ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Improved Interpretation of Feature Relevances: Iterated Relevance Matrix Analysis (IRMA)

Sofie Lövdal[1,2] and Michael Biehl[1] *

1- Univ. of Groningen - Bernoulli Inst. for Mathematics, Computer Science and Artificial Intelligence, Nijenborgh 9, 9747 AG Groningen - The Netherlands

2- University Medical Center Groningen (UMCG), Dept. of Nuclear Medicine and Molecular Imaging, Hanzeplein 1, 9713 GZ Groningen -The Netherlands

Abstract. We introduce and investigate the iterated application of Generalized Matrix Relevance Learning for the analysis of feature relevances in classification problems. The suggested Iterated Relevance Matrix Analysis (IRMA), identifies a linear subspace representing the classification specific information of the considered data sets in feature space using Generalized Matrix Learning Vector Quantization. By iteratively determining a new discriminative direction while projecting out all previously identified ones, all features carrying relevant information about the classification can be found, facilitating a detailed analysis of feature relevances. Moreover, IRMA can be used to generate improved low-dimensional representations and visualizations of labeled data sets.

1 Introduction

Prototype-based systems such as Learning Vector Quantization (LVQ) [1, 2] can serve as genuinely interpretable and transparent classification tools [3]. In combination with the use of adaptive distance measures [4, 5], they provide valuable insights into the structure of the problem at hand and into the relevance of features for the actual classification task. However, the presence of correlated features or multiple sets of features providing similar performance can lead to ambiguous relevance assignments and non-unique outcome of training. This frequently complicates the interpretation of relevance learning, see e.g. [6, 7].

Here, we suggest and study an extension of Generalized Matrix LVQ (GMLVQ) [4, 5]. We show that the successive removal of dominantly relevant directions in feature space and subsequent re-training of GMLVQ with the remaining information allows to infer the most class-relevant subspace. This *Iterated Relevance Matrix Analysis* (IRMA) facilitates the detailed analysis of feature relevances - especially in presence of multiple weakly relevant features. Moreover, the discriminative low-dim. representation and visualization of labeled data sets can be enhanced compared with the basic GMLVQ approach [4, 5]. A similar approach has been considered earlier for Support Vector Machines [11]. In Section 2 we introduce the suggested procedure. The illustrative application of IRMA to example artificial and benchmark data sets is presented in Sec. 3 and we summarize and discuss possible extensions in Sec. 4.

*S.L. acknowledges support by Stichting ParkinsonFonds.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2 Iterated Relevance Matrix Learning

An LVQ system assigns N-dim. feature vectors $\mathbf{x} \in \mathbb{R}^N$ to one of C classes labeled by $S \in \{1, 2, \ldots, C\}$. The *nearest prototype* classification is based on the distances of $\mathbf{x}$ from a set of M prototypes $\{\mathbf{w}_j \in \mathbb{R}^N\}_{j=1}^M$. Each prototype represents one of C classes as denoted by the labels $S(\mathbf{w}_j) \in \{1, 2, \ldots, C\}$.

GMLVQ in its basic variant [4] employs a global distance measure of the form

$$d(\mathbf{w}_j, \mathbf{x}) = (\mathbf{x} - \mathbf{w}_j)^\top \Lambda (\mathbf{x} - \mathbf{w}_j), \quad \text{with } \Lambda = \Omega^\top \Omega. \tag{1}$$

Here, the relevance matrix $\Lambda \in \mathbb{R}^{N \times N}$ is re-parameterized in terms of an auxiliary matrix $\Omega \in \mathbb{R}^{N \times N}$ as to guarantee that Λ is symmetric and positive semi-definite with $d(\mathbf{w}_j, \mathbf{x}) \geq 0$. Extensions to local relevance matrices or rectangular Ω have been considered in the literature [4, 5].

