Mitigating the Bias in Data for Fairness Using an Advanced Generalized Learning Vector Quantization Approach – FA(IR)^{2}MA-GLVQ

M. Kaden[1], A. Engelsberger[1], R. Schubert[1], S.S. Lövdal[2,3], E.L. van den Brandhof[2,3], M. Biehl[2], and T. Villmann[1,4] *

1- Saxon Institute for Computational Intelligence and Machine Learning (SICIM), Mittweida University of Applied Sciences, Saxony, Germany
[kaden1|engelsbe|schuber6|villmann]@hs-mittweida.de

2- University of Groningen, The Netherlands
3- University Medical Center Groningen (UMCG), The Netherlands
[m.biehl|s.s.lovdal|e.l.van.den.brandhof]@rug.nl

4- Technical University Freiberg, Saxony, Germany

Abstract. We propose a bias detection and mitigating scheme for data in the context of classification tasks based on learning vector quantizers (LVQ) as classifier. For this purpose generalized LVQ endowed with an advanced matrix adaptation scheme is used for bias detection. The bias removal from data is realized applying a nullspace data projection using the adjusted matrix. The usefulness of the approach is demonstrated and illustrated in terms of two real world datasets.

1 Introduction

The analysis of data by machine learning (ML) methods is an ongoing topic of increasing importance. One crucial aspect is that data or ML-systems contain and reflect biases. This relates directly to fairness in ML-based support systems when they are involved in processes regarding humans, concerning personality characteristics, career decisions, etc. Frequently, it is assumed or claimed that the data are biased without specifying the kind of bias. Furthermore, explainable/interpretable AI is assumed to contribute to more fairness [6].

Handling these problems is challenging and has to be tackled by modern approaches. At least two main processing directions can be identified: (a) modification of ML-models to deal with/ignore the bias, or (b) to remove the bias from the data. Both directions have advantages and disadvantages: While ignoring-bias-models leave data unchanged but usually add penalties if hidden bias information is used, the removal ansatzes modify the data in such a way that the bias information is no longer present in the data, neither directly nor by implicit correlations. In this contribution, we focus on the latter strategy. In particular we consider an interpretable and robust classifier approach – the Generalized Learning Vector Quantization (GLVQ [10]) method and show how it is used to solve the task. We compare our approach with a recently proposed fairness variant of GLVQ (FairGLVQ [11]), which favors the first bias-strategy.

*M.K. is supported by the IAI-XPRESS and the DAIMLER project which both are parts of the interdisciplinarry project Artificial Intelligence Meets Space (AIMS, No. 50WK2270E) funded by the German Space Agency (DLR). A.E. is supported by the PAL-project (*Perspektive Arbeit Lausitz*) of UAS Mittweida, and S.L. by Stichting ParkinsonFonds.

The paper is structured as follows: First, we reconsider the bias-problem and describe the challenge in a more formal way. Thereafter, we briefly recapitulate variants of GLVQ as well as FairGLVQ followed by the explanation of our proposed approach. Numerical results demonstrate the ability of the method and concluding remarks give further perspectives.

2 The Bias Problem and Fairness

Inherent characteristic data structures or patterns are essential for succesful application of ML-methods to solve data analysis and processing tasks. Without any structural information, successful machine learning becomes impossible. Bias is usually understood as an unwanted distortion of the data in which some aspects of a dataset are given more weight and/or representation than others and which is related in social context to fairness. A skewed outcome, low accuracy levels, and analytical errors result from a dataset that is biased and therefore does not represent a model's use case accurately. Thus, a final ML-model can be affected by a specific bias due to the training with the biased data.

Now we suppose a hypothesis $H_0(b)$ that a *specific* bias b is contained implicitly in the data $\mathcal{X} \subset \mathbb{R}^n$ (suspected bias), influencing a model response $r(\mathbf{x})$ for a given task. Hence, the assumption is that the bias information $b(\mathbf{x})$ of a sample can be extracted from $r(\mathbf{x})$ with higher precision than by random guessing. If the bias extraction only results in a random guess, we refer to the response model as *fair with respect to the hypothetic / suspected bias.*

Here, we will concentrate on classification models $\mathscr{C}(\mathcal{P})$ depending on the model parameters $\mathcal{P}$, i.e. the response is a class label $c(\mathbf{x}|\mathscr{C}(\mathcal{P})) \in \mathcal{C} = \{1, \ldots, C\}$ for a presented sample $\mathbf{x}$. Accordingly, the tasks are: (I) detection of an expected bias b in the data with respect to a classification problem; (II) If bias is detected, remove this bias from the data such that the classification decision cannot benefit from this information. The second task is to search for a transformation T_b of the data such that the resulting data $\mathcal{X}_{\text{trans}} = T_b(\mathcal{X})$ is unbiased with respect to the considered hypothesis $H_0(b)$ in the context of the specified target classification.

