ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Adaptive matrix distances aiming at optimum regression subspaces

M. Strickert[a*], Axel J. Soto[bc], and Gustavo E. Vazquez[b]

[a] Institute for Vision and Graphics (IVG), University of Siegen, Germany
[b] Laboratory for Research and Development in Scientific Computing, DCIC, Universidad Nacional del Sur Bahía Blanca, Argentina
[c] Planta Piloto de Ingeniería Química, CONICET-UNS, Argentina
[*] Corresponding author: strickert@informatik.uni-siegen.de

Abstract. A new supervised adaptive metric approach is introduced for mapping an input vector space to a plottable low-dimensional subspace in which the pairwise distances are in maximum correlation with distances of the associated target space. The new formalism of multivariate subspace regression (MSR) is based on cost function optimization, and it allows assessing the relevance of input vector attributes. An application to molecular descriptors in a chemical compound database is presented for targeting octanol-water partitioning properties.

Keywords. Data-driven metric, feature rating, informative subspace.

1 Introduction

The connection of data vectors with a specific target is a fundamental problem in data analysis. Input data of real valued vectors are fundamental objects in many scientific fields, for example, ranging from spectrum data and gene expression data in medicine and biology via sensor data in engineering sciences to compound fingerprints in chemistry. Targets can be categorial labels in classification tasks, real-valued dependent variables in regression problems or even vectors of properties in association scenarios.

The empirical assessment of target information is often a time consuming and expensive task, e.g., the identification of tissue types in histological samples requires manual work and wet-lab experiments. Due to this careful assessment it is assumed that those targets assigned to the sample vectors reflect, up to a few mislabelings, a reliable (constant) ground truth that should not be further transformed. In contrast to this, the data vectors live in a space of measurements that usually quantify general properties, but which preferably should be predictive of the targets, for example, by applying an appropriate transformation.

A number of different techniques exists that allow a link between the input space and the target space, such as linear discriminant analysis (LDA) for discrete class labels [3], generalized linear models (GLM) for regression tasks [2], and canonical correlation analysis (CCA) for association problems [1]. These are well-established linear models. Complementary, neural networks like feedforward networks provide a nonlinear connection between input space and the target, but they do require a choice of architectural parameters in the hidden

layer or the selection of an appropriate learning algorithm, making it difficult to assess stability and reliability.

The approach presented here allows to optimize the input vector representations for matching the target relationships. More precisely, a matrix distance with the structure of the Mahalanobis distance is adapted to yield maximum correlation of the pairwise vector distances and the associated pairwise target distances. The approach thus offers an alternative way for solving linear inverse models, such as calculated by the Moore-Penrose pseudoinverse. The new model will be called multivariate subspace regression (MSR) in the following.

Adaptive matrix metrics have proved to be useful for k-nearest neighbors [7] and learning vector quantization with local metrics [4]. Recently a feature ranking method based on a class discriminant function has been proposed as robust alternative to LDA [6] used for complementing hard feature selection strategies of evolutionary algorithms (EA) for assessing molecular descriptors for biological and physicochemical property prediction essential in drug design [5]. So far the metric-driven feature rating scheme was limited by a simplifying two-class assumption of low and high octanol-water partitioning coefficient (logP) targets. The new formalism presented in the following overcomes these limitations for dealing with feature rating in general regression contexts. A database of 439 chemical compounds used to study the influence of the underlying 73 molecular descriptors on the logP regression task.

2 Methods

Let N input vectors be given as $\mathbf{x}^j \in \mathbf{X} \subset \mathbb{R}^M, \mathbf{x}^j = (x^j_k)_{k=1\ldots M}, 1 \leq j \leq N$ with associated target vectors $\boldsymbol{l}^j \in \mathbf{L} \subset \mathbb{R}^q, \boldsymbol{l}^j = (l^j_k)_{k=1\ldots q}$. The transformable input space $\mathbf{X}$ shall be linked to the constant target space $\mathbf{L}$ by the relationship

$$S_{\mathrm{d}^v} = \mathsf{r}(\mathbf{D}_\mathbf{L}, \mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}) = \max . \tag{1}$$

Therein, $\mathbf{D}_\mathbf{L}$ is the square distance matrix between all pairs of target vectors, here defined by Euclidean distance; $\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}$ is the matrix of all input vector distances calculated by an adaptive metric depending on a parameter matrix $\boldsymbol{\lambda}$. Thus, parameters λ_i are sought that maximize the Pearson correlation (r) between input and target space.

