Inferring Feature Relevances from Metric Learning

Alexander Schulz* Bassam Mokbel* Michael Biehl†
Barbara Hammer*
*CITEC centre of excellence, Bielefeld University, Germany
†University of Groningen, Mathematics and Computing Science, The Netherlands

Preprint of the publication [1], as provided by the authors.

Abstract

Powerful metric learning algorithms have been proposed in the last years which do not only greatly enhance the accuracy of distance-based classifiers and nearest neighbor database retrieval, but also enable the interpretability of these operations by assigning explicit relevance weights to the single data components. Starting with the work [2], it has been noticed, however, that this procedure has limited validity in the important case of high data dimensionality or high feature correlations: the resulting relevance profiles are random to a large extend, leading to invalid interpretation and fluctuations of its accuracy for novel data. While the work [2] proposes a first cure by means of L2-regularization, it only preserves strongly relevant features, leaving weakly relevant and not necessarily unique features undetected. In this contribution, we enhance the technique by an efficient linear programming scheme which enables the unique identification of a relevance interval for every observed feature, this way identifying both, strongly and weakly relevant features for a given metric.

1 Introduction

Popular machine learning and data retrieval techniques crucially rely on a distance computation for the given data, including the k-nearest neighbor classifier, prototype based classification, unsupervised self-organizing maps, k-means clustering, or neighborhood-based retrieval. Often, the standard Euclidean metric is used as a default for these methods, and the techniques fail provided the chosen distance is not appropriate for the given task. Due to this fact, metric learning has made great strides in recent years: the aim of metric learning is to autonomously adjust the parameters of a given distance function such that it better suits the intended task. A large variety of methods covers techniques for k-nearest neighbor classifiers, met-

ric learning in regression, metric learning in prototype based classification, metric learning based on side information, learning methods for unsupervised clustering, hybrid techniques, etc. see e.g. [3–8]. Some approaches can be accompanied by theoretical guarantees as concerns their generalization ability, or limit behavior for online learning [9–14]. These techniques greatly enhance the performance of the methods for applications, since they enable practitioners to tailor the metric according to the given data and task at hand. The methods have in common that the by far most popular distance measure which is used for this purpose is given by a general quadratic form, corresponding to a linear transformation of the feature space. Albeit first techniques consider more general data structures [10, 15], a quadratic form is used in the majority of the techniques.

Besides an improved performance, a quadratic form has the benefit that it lends itself to an improved interpretability of the result: its diagonal corresponds to the relevance of the feature dimensions for the linear transformation underlying the quadratic form, such that this relevance profile can give hints about the importance of the observed features. This fact has been heavily used in the context of biomedical data analysis, for example [16–18]. It underlines the increasing importance of the interpretability of machine learning techniques, to enable not only excellent black-box behavior, but to also allow an interaction of human experts to benefit from the findings buried in the models [19–24]. When interpreting machine learning models based on their parametrization, however, certain minimum conditions have to be fulfilled, a crucial one being the uniqueness of the observed parameters. This is not guaranteed for metric learners provided high data dimensionality or large feature correlations are present, as first observed in the contribution [2]: feature correlations cause the fact that there exist changes in the quadratic form which do not change the resulting mapping. These effects are widely induced by a random initialization of the matrix, hence, relevance profiles which display spurious relevance peaks in large parts result. The approach [2] also proposed a first cure for this observation, a L2-regularization of the quadratic form, which is standard if the linear data transformation underlying the matrix is directly learned from given data e.g. via Thikonov regularization (also known as ridge regression), but which is not yet part of most metric learning algorithms.

Albeit the proposed regularization yields unique results, it diminishes the interpretability of the relevance profiles in the following sense: L2-regularization accounts for the fact that all so-called strongly relevant features (which cannot be replaced by others without a performance loss) correspond to peaks in the relevance profile. So-called weakly relevant features, which contribute to the classification but which could be substituted by alternatives, share their relevance, hence they are no longer distinguishable provided a large number of correlated features is present. Hence interpretability as concerns this potentially useful feature contributions is lost. The problem of weakly relevant features constitutes a classical hard

problem of feature selection, which is particularly complex provided sets of features rather than single features carry certain information [25–27]. One key technique which has been pioneered in the frame of the popular lasso refers to L1-regularization rather than L2-regularization [28]: L1-regularization favors sparse signals, resulting in minimal, but not necessarily unique feature sets which also include some weakly relevant signals. Based on this insight, recently, an efficient method has been proposed which identifies relevance intervals for all given features for a linear mapping based on L1-optimization [29].

In this contribution, we take the approach published in [29] as a starting point and we extend this technology towards the setting of relevance weights for metric learners. The key observation is that a quadratic form corresponds to an implicit linear transformation of the data, such that a vectorial extension of the approach [29] allows us to quantify the relevance of a feature dimension for a given distance computation. In the following, we will formally introduce this approach, and we will investigate its performance for two popular matrix learners for different data sets including artificial data with known ground truth as well as a variety of benchmarks.

2 Feature Relevance

Assume a mapping $f : \mathbb{R}^d \to Y$ is given such as a classifier with Y being the set of possible output classes. One very important way of interpreting any such mapping is offered by a judgment of the *relevance* of the given features for this mapping, i.e. a question related to feature selection. Here, a crucial distinction is offered by the notion of *strongly relevant features*, i.e. features which are of central relevance for the mapping f and which cannot be skipped without loss of information, and *weakly relevant features*, i.e. features which are of relevance for the mapping, but could be substituted by alternatives [27]. Formally, a feature X_i is strongly relevant provided the output $f(\mathbf{x})$ depends on the feature X_i even if all other features are known. A feature is weakly relevant if there exists a subset S of the other features, such that the output $f(\mathbf{x})$ depends on X_i provided S is known, but the feature is not strongly relevant. In all other cases, the feature is *irrelevant*.