We suggest the following procedure, referred to as the *bias-detection-mitigation-scheme* (BDMS):

1. classify the original data using a model $\mathscr{C}_{\text{orig}}(\mathcal{P})$ w.r.t. the given main classification task (MCT), yielding the accuracy acc_{orig}
2. classify the data using a model $\mathscr{C}_b(\mathcal{P})$, trained w.r.t. the pre-defined *specific* bias b, yielding the accuracy acc_b and compare it with random guessing:
   a) if acc_b is not significantly different from random guessing - STOP and accept the classifier and result from step 1
   b) if acc_b is better than random guessing - continue with 3.
3. mitigate/remove the bias from the data by the transformation $T_b(\mathcal{X})$ and (re-)train a model $\mathscr{C}_{\text{trans}}(\mathcal{P})$ from cleaned data $\mathcal{X}_{\text{trans}}$ w.r.t. the given target classification task, yielding the accuracy acc_{trans}
4. quantify the bias influence on the class prediction:
   a) calculate the fraction of data samples that are classified differently by the retrained model
   b) compare acc_{orig} with acc_{trans}

If the performances acc_{orig} and acc_{trans} deviate significantly, there is evidence that the supposed (and detected) bias influences the classification task, otherwise we can conclude that the detected bias has not led to an overvalued apparent performance of the target classification.

3 A GLVQ-approach for Data Bias-Removing

3.1 Matrix GLVQ for Classification – GMLVQ

A Generalized Matrix Relevance LVQ (GMLVQ) system generates a class decision for data $\mathcal{X} \subset \mathbb{R}^n$ based on class dependent prototypes $\mathcal{W} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^n$ with class labels $c(\mathbf{p}_k) \in \mathcal{C}$ such that each class is represented by at least one prototype. A data sample $\mathbf{x} \in \mathcal{X}$ is classified by the nearest prototype rule

$$\mathbf{x} \mapsto c(\mathbf{p}^*), \quad \mathbf{p}^* = \text{argmin}_{\mathbf{p}_k \in \mathcal{W}^* \subseteq \mathcal{W}} \left[d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}_k) \right] \tag{1}$$

$$\text{where} \quad d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}_k) = (\mathbf{x} - \mathbf{p}_k)^\top \mathbf{\Lambda} (\mathbf{x} - \mathbf{p}_k) \tag{2}$$

is a quadratic form with $\mathcal{W}^* = \mathcal{W}$ parameterized by $\mathbf{\Lambda} = \mathbf{\Omega}^\top \mathbf{\Omega}$ with $\mathbf{\Omega} \in \mathbb{R}^{m \times n}$ and $m \leq n$ [3]. For training data $\mathcal{T} = \{(\mathbf{x}_j, c(\mathbf{x}_j)) \in \mathcal{X} \times \mathcal{C}, j = 1, \ldots, N\}$ a loss function approximating the misclassification error in GMLVQ is

$$L_{\text{GMLVQ}}(\mathcal{T}, \mathcal{P}) = \sum_{j=1}^{N} \text{sgd}\left(\frac{d_{\mathbf{\Omega}}(\mathbf{x}_j, \mathbf{p}^+) - d_{\mathbf{\Omega}}(\mathbf{x}_j, \mathbf{p}^-)}{d_{\mathbf{\Omega}}(\mathbf{x}_j, \mathbf{p}^+) + d_{\mathbf{\Omega}}(\mathbf{x}_j, \mathbf{p}^-)} \right), \text{sgd}(z) = \frac{1}{1 + \exp(-z)}$$