The model parameters are obtained by maximizing the functional S_{d^v} using its gradient

$$\frac{\partial S_{\mathrm{d}^v}}{\partial \boldsymbol{\lambda}} = \frac{\partial \mathsf{r}(\mathbf{D}_\mathbf{L}, \mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})}{\partial \mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}} \cdot \frac{\partial \mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial \mathsf{r}(\mathbf{D}_\mathbf{L}, \mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})}{\partial (\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j}} \cdot \frac{\partial (\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j}}{\partial \boldsymbol{\lambda}} . \tag{2}$$

The required derivatives of the Pearson correlation are calculated by:

$$\frac{\partial \mathsf{r}(\mathbf{D}_\mathbf{L}, \mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})}{\partial (\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j}} = \frac{\left((\mathbf{D}_\mathbf{L})_{i,j} - \mu_{\mathbf{D}_\mathbf{L}}\right) - \frac{\mathscr{B}}{\mathscr{D}} \cdot \left((\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j} - \mu_{\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}}\right)}{\sqrt{\mathscr{C} \cdot \mathscr{D}}} . \tag{3}$$

Therein, $\mu_{\mathbf{D}_\mathbf{L}}$ and $\mu_{\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}}$ denote the mean values of the matrices, and the notations $\mathscr{B} = \sum_{i=1}^{N}\sum_{j=1}^{N}\big((\mathbf{D}_\mathbf{L})_{i,j}-\mu_{\mathbf{D}_\mathbf{L}}\big)\cdot\big((\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j}-\mu_{\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}}\big)$, $\mathscr{C} = \sum_{i=1}^{N}\sum_{j=1}^{N}\big((\mathbf{D}_\mathbf{L})_{i,j}-\mu_{\mathbf{D}_\mathbf{L}}\big)^2$ and $\mathscr{D} = \sum_{i=1}^{N}\sum_{j=1}^{N}\big((\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j}-\mu_{\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X}}\big)^2$ are used.

For optimization the quasi Newton Broyden-Fletcher-Goldfarb-Shanno method was taken. Optimization was stopped, when the improvement of subsequent evaluations of S_{d^v} dropped below 10^{-8}.

Because of its flexibility, the input vectors $\boldsymbol{x}^i$ and $\boldsymbol{x}^j \in \mathbf{X}$ are chosen to be compared by a matrix metric with Mahalanobis structure in this work:

$$(\mathbf{D}^{\boldsymbol{\lambda}}_\mathbf{X})_{i,j} = \mathrm{d}^v\left(\boldsymbol{x}^i, \boldsymbol{x}^j|\boldsymbol{\lambda}\right) = \sqrt{(\boldsymbol{x}^i - \boldsymbol{x}^j)^\mathsf{T}\cdot\boldsymbol{\lambda}\cdot\boldsymbol{\lambda}^\mathsf{T}\cdot(\boldsymbol{x}^i - \boldsymbol{x}^j)}. \tag{4}$$

Unlike Mahalanobis distance there is no inverse covariance matrix employed, instead, the outer self-product of the parameter matrix $\boldsymbol{\lambda} \in \mathbb{R}^{M\times u}$ defines an adaptive matrix $\boldsymbol{\Lambda} = \boldsymbol{\lambda}\cdot\boldsymbol{\lambda}^\mathsf{T}$. This positive-definite matrix $\boldsymbol{\Lambda}$ contains components that weigh the influence of attribute pairs (g, k) in the data space. Its maximum rank is u if the number of input dimensions M is larger than the u-dimensional subspace defined by $\mathbf{X}^\mathsf{T}\cdot\boldsymbol{\lambda}$. This subspace is an informative representation of the input space focused on the target association. Since, in principle, any dimension u can be chosen it is more flexible than inverse linear models which require the same dimensionality as the target space. As a very general recommendation, a choice of $u \leq M$ and $u \leq N$, or $u \leq 3$ for visualization is possible, depending on the desired representation accuracy expressed by S_{d^v}.

The derivative of Eqn. 4, useful for gradient-based optimization, is

$$\frac{\partial \mathrm{d}^v\left(\boldsymbol{x}^i, \boldsymbol{x}^j|\boldsymbol{\lambda}\right)}{\partial\boldsymbol{\lambda}} = \frac{(\boldsymbol{x}^i - \boldsymbol{x}^j)\cdot\left((\boldsymbol{x}^i - \boldsymbol{x}^j)^\mathsf{T}\cdot\boldsymbol{\lambda}\right)}{\mathrm{d}^v\left(\boldsymbol{x}^i, \boldsymbol{x}^j|\boldsymbol{\lambda}\right)}\,. \tag{5}$$