Given a set of data $\{\mathbf{x}^\mu, S^\mu\}_{\mu=1}^P$, prototypes $\mathbf{w}_j$ and matrix Ω are optimized in a training process which is guided by the minimization of the cost function

$$E = \sum_{\mu=1}^P \phi \left[\frac{d^\Lambda(\mathbf{w}_+, \mathbf{x}^\mu) - d^\Lambda(\mathbf{w}_-, \mathbf{x}^\mu)}{d^\Lambda(\mathbf{w}_+, \mathbf{x}^\mu) + d^\Lambda(\mathbf{w}_-, \mathbf{x}^\mu)} \right], \quad \text{with } \phi(z) = z \text{ in the following.} \tag{2}$$

For a given example $\{\mathbf{x}^\mu, S^\mu\}$, $\mathbf{w}_+$ denotes the *closest correct* prototype with $d(\mathbf{w}_+, \mathbf{x}_i) \leq d(\mathbf{w}_j, \mathbf{x}^\mu)$ among all $\mathbf{w}_j$ with $S(\mathbf{w}_j) = S^\mu$. Correspondingly, $\mathbf{w}_-$ is the *closest wrong* prototype carrying a label different from S^μ. In practice, GMLVQ ensures that the data points are linearly mapped by Ω into a space where classes are separated as well as possible. An additional normalization of the form $\sum_{i=1}^N \Lambda_{ii} = \sum_{i=1}^N \Omega_{ij}^2 = 1$ is imposed in order to avoid numerical instabilities and support comparability [4]. The resulting diagonal entries Λ_{jj} quantify the relevance of dimension j, provided all features x_j are of the same magnitude [4]. Throughout the following we achieve this by applying a z-score transformation based on the actual set of training data.

The symmetric semi-definite relevance matrix can be expressed as:

$$\Lambda = \sum_{j=1}^N \lambda_j \mathbf{v}_j \mathbf{v}_j^\top, \text{ with } \Lambda \mathbf{v}_j = \lambda_j \mathbf{v}_j \text{ and } \Omega = \sum_{j=1}^N \sqrt{\lambda_j} \mathbf{v}_j \mathbf{v}_j^\top \tag{3}$$

being a canonical solution of $\Lambda = \Omega^\top \Omega$. We assume that eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N$. After training, the relevance matrix typically assumes a low rank and is dominated by a few leading eigenvectors, see [8] for a detailed discussion and analysis. This property facilitates e.g. the discriminative visualization of the data set in terms of projections onto the first eigenvectors [4, 5].

In two-class problems, for instance, the training typically identifies a single, most discriminative direction $\mathbf{v}_1^{(0)}$ with $\lambda_1^{(0)} \approx 1$ and $\Lambda^{(0)} \approx \mathbf{v}_1^{(0)} \mathbf{v}_1^{(0)\top}$. Here and in the following the superscript (0) refers to the results of a first, unrestricted GMLVQ training. In such a situation, the eigenvectors $\mathbf{v}_j^{(0)}$ with $j \geq 2$ form an arbitrary basis of the space orthogonal to $\mathbf{v}_1^{(0)}$ with no particular order. Note, however, that the corresponding $(N-1)$-dim. subspace very likely still contains relevant information about the classes, reflecting the potential ambiguity of the relevance assignment. The selection of a particular $\mathbf{v}_1^{(0)}$ may depend strongly on the actual training data set, possibly leading to an overfitted relevance analysis.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

In order to obtain more comprehensive insights, we can perform a second GMLVQ training process which is restricted to an orthogonal subspace by considering a distance measure of the form (1) with $\Lambda^{(1)} = \Omega^{(1)\top}\Omega^{(1)}$ under the constraint that $\Omega^{(1)}\mathbf{v}_1^{(0)} = 0$. This can be achieved by applying the projection $\Omega^{(1)} \rightarrow \Omega^{(1)}\,[I - \mathbf{v}_1^{(0)}\mathbf{v}_1^{(0)\top}]$ after each update step, followed by the normalization of $\Omega^{(1)}$. In other words, this projection ensures that contributions corresponding to $\mathbf{v}_1^{(0)}$ are disregarded in the feature space. Now, the leading eigenvector $\mathbf{v}_1^{(1)}$ of the resulting $\Lambda^{(1)}$ represents the most discriminative direction orthogonal to $\mathbf{v}_1^{(0)}$. The degree to which $\mathbf{v}_1^{(1)}$ carries class relevant information can be evaluated in terms of a performance measure of the restricted classifier, e.g. by the balanced accuracy $BAC^{(1)}$, estimated in an appropriate validation procedure. Obviously, we can apply the idea iteratively and obtain a sequence of vectors $\mathbf{v}_1^{(j)}$ each of which is orthogonal to all $\mathbf{v}_1^{(i)}$ with $i = 0, 1, \ldots, j-1$. In each step $j \geq 1$ of this *Iterated Relevance Matrix Analysis* (IRMA) we perform GMLVQ training where the projection