where $\mathbf{p}^+$ and $\mathbf{p}^-$ are the class dependent best matching correct/incorrect prototypes for $\mathbf{x}_j$ according (1) with $\mathcal{W}^* = \mathcal{W}_j^+ = \{\mathbf{p}_k \in \mathcal{W} | c(\mathbf{p}_k) = c(\mathbf{x}_j)\}$ and $\mathcal{W}^* = \mathcal{W}_j^- = \{\mathbf{p}_k \in \mathcal{W} | c(\mathbf{p}_k) \neq c(\mathbf{x}_j)\}$, respectively. Hence, the GMLVQ model is determined by the parameter set $\mathcal{P} = \{\mathcal{W}, \mathbf{\Omega}\}$. Optimization takes place as stochastic gradient descent learning with respect to all parameters. Important to note that GMLVQ constitutes a non-linear classifier if at least three prototypes are contained in $\mathcal{W}$. Further, the optimized relevance matrix $\mathbf{\Lambda} = \mathbf{\Omega}^\top \mathbf{\Omega}$ reflects the correlation between the data attributes contributing to a correct classification and, hence, is frequently also denoted as classification correlation matrix (CCM). The choice $\mathbf{\Lambda} = \mathbf{I}$ without $\mathbf{\Omega}$-optimization yields standard GLVQ.

Recently, Iterated Relevance Matrix Analysis (IRMA) was proposed [8], which recursively determines a linear subspace representing the classification specific information of the considered datasets using GMLVQ. IRMA makes use of the representation $\mathbf{\Lambda} = \sum_{l=1}^{n} \lambda_l^{(0)} \mathbf{v}_l^{(0)} \mathbf{v}_l^{(0)\top}$ where $\mathbf{v}_l^{(0)}$ are the eigenvectors of $\mathbf{\Lambda}$ and $\lambda_1^{(0)} \geq \lambda_2^{(0)} \geq \ldots \geq \lambda_n^{(0)} \gtrapprox 0$ the corresponding eigenvalues. For binary classification problems, $\mathbf{\Lambda}$ is typically dominated by a single leading eigenvector [2]. We define the projector $\mathbf{P}^{(0)} = \mathbf{I} - \mathbf{v}_1^{(0)} \mathbf{v}_1^{(0)\top}$. Now, we can perform another GMLVQ training using the quadratic form (2) for the projected data $\mathbf{x}_j^{(1)} = \mathbf{P}^{(0)} \mathbf{x}_j^{(0)}$ with $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, i.e. we apply GMLVQ to the nullspace of $\mathbf{\Omega}^{(0)} = \sqrt{\lambda_1^{(0)}} \mathbf{v}_1^{(0)} \mathbf{v}_1^{(0)\top}$. Applying this scheme iteratively we obtain

a sequence of vectors $\mathbf{v}_1^{(i)}$ where each $\mathbf{v}_1^{(i)}$ is orthogonal to the vectors $\mathbf{v}_1^{(\hat{i})}$ with $\hat{i} = 0, \ldots, i-1$. In each iteration step i we obtain $\mathbf{v}_1^{(i)}$, $\mathbf{P}^{(i)} = \mathbf{I} - \sum_{k=0}^{i-1} \mathbf{v}_1^{(k)} \mathbf{v}_1^{(k)\top}$ and $\mathbf{\Omega}^{(i)} = \sqrt{\lambda_1^{(i)}} \mathbf{v}_1^{(i)} \mathbf{v}_1^{(i)\top}$. The corresponding subspace

$$V_K = \text{span}\left\{\mathbf{v}_1^{(0)}, \mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_1^{(K)}\right\} \text{ with associated projections } x_j^{(i)} = \mathbf{x}_j^\top \cdot \mathbf{v}_1^{(i)} \quad (3)$$

approximately covers all relevant data information regarding the classification task for sufficiently large $K \leq n$. Thus, $N_K = \mathbb{R}^n \setminus V_K$ is the nullspace of the projector $\mathbf{P}^{(K)}$ with the null-space projector $\mathbf{P}_0^{(K)} = \sum_{i=1}^{K} \sqrt{\lambda_1^{(i)}} \mathbf{v}_1^{(i)} \mathbf{v}_1^{(i)\top}$. In consequence, the projected data $\hat{\mathbf{x}}_j = \mathbf{P}_0^{(K)} \cdot \mathbf{x}_j$ do not contain any further useful specific information for the class discrimination. As explained in [8], this scheme can easily be extended to multiple-class scenarios.