If regression targets are modeled in a one-dimensional subspace $\boldsymbol{p} = \mathbf{X}^\mathsf{T}\cdot\boldsymbol{\lambda}$, the projected scalar values obviously depend on the data vectors and the parameter vector. Arbitrary scaling and shifting of the projections $\boldsymbol{p}$ are matched to fit by choosing α and β in $\hat{\boldsymbol{p}} = \alpha\cdot\boldsymbol{p} + \beta$ such that

$$F = \sum_{i=1}^{N}\big(l_i - (\alpha\cdot p_i + \beta)\big)^2 = \min\,. \tag{6}$$

3 Results

A compound data set with 73 molecular features and associated logP values for 439 chemical compounds has been taken for the analysis, online available at `http://dig.ipk-gatersleben.de/sardux/sardux.html` [6]. Therein, an independent test set of 30 compounds has been defined that covers the range of logP values uniformly and that is not confined in the convex hull of the training data.

Two relevant cases are considered here: a multidimensional regression task on the scalar logP target values and a regression involving three disjoint classes.

While the first application shall illustrate its competitiveness with state-of-the-art inverse linear models, the second application unfolds its unique use for mapping data related to the three-dimensional space of independent (orthogonal) class labels onto a two-dimensional subspace.

The only model parameter needed to be chosen is the dimensionality of the subspace, i.e. one and two in these examples. The stability is assessed by running the optimization 10 times with randomly initialized parameter vectors $\boldsymbol{\lambda}$.

For the multidimensional regression problem, well-proven tools are available for comparison to the averages and standard deviations of 10 MSR runs: the matrix left division operator '\' based on MATLAB Householder reflections and the R:limSolve package implementing the Moore-Penrose pseudoinverse. Comparison Table 1 shows that MSR is better than MATLAB, and slightly worse then R:limSolve only for the training data. The low standard deviations of MSR indicate a very good reproducibility. The left panel of Figure 1 shows the MSR regression result of a model of median performance on the training data, using projections transformed according to Eqn. 6.

r^2	MSR	MATLAB(7.5.0):'\'	R:limSolve(11.09)
train	0.9357 ± 0.0001	0.9231	0.9361
test	0.8704 ± 0.0004	0.8413	0.8660

Table 1: Regression results of the new method compared to approaches based on pseudoinverse calculations.

The problem with three disjoint classes has been created by splitting the logP values into the lower, middle, and upper 33.3% quantile, assigning three-dimensional targets $(0, 0, 1)$ for logP < 1.78, $(0, 1, 0)$ for $1.78 \leq$ logP < 3.0132, and $(1, 0, 0)$ for logP ≥ 3.0132 values. Note that this is different from assigning integer class labels 1, 2, and 3, which, for example, would induce a closer relationship of the class labels 1 and 2 compared to 1 and 3. The right panel of Figure 1 shows the two-dimensional transformation of the data space aiming at arranging the projections according to the target relationships. Despite of logP being a continuous regression variable naturally reflected in the molecular descriptor vectors MSR is able to render a good separation with only decent overlap of the projections by a linear transformation of the 73 input vector attributes.

Figure 2 shows the attribute relevance profiles corresponding to the two regression tasks. At first glance a high degree of similarity can be detected, such as the highly important molecular van der Waals volume (Mv). Yet, descriptors like atomic polarizability (Sp) and the number of sulfor atoms(nS) show quite a different influence on the specific task. These results are quite certain, because the box plots display a high reproducibility of the model runs. As interesting to chemists, the profiles also indicate that many other variables do only have minor relevance for the regression tasks.

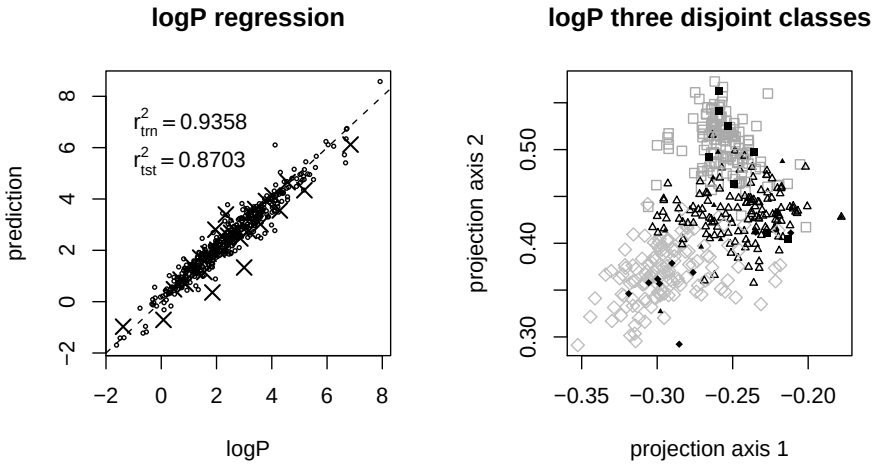

Fig. 1: MSR projection subspaces of logP data set. Left: 1D regression task (big crosses: test data). Right: 3D disjoint label separation in 2D (filled symbols: test data).