While strongly relevant features can be detected e.g. based on efficient estimates of the mutual information, weakly relevant features are harder to detect since they require the investigation of all subsets of features. In the context of a linear or generalized linear mapping f, a popular technology for feature selection is offered by lasso and variants [30]. Assume $f(\mathbf{x}) = \boldsymbol{\omega}^T\mathbf{x}$. Then lasso relies on a L1-regularized optimization of the mapping parameters

$$\min \frac{1}{2} \cdot \sum_{i=1}^{n} (f(\mathbf{x}^i) - y_i)^2 \text{ such that } \sum_{j=1}^{d} |\omega_j| \leq s \tag{1}$$

for a sparsity constant $s > 0$. The constraint can be integrated into the objective as a penalty term with a fixed weighting. Ridge regression penalizes the L2 instead of L1 norm, and elastic net addresses a mixture

of both objectives [30], where, depending on the weighting of the penalty, different degrees of sparsity can be enforced. It is possible to infer the relevance of a given feature X_i from the size of the resulting weight $|\omega_i|$, whereby a varying penalty also can shed some light on the question whether the feature is weakly / strongly relevant.

Inspired by this observation, we will use L1 regularization for the valid interpretation of relevance terms of a given mapping. Thereby, we separate the question of how to train the mapping and how to interpret the feature relevance. This strategy, together with a slight reformulation of the regularization, enables us to derive intervals for the possible relevance range of a given feature. The first step is to reduce the problem of feature relevance determination for a metric learner to the equivalent question for a linear data transformation.

3 Metric learning as linear data transformation

Metric learning has been introduced in distance based machine learning models as a means to autonomously adjust the underlying distance measure to the given task at hand [3, 6]. Here, we focus on two popular metric learning schemes only. Since the proposed technique for metric interpretation is separated from the metric learning step itself, the proposed regularization for feature relevance determination can be used for every metric adjustment scheme which arrives at a general quadratic form, as utilized in the following.

We rely on a distance measure which is given by a general quadratic form

$$d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, (\mathbf{x}^i, \mathbf{x}^j) \mapsto (\mathbf{x}^i - \mathbf{x}^j)^T \mathbf{\Lambda} (\mathbf{x}^i - \mathbf{x}^j) \tag{2}$$

with the positive semi-definite matrix $\mathbf{\Lambda} = \mathbf{\Omega}^T \mathbf{\Omega}$. Large margin nearest neighbor (LMNN) [31] adjusts this matrix in such a way that the k-NN error induced by this distance is optimized. More precisely, it fixes the k nearest neighbors for every given data point, and adjusts the matrix $\mathbf{\Omega}$ such that points with the same label in this neighborhood are close, while points with a different label are separated by a distance term with a margin at least one. Generalized matrix learning vector quantization (GMLVQ) relies on a prototype-based winner-takes-all scheme rather than lazy learning [9]. Together with the prototype locations, the matrix $\mathbf{\Omega}$ is adjusted such that the distance of a given data point with correct labeling versus its distance to a prototype with incorrect labeling is minimized.

By optimizing $\mathbf{\Omega}$, both methods do not only increase the classification accuracy, but they also offer an interpretation of the feature relevance in terms of its diagonal $\Lambda_{ii} = \sum_j \Omega_{ji}^2$ or related terms, since the metric (2) corresponds to the linear data transformation

$$\mathbf{x} \mapsto \mathbf{\Omega} \mathbf{x}. \tag{3}$$

Hence interpretation of the matrix relevance terms reduces to the interpretation of this linear mapping.

In the following, we will employ the unique eigendecomposition of the matrix $\mathbf{\Lambda}$ as the linear data mapping $\mathbf{\Omega}$, i.e. the eigenvectors scaled with the square root of the eigenvalues.

4 Linear Bounds

Given the parameters $\mathbf{\Omega}$ that define a linear mapping $\mathbf{\Omega x}$ of a general quadratic form (2), we are interested in the interpretation of the mapping parameters $\mathbf{\Omega}$. First, we decompose the problem into one-dimensional mappings based on the following observation: Each row $\boldsymbol{\omega}$ of $\mathbf{\Omega}$ constitutes an independent mapping of the data into a one-dimensional subspace. Hence we can interpret each of these rows independently. After having obtained relevance bounds for the individual mappings $\boldsymbol{\omega}$, we can sum the absolute values of them in order to obtain relevance bounds for the whole mapping $\mathbf{\Omega}$.

In a particular mapping, the parameter value $|\omega_j|$ is often directly interpreted as the relevance of feature X_j, provided the input features have the same scaling. However, this can be highly problematic, as pointed out in [2], because features in high-dimensional data are often correlated, and thus the absolute value of ω_j can be misleading. In [2], the authors formalize mapping invariances for the given data to underline this observation, which we will briefly recap here.

For given data vectors $\mathbf{x}_i$ in a matrix $\mathbf{X}$, the central notion of invariance is defined as follows: Given a mapping $f(\mathbf{x}) = \boldsymbol{\omega}^\top \mathbf{x}$, we define that the parameters $\boldsymbol{\omega}$ are *equivalent* to $\boldsymbol{\omega}'$ iff

$$\boldsymbol{\omega}^\top \mathbf{X} = (\boldsymbol{\omega}')^\top \mathbf{X} \tag{4}$$

i.e. the data mapping remains unchanged under a substitution of $\boldsymbol{\omega}$ by $\boldsymbol{\omega}'$. Note that this formulation addresses the behavior of the mapping for given data, without referring to a predefined criterion, such as the accuracy. The conditions for which $\boldsymbol{\omega}$ is equivalent to $\boldsymbol{\omega}'$ are exactly described in [2]: two vectors $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ are equivalent iff the difference vector $\boldsymbol{\omega} - \boldsymbol{\omega}'$ is contained in the null space of the data covariance matrix $\mathbf{X}\mathbf{X}^\top$. The covariance matrix has eigenvectors $\mathbf{v}_i$ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_I > \lambda_{I+1} = \ldots = \lambda_d = 0$ sorted according to their size, whereby I denotes the number of non zero eigenvalues.