$$\Omega^{(j)} \rightarrow \Omega^{(j)} \left[I - \textstyle\sum_{i=0}^{j-1} \mathbf{v}_1^{(i)}\mathbf{v}_1^{(i)\top} \right] \tag{4}$$

is applied after each update together with the appropriate normalization. We will refer to the unrestricted GMLVQ training as the 0-*th iteration*. The key step (4) is reminiscent of the subspace correction in [9], where it however serves a different purpose.

The procedure can be terminated when the classifier in iteration $(k+1)$ achieves only random or near random classification performance as signaled by, for example, a $BAC^{(k+1)} \approx 0.5$. The obtained subspace

$$V = \text{span}\{\mathbf{v}_1^{(0)}, \mathbf{v}_1^{(1)}, \ldots \mathbf{v}_1^{(k)}\} \;\text{ with associated projections } y_i^\mu = \mathbf{x}^\mu \cdot \mathbf{v}_1^{(i)} \tag{5}$$

can be interpreted as to contain (approximately) all class relevant information in feature space. Hence, it can serve for further analysis of feature relevances. An obvious application could be the low-dim. representation of labeled data sets in terms of the y_i^μ, e.g. for the purpose of two- or three-dim. visualizations.

3 Illustrative applications

Artificial Data: We first consider an extremely simple and clear-cut artificial two-class data set illustrated in Fig. 1 (a). Feature vectors $\mathbf{x} \in \mathbb{R}^4$ comprise two informative components x_1, x_2 in which each class corresponds to an elongated Gaussian cluster. The remaining components are independently drawn from an isotropic zero mean, unit variance normal density. As can be seen in panel (a), feature x_2 should be sufficient to separate the classes with almost 100% accuracy. However, classes also separate along x_1, albeit less perfectly. Unrestricted GMLVQ with one prototype per class realizes near perfect classification with $BAC^{(0)} \approx 0.99$ (w.r.t. training and test) in a balanced data set of 600 samples, where the training set contains 150 randomly drawn examples and the remaining 450 form a test set. Projections on the leading eigenvectors are shown in panel

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

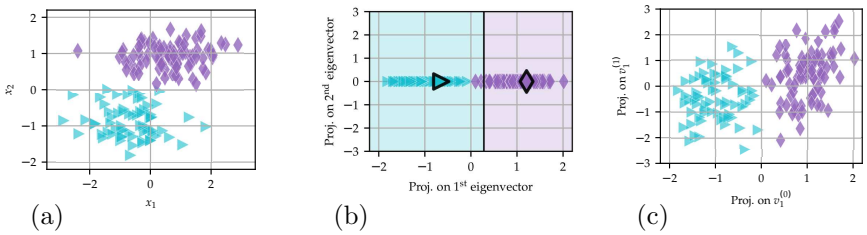

Fig. 1: Artificial data: original features x_1, x_2 of the data set (a), projections on $v_1^{(0)}$, $v_2^{(0)}$ of unrestricted GMLVQ (b), and projections on the eigenvectors $v_1^{(0)}$ and $v_1^{(1)}$ of the unrestricted system and the first iteration of IRMA in (c).

(b) of Fig. 1. The dominating eigenvector is $\mathbf{v}_1^{(0)} \approx (0.19, 0.98, 0.1, 0.05)^\top$ corresponding to $\Lambda_{jj}^{(0)} \approx \delta_{j,2}$. The orthogonal $\mathrm{v}_2^{(0)}$ is essentially random as indicated by the absence of a separation of classes, resulting in an effectively one-dim. visualization.