3.2 Application of IRMA-scheme to Remove Bias Information in GLVQ-Classification

In the following, we suppose bias classes $c_b(\mathbf{x}) \in \mathcal{C}_b = \{1_b, \ldots, C_b\}$ of the data corresponding to a bias classification task – BCT, whereas the MCT task is to discriminate the data into classes $\mathcal{C} = \{1, \ldots, C\}$. Further, we suppose available training data $\mathcal{T}_b = \{(\mathbf{x}_j, c(\mathbf{x}_j), c_b(\mathbf{x}_j)) \in \mathcal{X} \times \mathcal{C} \times \mathcal{C}_b, j = 1, \ldots, N\}$. Finally, we suspect a bias hypothesis $H_0(b)$: The data $\mathcal{X} \subset \mathbb{R}^n$ may contain information regarding the bias type b influencing a main classification task (MCT). An approved hypothesis $H_0(b)$ indicates that the bias contributes to the MCT.

To validate $H_0(b)$ by means of GMLVQ as the classifier in use, we follow the scheme proposed in Sect. 2: First, a GMLVQ with prototypes $\mathcal{W} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^n$ is trained for the MCT using the training sample pairs $(\mathbf{x}_j, c(\mathbf{x}_j))$ yielding predictions $c^0(\mathbf{x}_k)$ for test data $\mathcal{T}_{\text{test}} = \{\mathbf{x}_k \in \mathbb{R}^n, k = 1, \ldots, N_{\text{test}}\}$. In the second step, a GMLVQ followed by the IRMA-scheme is applied to the BCT generating the projector $\mathbf{P}^{(K)}$ with the corresponding nullspace-projector $\mathbf{P}_0^{(K)}$ identified as the data transformation $T_b(\mathcal{X})$ to mitigate/remove the bias as explained in sec. 2. The last step is to train a new GMLVQ for the MCT with the same parameter setting as for the GLVQ in the first step but using the nullspace-projected training sample pairs $\left(\mathbf{P}_0^{(K)} \cdot \mathbf{x}_j, c(\mathbf{x}_j)\right) \in \mathcal{X}_{\text{trans}} \times \mathcal{C}$ and determine the predictions $c^1\left(\mathbf{P}_0^{(K)} \cdot \mathbf{x}_k\right)$ for the test data.

If the number N_b of data $\mathbf{x}_k \in \mathcal{T}_{\text{test}}$ with $c^0(\mathbf{x}_k) \neq c^1\left(\mathbf{P}_0^{(K)} \cdot \mathbf{x}_k\right)$ leads to a *bias-influence-ratio* (BIR) $r_b = N_b / N_{\text{test}} \ll 1$ we can conclude that the hypothesis $H_0(b)$ is not supported by the data and the opposite hypothesis $H_1(b)$ should be preferred, i.e. the suspected bias is not validated.

Note that the bias removal realized by the projector $\mathbf{P}_0^{(K)}$ is only a linear operation and, hence, strong non-linear bias would not be removed completely. Yet, in high-dimensional data, linear approximations of non-linear manifolds frequently give sufficient quality [7] whereas in low-dimensional settings the deviations usually become more substantial. The GLVQ obtained for the

dataset	BIR r_b	acc_{orig}	acc_{b}	acc_{trans}
SPD	0.558	63.7	75.2	60.7
PIMA	0.146	76.7	80.9	71.2

Table 1: Results of BDMS using the FA(IR)^{2}MA-GLVQ. The bias-influence ratios (BIR) $r_b \in [0, 1]$ reflect the number of differently evaluated samples after bias mitigation. The balanced accuracy values (acc) are given in %.

projected data still yields a non-linear classifier if more than two prototypes are used. We refer to this approach as FAIRness-IRMA-based GLVQ – FA(IR)^{2}MA-GLVQ.

4 Numerical Experiments

SPD – Student Performance Dataset The *student performance dataset* examines student performance in secondary education at two Portuguese schools in two subjects: *Portuguese* and *Math*, here we consider only the former. The attributes include student grades as well as demographic, social, and school-related features collected by the schools. More details about the features and further background information can be found in [5]. The data set consists of 649 data points for the *Portuguese* subject. The number of features is 30 and the categorical features are preprocessed by one hot-encoding. Moreover, for the target we chose the final grade (numeric from 0 to 20), which we grouped into the three classes: failed (< 10), mediocre ($10-13$), good to very good ($14-20$). Nevertheless, the data set is very unbalanced. As mentioned in [4] the grades could be biased by gender, as the information is given as a feature in the data set. Accordingly, the suspected bias b is *gender*.