Therefore, before interpreting the mapping parameters, the proposal in [2] is to choose one canonic representation $\boldsymbol{\omega}'$ of the equivalence class induced by a given $\boldsymbol{\omega}$: the vector $\boldsymbol{\omega}'$ results by dividing out the null space, such that $\boldsymbol{\omega}$ becomes $\boldsymbol{\omega}' = \Psi\boldsymbol{\omega}$ where the matrix

$$\Psi = \mathrm{Id} - \sum_{i=I+1}^{d} \mathbf{v}_i \mathbf{v}_i^\top$$

corresponds to a projection of $\boldsymbol{\omega}$ to the eigenvectors with nonzero eigenvalues only, induced by the eigenvectors $\mathbf{v}_i$ of the covariance matrix $\mathbf{X}\mathbf{X}^\top$. Hence, the eigenvectors with eigenvalue zero are divided out. In [2], the authors show that choosing a representative in this way corresponds to a vector in the equivalence class with the smallest L_2 norm.

It is therefore no longer possible to ascribe a misleading high value ω_j to an irrelevant feature, e.g. due to unfavorable effects in the data. Instead, only relevant features are

expressed in the mapping parameters. Although this approach provides a unique representative of every equivalence class, it is problematic regarding direct interpretability of the values: Sets of correlated features share their total relevance, which can lead to large groups of weakly expressed relevance. This is undesirable, since a single feature may be weighted consistently low, only because it is highly correlated to a large number of others, despite the fact that the information provided by this feature (or any equivalent one) might be of high relevance for the linear mapping outcome.

Hence, we propose an alternative approach to determine representatives which are equivalent to a certain parameter vector $\boldsymbol{\omega}$, while allowing for an intuitive direct interpretation of the weights as feature relevances. In essence, instead of choosing the representative with smallest L_2 norm, we will use the L_1 norm. Unlike the former, the latter induces a set of equivalent weights which have minimal L_1 norm. This is beneficial, since we can infer the minimum and maximum relevance of each feature by looking at the minimum and maximum weighting of the feature within this set. In the following, we will formalize this concept.

4.1 Formalizing the Objective

Given a parameter vector $\boldsymbol{\omega}$ of a mapping, we are looking for equivalent vectors of the form

$$\boldsymbol{\omega}' = \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \tag{5}$$

where real-valued parameters α_i add the null space of the mapping to the vector $\boldsymbol{\omega}$. Similar to the approach in [2], we choose minimum vectors only, in order to avoid arbitrary scaling effects of the null space. However, we use the L_1 norm instead of the L_2 norm:

$$\mu \leftarrow \min_{\boldsymbol{\alpha}} \left\| \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \right\|_1 . \tag{6}$$

The minimum value μ is unique per definition. However, uniqueness of the corresponding vector $\boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i$ is not guaranteed. To illustrate this fact, we refer to a simple observation: assume identical features $X_i = X_j$ and a weighting ω_i and ω_j. Then any weighting $\omega_i' = t \cdot \omega_i + (1-t)\omega_j$ and $\omega_j' = (1-t)\omega_i + t\omega_j$ yields an equivalent vector with the same L_1 norm.

Based on this observation, we can formalize a notion of minimum and maximum feature relevance for a given linear mapping: the *minimum feature relevance* of feature X_j is the smallest value of a weight $|\omega_j'|$ such that $\boldsymbol{\omega}'$ is equivalent to $\boldsymbol{\omega}$ and $|\boldsymbol{\omega}'|_1 = \mu$. Analogously, the *maximum feature relevance* of feature X_j is the largest value of a weight $|\omega_j'|$ such that $\boldsymbol{\omega}'$ is equivalent to $\boldsymbol{\omega}$ and $|\boldsymbol{\omega}'|_1 = \mu$. Thus, we arrive at the following mathematical optimization problems:

$$\underline{\omega}_j \leftarrow \min_{\boldsymbol{\alpha}} \left| \omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right| \tag{7}$$

$$\text{s.t.} \quad \left\| \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \right\|_1 = \mu$$

and

$$\overline{\omega}_j \leftarrow \max_{\boldsymbol{\alpha}} \left| \omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right| \qquad (8)$$
$$\text{s.t.} \quad \left\| \boldsymbol{\omega} + \sum_{i=I+1}^{d} \alpha_i \mathbf{v}_i \right\|_1 = \mu.$$

where $(\mathbf{v}_i)_j$ refers to entry j of $\mathbf{v}_i$. As solutions, we obtain a pair $(\underline{\omega}_j, \overline{\omega}_j)$ for each feature X_j indicating the minimum and maximum weight of this feature for all equivalent mappings that share the same L_1 norm. In the special case of linear mappings with the objective of mapping invariance, this strongly resembles the concept of strong and weak feature relevance.

However, this framework does not realize the notion of strong and weak feature relevance in a strict sense. The reason is that we aim for scaling terms as observed in the linear mapping, which are subject to L_1 regularization. Consequently, a set of two features with the same information content as a single feature are not treated as identical by this formulation. Instead, our formulation prefers the single feature because of the smaller scaling of the respective weight. A qualitative feature selection objective would treat such variables identically.