In the first IRMA iteration, the leading eigenvector of $\Lambda^{(1)}$ approaches the second relevant direction: $\mathbf{v}_1^{(1)} \approx (0.92, -0.16, -0.02, -0.36)^\top$ with $\Lambda_{jj}^{(1)} \approx \delta_{j,1}$. As expected, the performance drops compared to the unrestricted system: we observe a $BAC^{(1)}$ of 0.68 (training) and 0.70 (test). As shown in panel (c) of Fig. 1, the projections y_0, y_1, cf. Eq. (5), of the data set onto $\mathbf{v}_1^{(0)}$ and $\mathbf{v}_1^{(1)}$ display both relevant separating directions and reproduce the cluster structure of the original features x_1, x_2. Already in the second iteration of IRMA, the accuracy drops to $BAC^{(2)} \approx 0.54$ and 0.51 for training and test data, respectively. As expected, no further relevant directions can be identified.

Wisconsin Diagnostic Breast Cancer data: This benchmark data set from the UCI Machine Learning Repository [10] contains 569 samples with 30 features extracted from cells in an image of a fine needle aspirate of a breast mass (357 benign, 212 malignant). For illustration purposes, we train a GMLVQ system using 25% randomly sampled training data, and use the remaining 75% as a test set. Fig. 2 shows the projection of the training data into GMLVQ space at the end of training for the unrestricted system (iteration 0, (a)), and after the 1st iteration (b). Fig. 2 (c) shows the training data projected onto the leading eigenvector of the 0th and 1st iteration, where you can see a clear discrimination of the two classes along both coordinate axes. The leading eigenvalue of the GMLVQ system from both the 0th and 1st iteration is ≈ 1.0 respectively, meaning that there is no contribution from non-dominant eigenvectors from iteration 0 in iteration 1. The application of IRMA allows deeper insights into the feature relevances. For example, Fig. 3 shows that features 25 and 26 display significant $\Lambda_{jj} > 0$ in iteration (1), while they appear irrelevant in the unrestricted system (0). However, the performance of the two systems is virtually identical with $BAC^{(1)} \approx BAC^{(0)}$ (the BACs are estimated based on five random training and

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

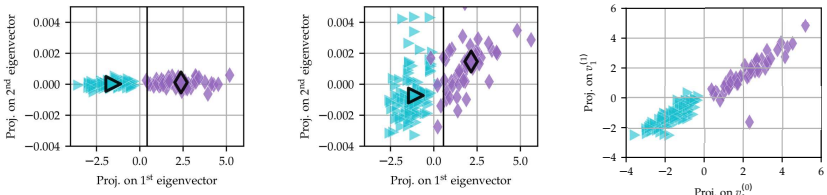

Fig. 2: Wisconsin data set: Projections after 0th (a), 1st iteration (b), and data projected onto leading eigenvectors of 0th and 1st iteration, respectively (c).

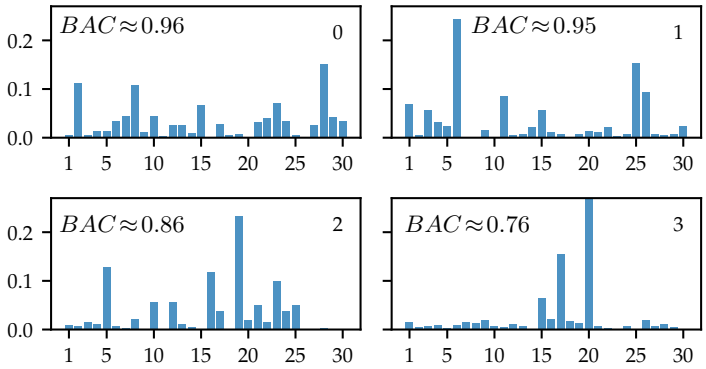

Fig. 3: Wisconsin data set: Diagonal of Λ per iteration (i), which is indicated as i in the upper right corner of each panel. In addition, the obtained BAC w.r.t. test data are shown. In iteration 3, $\Lambda_{20,20}$ is approx. 0.54.

test sets, while the relevance profiles correspond to a single split of the data). Hence, these features constitute examples of *weakly* relevant dimensions in the sense of the discussion given in [6, 7]: they enable successful classification in (1), but are replaced by other (combinations of) features in (0). Similarly, the single feature $j = 20$ with a relevance of 0.54 dominates the classification in iteration (3), while it plays only a minor role in the other classifiers.

