c\} \mid i = 1,\ldots,n\}$ be a training data set with d-dimensional elements to be classified $\mathbf{x}^k = (x_1^k,\ldots,x_d^k)$ and c classes. A set $\mathbf{W} = \{\boldsymbol{w}^1,\ldots,\boldsymbol{w}^K\}$ of prototypes in data space with class labels y^i is used for data representation, $\boldsymbol{w}^i = (w_1^i,\ldots,w_d^i, y^i) \in \mathbb{R}^d \times \{1,\ldots,c\}$.

The classification cost function to be minimized is given in the generic form [3]:

$$E_{\mathsf{GRLVQ}} := \sum_{i=1}^{n} g(q_{\boldsymbol{\lambda}}(\mathbf{x}^i)) \quad \text{with}$$

$$q_{\boldsymbol{\lambda}}(\mathbf{x}^i) = \frac{d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i) - d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i)}{d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i) + d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i)}, \quad d_{\boldsymbol{\lambda}}(\mathbf{x}) := d_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w}).$$

The classification costs of all patterns are summed up, whereby $q_{\boldsymbol{\lambda}}(\mathbf{x}^i)$ serves as quality measure of the classification depending on the degree of fit of the presented pattern $\mathbf{x}^i$ and the two closest prototypes, $\boldsymbol{w}^{i+}$ representing the same label as $\mathbf{x}^i$ and $\boldsymbol{w}^{i-}$ a different label. A sigmoid transfer function $g(x) = \mathrm{sgd}(x) = 1/(1+\exp(-x)) \in (0;1)$ is used [8]. Implicit degrees of freedom of the cost minimization are the prototype locations in the weight space and a set of adaptive parameters $\boldsymbol{\lambda}$ connected to the measure $d_{\boldsymbol{\lambda}}(\mathbf{x}) = d_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w})$ comparing pattern and prototype. In prior work, $d_{\boldsymbol{\lambda}}(\mathbf{x})$ was supposed to be a metric in mathematical sense, i.e. taking only non-negative values, conforming to the triangle inequality, with a distance of $d = 0$ only for $\boldsymbol{w} = \mathbf{x}$. These conditions enable intuitive interpretations of prototype relationships. However, if just a well-performing classifier invariant to certain features is wanted, distance conditions might be relaxed to a mere similarity measure to be plugged into the algorithm. Overall similarity maximization can be expressed in the GRLVQ framework by flipping the sign of the measure and then just keeping the minimization of E_{GRLVQ}. Since the iterative GRLVQ update implements a gradient descent on E, d must be differentiable almost everywhere, no matter if acting as distance or as similarity measure.

Partial derivatives of E_{GRLVQ} yield the generic update formulas for the closest correct and the closest wrong prototype and the metric weights:

$$\triangle \boldsymbol{w}^{i+} = -\gamma^+ \cdot \frac{\partial E_{\mathsf{GRLVQ}}}{\partial \boldsymbol{w}^{i+}} = -\gamma^+ \cdot g'(q_{\boldsymbol{\lambda}}(\mathbf{x}^i)) \times \frac{2 \cdot d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i)}{(d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i) + d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i))^2} \cdot \frac{\partial d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i)}{\partial \boldsymbol{w}^{i+}},$$

$$\triangle \boldsymbol{w}^{i-} = \gamma^- \cdot \frac{\partial E_{\mathsf{GRLVQ}}}{\partial \boldsymbol{w}^{i-}} = \gamma^- \cdot g'(q_{\boldsymbol{\lambda}}(\mathbf{x}^i)) \times \frac{2 \cdot d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i)}{(d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i) + d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i))^2} \cdot \frac{\partial d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i)}{\partial \boldsymbol{w}^{i-}},$$

$$\triangle \boldsymbol{\lambda} = -\gamma^{\lambda} \cdot \frac{\partial E_{\mathsf{GRLVQ}}}{\partial \boldsymbol{\lambda}} = -\gamma^{\lambda} \cdot g'(q_{\boldsymbol{\lambda}}(\mathbf{x}^i)) \times \frac{2 \cdot \partial d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i)/\partial \boldsymbol{\lambda} \cdot d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i) - 2 \cdot d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i) \cdot \partial d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i)/\partial \boldsymbol{\lambda}}{(d_{\boldsymbol{\lambda}}^+(\mathbf{x}^i) + d_{\boldsymbol{\lambda}}^-(\mathbf{x}^i))^2}.$$

Learning rates are γ^{λ} for the metric parameters λ_j, all initialized equally by $\lambda_j = 1/d$, $j = 1 \ldots d$; γ^+ and γ^- describe the update amount. Their choice depends on the used measure—generally, they should be chosen according to the relation $0 \leqslant \gamma^{\lambda} \ll \gamma^- \leqslant \gamma^+ \leqslant 1$and decreased within these constraints during training. Metric adaptation should be realized slowly, as a reaction to the quasi-stationary solutions for the prototype