Natural relaxations of our optimization problems are possible, as follows: By incorporating also eigenvectors which correspond to small eigenvalues in Eq. (5), we can enable an approximate preservation of mapping equivalence. Further, we can approximate the equality in Eq. (6) by allowing values below $(1+\epsilon)\mu$ instead of exactly μ, for some small $\epsilon > 0$. Such relaxations with small thresholds ϵ are strongly recommended for practical applications, where noise in the data has to be considered. We will refer to these straightforward approximations in our experiments.

4.2 Reformalization as Linear Programming Problem

To obtain a solution algorithmically, we reformulate our optimization problems as linear programming problems (LP). Problem (7) can be rephrased as the following equivalent LP, in which we introduce a new variable $\tilde{\omega}_k$ for every k, which takes the role of $|\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k|$:

$$\underline{\omega}_j \leftarrow \min_{\tilde{\boldsymbol{\omega}}, \boldsymbol{\alpha}} \tilde{\omega}_j, \qquad (9)$$
$$\text{s.t.} \quad \sum_{i=1}^{d} \tilde{\omega}_i \leq \mu$$
$$\tilde{\omega}_k \geq \omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k, \forall k$$
$$\tilde{\omega}_k \geq -\left(\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k \right), \forall k,$$

where μ is computed in (6) and the variables $\tilde{\omega}_i$ must be non-negative due to the constraints. For the optimum solution, we can assume that equality holds for one of the two constraints for every k. Otherwise, the solution could be improved due to the weaker constraints and the minimization of the objective. To rephrase problem (8), we use the

equivalent formulation

$$\max_{\tilde{\boldsymbol{\omega}},\boldsymbol{\alpha}} \left| \omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right|, \tag{10}$$

$$\text{s.t.} \quad \sum_{i=1}^{d} \tilde{\omega}_i \leq \mu$$

$$\tilde{\omega}_k \geq \omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k, \forall k$$

$$\tilde{\omega}_k \geq -\left(\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k \right), \forall k,$$

where new variables $\tilde{\omega}_k$ are introduced as well. Again, these serve as the absolute value $|\omega_k + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_k|$: any solution for which equality does not hold for one of the constraints can be improved due to the weaker constraints and maximization as the objective. Since an absolute value is optimized, this is not an LP yet. To obtain its solution, one can simply solve two LPs, where the positive and negative value of the objective is considered, respectively:

$$\overline{\omega}_j^{\pm} \leftarrow \max_{\tilde{\boldsymbol{\omega}},\boldsymbol{\alpha}} \pm \left(\omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right),$$

and we add the corresponding non-negativity constraint

$$\pm \left(\omega_j + \sum_{i=I+1}^{d} \alpha_i (\mathbf{v}_i)_j \right) \geq 0$$

At least one of these LPs has a feasible solution, and the final upper bound can be derived thereof as the maximum value

$$\overline{\omega}_j = \max\{\overline{\omega}_j^{+}, \overline{\omega}_j^{-}\}$$

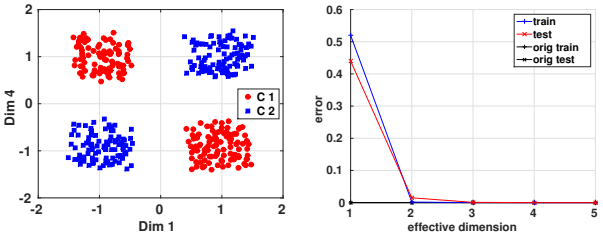

Figure 1: Two relevant features of the xor data set (left). Average classification error rates of GMLVQ with regularized metrics for the xor data set (right).

For each linear mapping, this formulation requires to solve $3d$ LP problems containing $2d$ constraints and $d - I$ variables. For this purpose, standard solvers can be applied.

5 Experiments

In this section we apply our proposed methods to four data sets from different domains. After describing the data, we explain the experimental setup and, finally, depict the results. For the evaluation, we employ the following data sets.

- The xor data set is artificially generated and consists of 4 clusters belonging to 2 classes constituting the XOR problem. One dimension is present 3 times with the addition of noise (features 1-3) and two identical irrelevant features are included (5-6). An image with features 1 and 4 is depicted in Fig. 1 (left).

- The wine data set consists of 256 features which are near-infrared spectra measur-

ing the alcohol content of 124 wine samples [32]. The set is split into 94 training and 30 test samples, where samples number 34,35 and 84 are discarded as outliers, similar to [33]. Additionally, we switch the role of training and test set to obtain a more challenging problem in terms of interpretation. Since this is originally a regression problem, we transform it into a classification problem by binning alcohol levels into 3 classes of similar size.

- The tecator data set [34] consists of 100 features that represent absorbances deduced from a spectometer. The goal is to predict the fat content of 215 meat samples. The set is split into 172 samples for training and 43 samples for evaluation, where we again switch the role of training and test set. Similarly as for the wine data, we bin the target variable into three classes to obtain a classification problem.

- The adrenal data set [16, 35] consists of 147 patients characterized by 32 steroid markers. The goal is to predict whether a patient has a benign or a malignant adrenal tumor. For this purposes, we split the data set into a training set containing 110 instances for training and 37 instances for evaluation.

5.1 Experimental setup

As a pre-processing step, we apply a z-score transformation to all our data sets, by removing the mean and standard deviation of the training data from each data set. It is important that the features in each data set have the same scaling so that we can interpret the weights of linear mappings.