Note that the test set accuracies decrease to $BAC^{(4)} \approx 0.71$ and $BAC^{(5)} \approx 0.57$. Here, we restrict the discussion to $V = \{\mathbf{v}_1^{(0)}, \mathbf{v}_1^{(1)}, \mathbf{v}_1^{(2)}, \mathbf{v}_1^{(3)}\}$ as the most discriminative subspace. Two features ($j = 9, 18$) display diagonal relevances $\Lambda_{jj}^{(i)} < 0.02$ for all $i \leq 3$ and, therefore, could be considered irrelevant. No features were rated relevant with $\Lambda_{jj}^{(i)} \geq 0.01$ for all $i \leq 3$.

4 Conclusion and Outlook

We have shown how IRMA based on GMLVQ with iterative subspace elimination can be used to find class-relevant subspaces for a two-class classification

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

problem. As an example, we have demonstrated that two mutually exclusive directions provide the same (highest) performance for the Wisconsin data set. Consequently, feature profiles from each relevant subspace can be taken into account for the final feature relevance analysis. This should be especially important for data sets with correlated or multiple weakly relevant features, or problems where only a small amount of training data is available. The issue of creating a (weighted) accumulated relevance profile reflecting the importance of a feature across all relevant subspaces will be addressed in future work. Note that the results of previous analyses depend strongly on the details of the method and the considered classifiers, compare e.g. [6] and [7].

Suitable criteria for the termination of the iteration will be the topic of forthcoming investigations. Similarly, the application of IRMA on multi-class problems is left as future work. Here, several dominant directions are expected per iteration, which can be removed simultaneously.

We also suggest that IRMA may be useful when building ensemble classifiers. Note that at each stage of IRMA a different classifier is obtained. In particular, the respective prototypes are placed in entirely different positions in feature space. Hence, it is non-trivial to construct a single classifier from the individual results. In general, the naive application of an LVQ classifier on the vectors $(y_0^\mu, y_1^\mu, \ldots, y_k^\mu)^\top$, cf. (5), will simply recover the unrestricted classifier by identifying y_0 as the most discriminative projection. Creating a weighted ensemble from all models (iterations) that achieve high performance, may result in a more robust performanc and would be of particular interest in the presence of subclusters within the classes.

References

[1] T. Kohonen. Self-Organizing Maps. Springer, Berlin, 1997.

[2] D. Nova and P.A. Estévez. A review of Learning Vector Quantization Classifiers. Neural Computing and Applications, 25: 511-524, 2014.

[3] S. Ghosh, P. Tino and K. Bunte. Visualisation and knowledge discovery from interpretable models. In: Intl. Joint Conf. on Neural Networks (IJCNN), 8 pages, 2020.

[4] P. Schneider, M. Biehl and B. Hammer. Adaptive relevance matrices in Learning Vector Quantization. *Neural Computation*, 21(12):3532-3561, 2009.

[5] K. Bunte, P. Schneider, B. Hammer et *al.* Limited rank matrix learning, discriminative dimension reduction, and visualization. *Neural Networks*, 26:159–173, 2012.

[6] C. Göpfert, L. Pfannschmidt, J.P. Göpfert, B. Hammer. Interpretation of linear classifiers by means of feature relevance bounds. *Neurocomputing*, 298: 69-79, 2018.

[7] C. Göpfert, L. Pfannschmidt and B. Hammer. Feature Relevance Bounds for Linear Classification. Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pages 187-192, 2017.

[8] M. Biehl, B. Hammer, F.-M. Schleif, P. Schneider, and T. Villmann. Stationarity of Matrix Relevance LVQ. In *Proc. Intl. Joint Conf. on Neural Networks (IJCNN)*, 8 pages, 2016.

[9] R. van Veen, N.R. Bari Tambolia, S. Lövdal et *al.* Subspace Corrected Relevance Learning with Application in Neuroimaging. Submitted, 2023.

[10] M. Lichman. UCI machine learning repository (2013). URL http://archive.ics.uci.edu/ml

[11] Q. Tao, D. Chu, J. Wang. Recursive support vector machines for dimensionality reduction. IEEE Transactions on Neural Networks. 19(1):189-93, 2008.