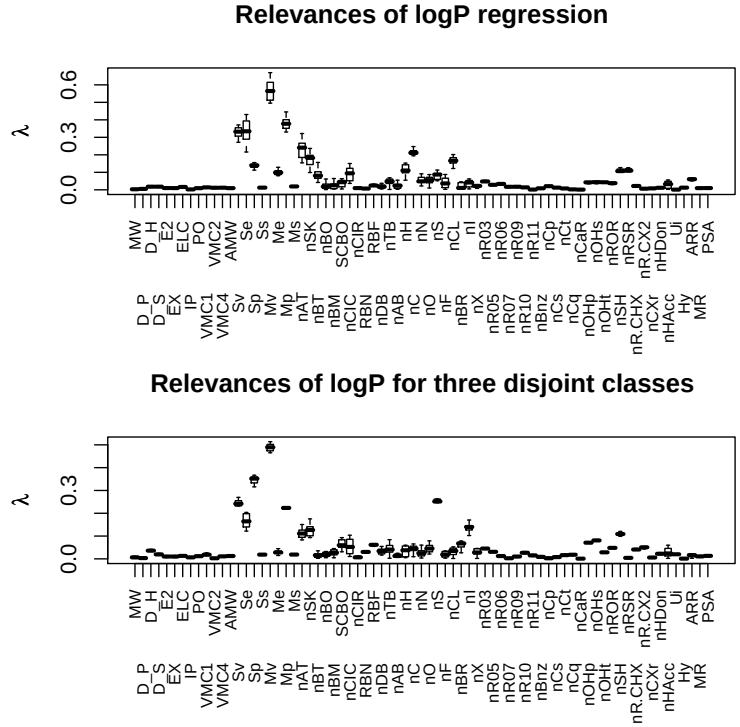

Fig. 2: Molecular descriptor relevance profiles given by $\sum_{i=1}^{u} |\boldsymbol{\lambda}^i - \mu_{\lambda^i}|$.
Top: 1D regression task ($u = 1$). Bottom: 3D disjoint label separation (u=2).

4 Conclusions

The proposed method adjusts a data metric of Mahalanobis structure for arranging the input vector relationships in good agreement to the target relationships. The metric parameters result from the optimization of a correlation-based cost function connecting input and target space. The distance can be re-interpreted as a mapping of the data vectors to a low-dimensional Euclidean space where points aim at reflecting the target relationships.

These transformed data points can be used as data replacement in subsequent analysis steps with standard Euclidean methods for classification and multivariate regression. In contrast to traditional feature assessment methods, the proposed adaptive matrix metric contains information not only about singular attributes, but about pairs of attributes. This is, for example, useful in combination with feedforward neural networks, because they integrate over input feature combinations in the hidden layer rather than utilizing single features.

Alternatively, the learned metric parameters can be used for identifying the relevance of pairs of input data attributes. As demonstrated for the logP prediction task, the rating may depend on the target of regression or multiple class labeling. The method has good empirical convergence properties and good potential for general data processing tasks.

Thanks to the anonymous reviewers. This work is kindly supported by BMBF grant ARG 08/016 and by MinCyT grant AL0811.

References

[1] C. Fyfe and G. Leen. Stochastic processes for canonical correlation analysis. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 245–250. D-facto Publications, 2006.

[2] P. McCullagh and J. Nelder. *Generalized Linear Models.* Chapman and Hall, 1989.

[3] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley-Interscience, 2004.

[4] P. Schneider, M. Biehl, and B. Hammer. Distance learning in discriminative vector quantization. *Neural Computation*, 21(10):2942–2969, 2009.

[5] A. Soto, R. Cecchini, G. Vazquez, and I. Ponzoni. Multi-Objective Feature Selection in QSAR using a Machine Learning Approach. *QSAR and Combinatorial Science*, 28(11-12):1509–1523, 2009.

[6] M. Strickert, A. Soto, J. Keilwagen, and G. Vazquez. Towards matrix-based selection of feature pairs for efficient ADMET prediction. In *Proceedings of the 9th Argentine Symposium on Artificial Intelligence (ASAI 2009)*, pages 83–94, 2009.

[7] K. Weinberger and L. Saul. Fast solvers and efficient implementations for distance metric learning. In A. McCallum and S. Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1160–1167. Omnipress, 2008.