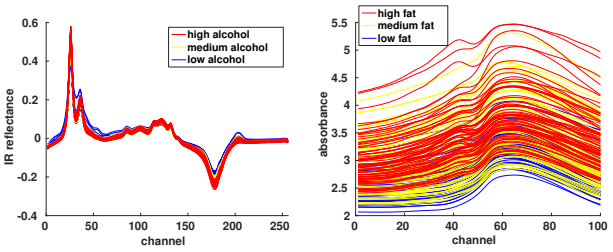

Figure 2: Spectra of the data sets wine (left) and tecator (right).

We train the GMLVQ model always using one prototype per class, except for the xor data set, where we use two. For the LMNN model, we use the parameters suggested by a parameter search procedure provided by the original authors.

A crucial parameter in our framework is the size of the assumed null space of the data. In order to obtain a sensible choice for this parameter, we first train a metric learning algorithm and then utilize the following scheme:

1. Create a set S of candidate values for the size of the null space. This can simply be all possible values, or a guess based on the eigenspectrum of the data.

2. For each element in S, apply our proposed interpretation framework to the previously learned metric, resulting in $2d$ relevance mappings for each row of the trained metric.

3. Compute the classification accuracy on the train and test set for each of these $2d$ mappings and average them. Select the size of the null space as the one with a small test error along with the largest null space.

We also employ the term 'regularized relevance profile' when we refer to the resulting relevance bounds of our approach.

Since the null space is often large, we will recall in the following the size of the *effective dimension* which is the number of dimensions minus the size of the null space. Additionally, due to noise, we soften the minimum norm conditions in (9) and (10) by allowing solutions smaller then $1.01 \cdot \mu$, in the following.

As concerns the complexity of the metric learning scheme for high-dimensional data, the computation of a full rank matrix $\mathbf{\Lambda} \in \mathbb{R}^{d \times d}$ can be costly. However, $\mathbf{\Lambda}$ can be forced to have a low rank [5]. This can be done by defining $\mathbf{\Lambda} = \mathbf{\Omega}^T \mathbf{\Omega}$ with $\mathbf{\Omega} \in \mathbb{R}^{l \times d}$, where $l \leq d$ restricts the rank.

5.2 Synthetic Data

In order to demonstrate the problem of directly interpreting linear weights of a trained metric as relevances, we employ the synthetic data set xor.

We train a GMLVQ method that results in a zero prediction error on the training and test set. The resulting three mappings of the metric with the largest scaling are depicted in the first row of Fig. 3. Basically, only one of these mappings has a high scaling so that the classification model uses approximately a one-dimensional subspace to solve the classification task.

A direct interpretation of this linear vector $|\boldsymbol{\omega}^3|$ would suggest that feature 4 is the most important one, features 1 and 3 have only half the relevance and features 2,5 and 6 are not useful for the task. However, for this data set we know that features 1-3 have the same explanatory power and if considered alone, each of them is as important as feature 4. It follows that, for this example, a direct interpretation of the linear weights is misleading, particularly for the weakly relevant features.

In order to obtain a valid interpretation for the relevance of the features, we apply our proposed framework. We estimate the classification accuracy of the regularized mappings for all possible sizes of the null space as described in subsection 5.1. The resulting curves are depicted in Fig. 1 (right). It is apparent from the Figure, that the smallest effective dimension size with a zero test error is 3, although the test error for 2 dimensions is only slightly larger. Nevertheless, we employ 3 for our proposed framework. The resulting relevance bounds are shown in the second row of Fig. 3, where black bars depict weakly and white bars strongly relevant features.

The results show that the bounds for the first two one-dimensional mappings $|\boldsymbol{\omega}^1|$ and $|\boldsymbol{\omega}^2|$ have vanished. Formally speaking, this implies that the same mappings can be obtain with an almost zero L1 norm, meaning that these two mappings map the training data to zero. More interestingly are the resulting bounds for $|\boldsymbol{\omega}^3|$: The framework has identified feature 4 as a strongly relevant feature and has found that features 1-3 can be re-

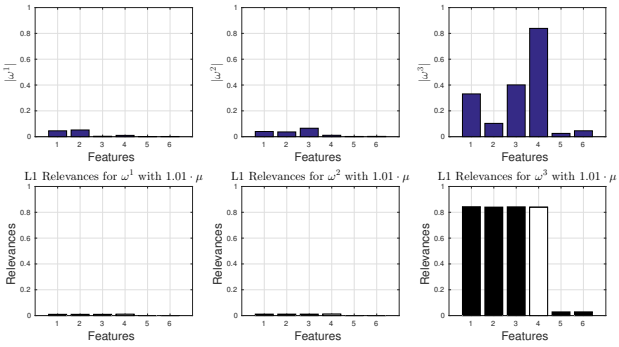

Figure 3: Results of our proposed approach for the xor data set. The first row shows the original linear mappings, the second row depicts the resulting upper (in black) and lower bounds (in white).

placed but that each of them can explain as much of the target variable as feature 4. This explanation is precisely how we generated the data. Features 5 and 6 have almost 0 upper bounds, merely reflecting noise.

In order to have a comparison to relevance interpretation in literature, we apply the methods Lasso, Elastic Net and Ridge Regression to our resulting mapping by defining $\hat{y}_i = \boldsymbol{\omega}^\top \mathbf{x}_i$. Then, we can apply the formulation in equation (1) for Lasso and according ones for Elastic Net and Ridge Regression to obtain an interpretation for the feature weights based on these methods. The results are depicted in Fig. 4 for the Lasso (left), Elastic Net (middle) and Ridge Regression (right). We interpret only $|\boldsymbol{\omega}^3|$ this way, since, as we saw previously, this mapping contains the relevant information for discriminating the classes.