PIMA – Pima Indians Diabetes Dataset The second dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases provided by [1]. The primary purpose is to predict whether a patient suffers from diabetes using specific diagnostic measurements provided in the data. The dataset contains 768 samples of 7 medical predictor variables like body mass index, insulin level, glucose concentration, etc. for female patients who are all at least 21 years old and are of Pima Indian descent. It is assumed that the age could be influencing the diabetes diagnostics, i.e. age could be a bias. The provided age information for the samples is grouped into young age (< 31) and older (≥ 31). Thus, here the suspected bias b is *age*.

Numerical Results

Both datasets were z-score normalized. All results we report are obtained by 5-fold cross validation. We applied the BDMS using the FA(IR)^{2}MA-GLVQ introduced above for both datasets with exactly the same parameters. The results are presented in Tab. 4. As expected, both datasets suffer from the suspected bias (non-vanishing r_b-ratios), i.e. the bias hypothesis is validated by the data: SPD is heavily biased regarding gender whereas PIMA is slightly biased w.r.t. age. Interestingly, the overall accuracy is not affected to the same

degree. The reason for this behavior could be that the MCT is challenging and bias mitigation does not influence the performance as much as expected.

5 Concluding Remarks

In this paper we have shown that GMLVQ together with the IRMA procedure for data related class information extraction can be used to successfully mitigate suspected bias in the data regarding the given classification task.

We remark that a similar GMLVQ-based null-space projection approach was suggested in [12] and in [13] for leveling the influence of different data sources in federated learning. Further, an iterated nullspace evaluation for bias detection was also proposed for text analysis in context of linear classifiers [9], while GMLVQ constitutes a non-linear classifier if the overall number of prototypes is greater than two.

References

[1] A. Asuncion and D. Newman. Indian diabetes data set (PIMA), http://archive.ics.uci.edu/ml/.

[2] M. Biehl, B. Hammer, F.-M. Schleif, P. Schneider, and T. Villmann. Stationarity of matrix relevance LVQ. In *Proc. of the International Joint Conference on Neural Networks 2015 (IJCNN)*, pages 1–8, Los Alamitos, 2015. IEEE Computer Society Press.

[3] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[4] P. Cortez. Student Performance Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TG7T.

[5] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, 2008.

[6] L. Deck, J. Schöffer, M. De-Arteaga, and N. Kühl. A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT'24), Rio de Janeiro Brazil*, pages 1579–1595, 2024.

[7] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, New York, 2007.

[8] S. S. Lövdal and M. Biehl. Iterated Relevance Matrix Analysis (IRMA) for the identification of class-discriminative subspaces. *Neurocomputing*, 577(127367):1–6, 2024.

[9] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: guarding protected attributes by iterative nullspace projection. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

[10] A. S. Sato and K. Yamada. Generalized learning vector quantization. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 423–429. MIT Press, 1995.

[11] F. Störck, F. Hinder, J. Brinkrolf, B. Paaßen, V. Vaquet, and B. Hammer. FairGLVQ: Fairness in partition-based classification. In T. Villmann, M. Kaden, T. Geweniger, and F.-M. Schleif, editors, *Proceedings of the 15th Workshop on Self-Organizing Maps, Learning Vector Quantization and Beyound (WSOM+ '2024)*, pages 141–151, 2024.

[12] R. van Veen, N. Bari Tambolia, S. Lövdal, S. Meles, R. Remken, G.-J. de Vries, D. Arnaldi, S. Morbelli, P. Clavero Ibarra, J. Obeso Inchausti, M. Rodriguez Oroz, K. Leenders, T. Villmann, and M. Biehl. Subspace corrected relevance learning with application in neuroimaging. *Artificial Intelligence in Medicine*, 149(102786):1–12, 2024.

[13] T. Villmann, D. Staps, J. Ravinchandran, S. Saralajew, M. Biehl, and M. Kaden. A learning vector quantization architecture for transfer learning based classification by means of nullspace evaluation. In T. Bouadi, E. Fromont, and E. Hüllermeier, editors, *Advances in Intelligent Data Analysis XX – Proceedings of the 20th Symposium on Intelligent Data Analysis (IDA 2022)*, volume 13205 of *Lecture Notes in Computer Science*, pages 354–364. Springer, 2022.