The progress of the coefficient weights for

Table 1: Classification error rates ranging between 0 and 1 for all data sets. If not specified differently, the classification model is GMLVQ.

	xor	wine	tecator	GMLVQ on adrenal	LMNN on adrenal
train er	0.00	0.00	0.07	0.04	0.00
test er	0.00	0.29	0.16	0.03	0.05

all three frameworks implies that feature 4 is particularly relevant. However, the results differ for the three weakly relevant features 1-3: Elastic net and Ridge regression require all three features equally weighted and hence do not show that each of them can actually be neglected. The Lasso identifies feature 1 as particularly important, followed by feature 3 and 2. This order seems arbitrary and is an artifact of the noise which was added to all three features. Hence, we argue that these frameworks cannot provide the same information as our formalization.

5.3 Near-Infrared spectral data

Spectral data have many correlated features and hence a large null space. For such data, it can be particularly misleading to directly interpret the weights of linear mappings. Hence, our method should be well suited in this case.

First of all we train a GMLVQ model for each of the two data sets wine and tecator where we restrict the rank of the matrix $\boldsymbol{\Omega}$ to two since GMLVQ tends to utilize only a low rank matrix in the end, usually. This does not harm the training accuracy as can be observed in Table 1 but tends to improve the generalization. Subsequently, we apply

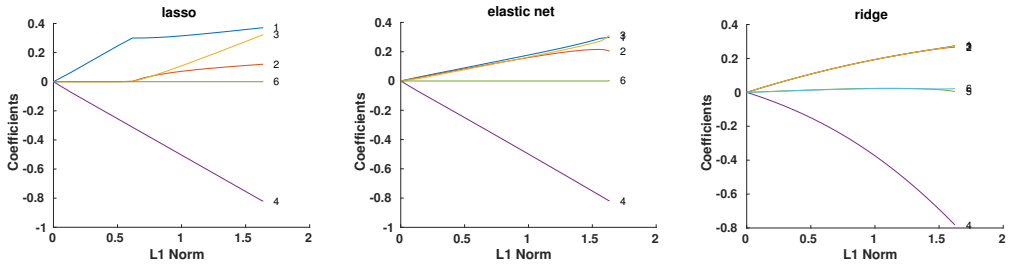

Figure 4: Employing the xor data set, estimates of the coefficients for different values of the L1 norm (x-axis) are shown. The methods lasso (left), elastic net (middle) and ridge regression (right) are utilized.

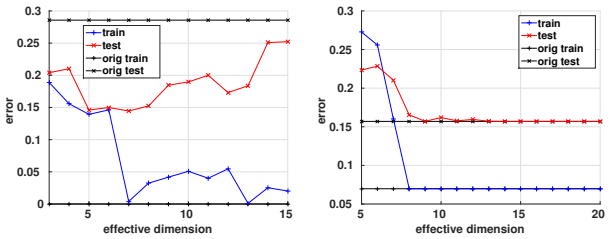

Figure 5: Average classification error rates of GMLVQ with regularized metrics for the wine (left) and tecator (right) data set, both for set S.

our approach to compute relevance bounds for the according learned metric. As previously, we determine a suitable size of the effective dimension using the scheme described in subsection 5.1. The corresponding images are shown in Fig. 5: left for the wine and right for the tecator data set.

The smallest test error is obtained with an effective dimensionality of 7 for the wine data set and with an effective dimensionality of 9 for the tecator data set. It is particularly interesting, that for the wine data set the regularized metric achieves a better performance then the original metric: while the training error stays the same, the test error drops from 0.29 to 0.14, which is a factor 2. The resulting relevance bounds for the wine data set are depicted in Fig. 6 and for the tecator data set in Fig. 7. For the wine data, the training procedure of the classifier yielded a rank one matrix, hence we use only a one-dimensional mapping for interpretation, in this case. For the tecator data set, both mappings are utilized, and the resulting relevance bounds for both mappings are displayed in Fig. 8.

Interestingly, the relevance bounds for both data sets contain only very few irreplaceable features, while the upper bounds are peaked, which implies that a few features can already explain the mapping to a large extent. Particularly for the wine data set, much noise is removed from the original mapping, i.e. many features have a low upper bound. Fig. 9 displays the original mapping (top) and the averaged mapping over all $\overline{\omega}, \underline{\omega}$ (bottom). Here it is apparent that, while the original mapping has many non-zero values, the averaged regularized profile is extremely

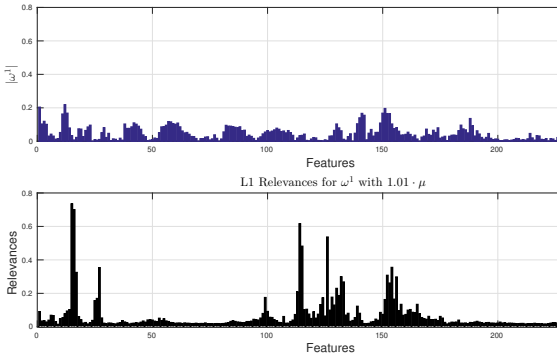

Figure 6: Results of our proposed approach for the wine data set. The first row shows the original linear mapping, while the second row depicts the resulting upper relevance bounds. The lower bounds are all zero, in this case.

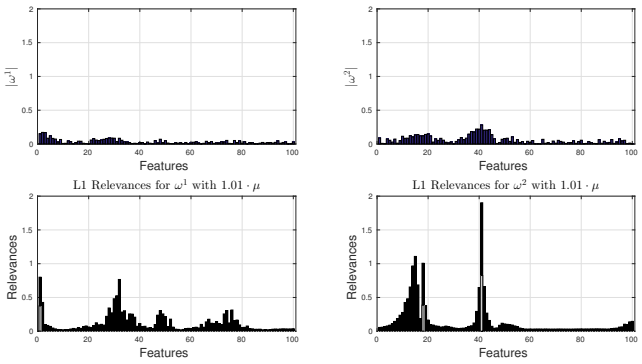

Figure 7: Results of our proposed approach for the tecator data set. The first row shows the original linear mapping, the second row depicts the resulting upper and lower relevance bounds.

sparse. Furthermore, the classification error with the averaged map accounts to 0 on the training and to 0.14 on the test set, which is comparable to the averaged error of the regularized mappings $\overline{\omega}, \underline{\omega}$.

5.4 Biomedical data

For the adrenal data set we compare two metric learning approaches: The GMLVQ and LMNN. Both use the same functional form for computing distances, hence, we can apply our approach to both trained relevance matrices. First, we train both models with the restriction to rank two relevance matrices. This restriction did not harm the classification performance in our experiments, as compared to training without this restriction. The classification errors are depicted in Table 1. Both approaches achieve a comparable performance, where the LMNN model is better on the training data while GMLVQ is superior on the test data.

For these models, we compute the classification errors based on different sizes of the effective dimension. These results are depicted in Fig. 10. Good performances are achieved with an effective dimensionality of 15 for the GMLVQ model and of 20 for the LMNN algorithm. The according relevance bounds for both relevance matrices are shown in Fig. 11.

Both trained distance metrics agree on a few features, such as 19, while they emphasize often different ones. This might explain the different error curves in Fig. 10 and the different classification performance on the training and test set.

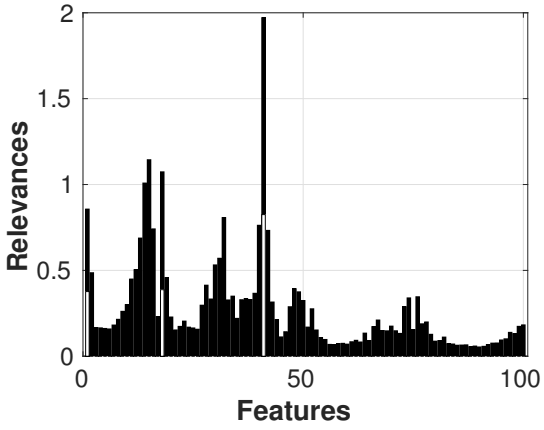

Figure 8: Summed lower and upper bounds for the tecator data set.

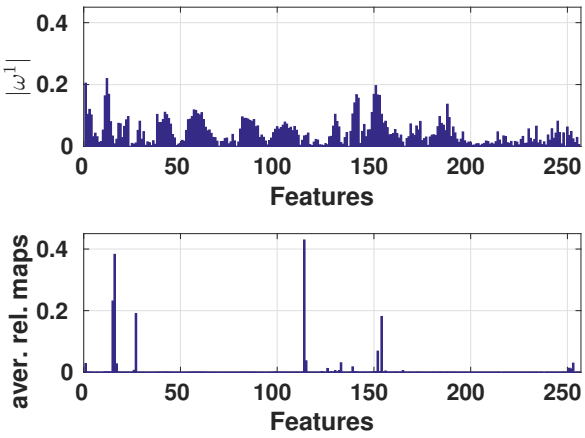

Figure 9: Absolute values of the original mapping (top row) together with the absolute value of the averaged regularized mappings (bottom row).

6 Conclusion

In this contribution, we present an approach to obtain a valid interpretation of the feature relevance from a trained metric learning model. The results show that this procedure does not only provide a meaningful interpretation of the learned data transformation, but it can even improve the classification performance, in cases of a particular large null space.

In future work we will employ this proposal to obtaining insights into specific data sets, as well as to provide more information about weakly relevant features.

Acknowledgment

Funding from DFG under grant number HA2719/7-1 and by the CITEC center of excellence is gratefully acknowledged.

References

[1] A. Schulz, B. Mokbel, M. Biehl, and B. Hammer, "Inferring feature relevances from metric learning," in *2015 IEEE Symposium Series on Computational Intelligence*, Dec 2015, pp. 1599–1606.

[2] M. Strickert, B. Hammer, T. Villmann, and M. Biehl, "Regularization and improved interpretation of linear data mappings and adaptive distance measures," in *IEEE SSCI CIDM 2013*. IEEE Computational Intelligence Society, 2013, pp. 10–17.

[3] A. Bellet, A. Habrard, and M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data," *ArXiv e-prints*, Jun. 2013.

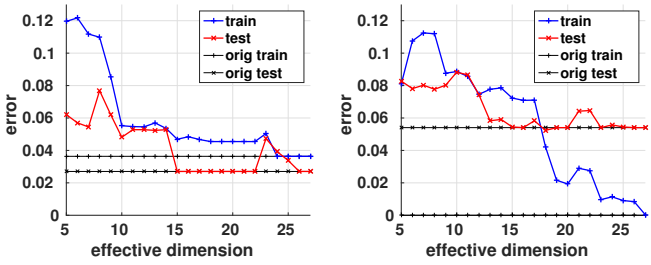

Figure 10: Average classification error rates of GMLVQ (left) and LMNN (right) with regularized metrics for the adrenal data set.

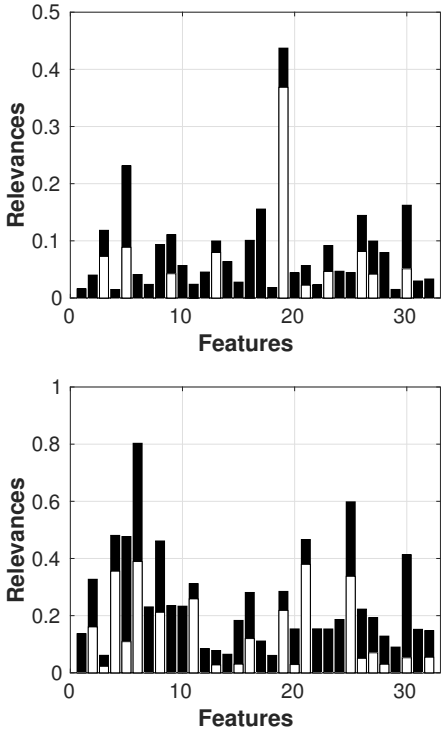

Figure 11: Relevance bounds for a GMLVQ model (top) and a LMNN model (bottom), both trained on the adrenal data set.

[4] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. IEEE Computer Society, 2012, pp. 2288–2295.

[5] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl, "Limited rank matrix learning, discriminative dimension reduction and visualization," *Neural Networks*, vol. 26, pp. 159–173, 2012.

[6] B. Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[7] A. Takasu, D. Fukagawa, and T. Akutsu, "Statistical learning algorithm for tree similarity," in *IEEE Int. Conf. on Data Mining, ICDM*, 2007, pp. 667–672.

[8] S. S. Bucak, R. Jin, and A. K. Jain, "Multiple kernel learning for visual object recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1354–1369, 2014.

[9] P. Schneider, M. Biehl, and B. Hammer, "Adaptive relevance matrices in learning vector quantization," *Neural Computation*, vol. 21, pp. 3532–3561, 2009.

[10] A. Bellet and A. Habrard, "Robustness and generalization for metric learning,"

Neurocomputing, vol. 151, pp. 259–267, 2015.

[11] R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning: Theory and algorithm," in *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 862–870.

[12] Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds., *Advances in Neural Information Processing Systems 22. Proceedings from 7-10 December 2009, Vancouver, British Columbia, Canada.* Curran Associates, Inc., 2009.

[13] B. Arnonkijpanich, A. Hasenfuss, and B. Hammer, "Local matrix adaptation in topographic neural maps," *Neurocomputing*, vol. 74, no. 4, pp. 522–539, 2011.

[14] M. Biehl, B. Hammer, F. Schleif, P. Schneider, and T. Villmann, "Stationarity of matrix relevance LVQ," in *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-17, 2015.* IEEE, 2015, pp. 1–8.

[15] B. Mokbel, B. Paassen, F.-M. Schleif, and B. Hammer, "Metric learning for sequences in relational {LVQ}," *Neurocomputing*, no. to appear, pp. –, 2015.

[16] W. Arlt, M. Biehl, A. E. Taylor, S. Hahner, R. Libe, B. A. Hughes, P. Schneider, D. J. Smith, H. Stiekema, N. Krone, E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. H. L. Shackleton, X. Bertagna, M. Fassnacht, and P. M. Stewart, "Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors," *J Clinical Endocrinology and Metabolism*, vol. 96, pp. 3775–3784, 2011.

[17] G. de Vries, S. C. Pauws, and M. Biehl, "Insightful stress detection from physiology modalities using learning vector quantization," *Neurocomputing*, vol. 151, pp. 873–882, 2015.

[18] F.-M. Schleif, B. Hammer, M. Kostrzewa, and T. Villmann, "Exploration of mass-spectrometric data in clinical proteomics using learning vector quantization methods," *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 129–143, 2008.

[19] A. A. Freitas, "Comprehensible classification models: A position paper," *SIGKDD Explor. Newsl.*, vol. 15, no. 1, pp. 1–10, Mar. 2014.

[20] V. V. Belle and P. Lisboa, "White box radial basis function classifiers with

component selection for clinical prediction models," *Artificial Intelligence in Medicine*, vol. 60, no. 1, pp. 53–64, 2014.

[21] C. Rudin and K. L. Wagstaff, "Machine learning for science and society," *Machine Learning*, vol. 95, no. 1, pp. 1–9, 2014.

[22] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher, "Pitfalls of predicting complex traits from SNPs." *Nature reviews. Genetics*, vol. 14, no. 7, pp. 507–515, Jul. 2013.

[23] P. Langley, "The changing science of machine learning." *Machine Learning*, vol. 82, no. 3, pp. 275–279, 2011.

[24] P. J. G. Lisboa, "Interpretability in machine learning - principles and practice," in *WILF*, ser. Lecture Notes in Computer Science, F. Masulli, G. Pasi, and R. R. Yager, Eds., vol. 8256. Springer, 2013, pp. 15–21.

[25] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time," *Journal of Machine Learning Research*, vol. 8, pp. 589–612, 2007.

[26] B. Frénay, G. Doquire, and M. Verleysen, "Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification," *Neurocomputing*, vol. 112, pp. 64–78, 2013.

[27] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[29] B. Frénay, D. Hofmann, A. Schulz, M. Biehl, and B. Hammer, "Valid interpretation of feature relevance for linear data mappings," in *2014 IEEE SSCI, CIDM 2014, Orlando, FL, USA, December 9-12, 2014*, 2014, pp. 149–156.

[30] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2 2010.

[31] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[32] UCL, "Spectral wine database," 2007, provided by Prof. Marc Meurens.

[33] C. Krier, D. François, F. Rossi, and M. Verleysen, "Feature clustering and mutual information for the selection of variables in spectral data," in *15th ESANN 2007, Bruges, Belgium, April*

25-27, 2007, Proceedings, 2007, pp. 157–162.

[34] "Tecator meat sample dataset."

[35] M. Biehl, P. Schneider, D. Smith, H. Stiekema, A. Taylor, B. Hughes, C. Shackleton, P. Stewart, and W. Arlt, "Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors," in *20th European Symposium on Artificial Neural Networks, ESANN 2012, Bruges, Belgium, April 25-27, 2012*, 2012.