Causal Relevance Learning for Robust Classification under Interventions

Ernest Mwebaze[1,2] and Michael Biehl[2] and John A. Quinn[1]

[1]Faculty of Computing & IT, Makerere University
P.O. Box 7062, Kampala, Uganda.

[2] Johann Bernoulli Institute for Mathematics and Computer Science
Univ. of Groningen P.O. Box 407, 9700AK Groningen, The Netherlands

Abstract. In some classification problems the distribution of the test data is different from that of the training data because of external manipulations to the variables we observe. We propose a classification scheme which is robust to outside interventions by identifying causes in the training data, given that causes of a target variable remain predictive even when the data is manipulated. We do this by extending Relevance Learning Vector Quantization (RLVQ), a classification scheme that learns a relevance profile for the classification task presented. Our proposed algorithm, Causal-RLVQ, learns a relevance profile that weights causally relevant features more strongly. The algorithm can determine a trade-off between robustness to intervention and accuracy on non-manipulated data, yielding RLVQ as a special case.

1 Introduction

The task of performing classification on data for which some of the variables have been externally manipulated is a specific case of the dataset shift problem. This can occur when deploying a classification in many practical settings; for example, we may be trying to classify whether a region is at risk of a disease outbreak based on some environmental and demographic factors, when some of those factors have been directly influenced by other parties in ways not seen in the training data.

Finding the causes of a target variable is primarily useful in order to predict the effects of interventions on that variable, and in the last decade a number of methods have been developed to do this on purely observational data [1]. It therefore makes sense for prediction and causal discovery to be tightly coupled, and the algorithm we describe in this work carries out both tasks simultaneously in a prototype-based learning scheme.

In this paper we introduce a scheme that extends relevance learning vector quantization (RLVQ) [2]. RLVQ generalises the distance measure of input data such that the features relevant to the target variable are weighted more heavily. Our extension, Causal RLVQ, introduces a new parameter α which determines how far to bias the relevance weights towards causative features. When $\alpha > 0$, causative features are favoured, giving us robust classification at test time under such cases where the new data is suspected to have been intervened upon. We do this by trying to identify so called "V-structures" (section 2) in the data.

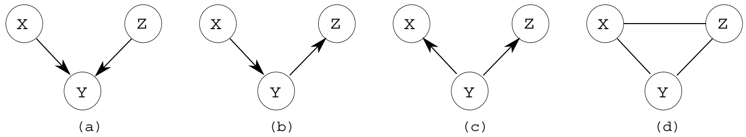


Fig. 1: Conditional independence configurations of 3 variables

2 Identifying causes in observational data

Techniques for recovering causal structure from observational data broadly fall into Bayesian model search (e.g. [3]) or constraint-based/hypothesis testing techniques [1] (the latter having recently been found able to deal even with the extreme 2-variable case [4]). A fundamental concept in this work is conditional independence. When we have sets of three or more variables, conditional independence properties can help to rule in or out particular causal configurations of the variables. X is conditionally independent of Y given Z, written $X \perp\!\!\!\perp Y \mid Z$, if $P(X|Y,Z) = P(X|Z)$.

Figure 1 shows some configurations of three variables X, Y and Z: (a) *Collider* ($X \perp\!\!\!\perp Z$ but $X \not\perp\!\!\!\perp Z \mid Y$), (b) *Chain* ($X \perp\!\!\!\perp Z \mid Y$), (c) *Fork* ($X \perp\!\!\!\perp Z \mid Y$), and (d) the fully connected configuration (can have any conditional independence properties). Note that the chain and fork are in the same conditional independence class, but the collider is uniquely specified by its conditional independence properties. The collider is therefore an important structure in causal discovery. In hypothesis tests, such as the prototypical Inductive Causation algorithm [1, §2.5], colliders are identified by looking for any variables X and Z which are unconnected (dependent under every conditioning set) with each other, but both of which are connected with a third variable Y. When such a configuration is found, the edges are orientated as in Figure 1(a). This is the approach that we adopt here, using the RLVQ framework itself rather than hypothesis tests.

3 RLVQ

The basic LVQ scheme is set up as follows; Given a dataset $D = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^{P}$ where $\mathbf{x}^\mu \in \mathbf{R}^N$ and the labels $y^\mu \in 1, 2, \ldots C$ correspond to one of the classes, the LVQ scheme is parameterized by a set of prototype vectors $\mathbf{w} = \left\{\mathbf{w}^j, c(\mathbf{w}^j)\right\}_{j=1}^{M}$ with the prototype vectors $\mathbf{w}^j$ having labels $c(\mathbf{w}^j) \in 1, 2, \ldots C$.

For a particular dissimilarity/distance measure $d(\mathbf{x},\mathbf{w})$, the LVQ classifier employs a Winner-Takes-All scheme where an arbitrary input is assigned to the class $c(\mathbf{w}^L)$ of the closest prototype with $d(\mathbf{x}, \mathbf{w}^L) \leq d(\mathbf{x}, \mathbf{w}^j)$ *for all* j.

Many modifications of Kohonen's original formulation [5] have been suggested with the aim of achieving better convergence and generalization behavior. A specific set of modifications have been towards accounting for heterogeneous datasets where features can have different meanings and magnitudes. These are the class of relevance learning schemes which employ adaptive scaling factors for each dimension in the feature space. For our purposes it will suffice to give the general formulation of RLVQ.

For RLVQ we can consider a generalized Euclidean distance of the form

$$d(\mathbf{x}, \mathbf{w}^J) = \sum_{j=1}^{N} \lambda_j (x_j - w_j^J)^2, \tag{1}$$

as the dissimilarity measure where λ_j are the adaptive relevance factors. The special case $\lambda_j = 1/N$ for all $j = 1, \ldots N$ is analogous to the original LVQ1 formulation. Each update of the winning prototype w_j is accompanied by a corresponding update in the relevance factor $\lambda_j(t)$ as follows;

$$\lambda_j(t) = \lambda_j(t-1) - \eta_\lambda \phi \cdot (x_j - w_j^J)^2 \tag{2}$$

where $\lambda_j(t)$ is restricted to non-negative values and obeys the normalization $\sum_{j=1}^{N} \lambda_j = 1$.

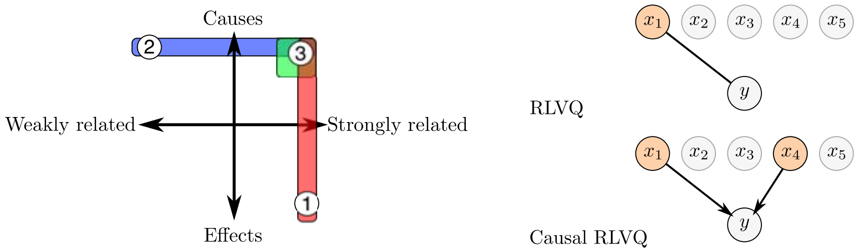
The λ update hence decreases the relevance factor λ_j if the winning prototype $\mathbf{w}^J$ does represent the correct class but the contribution $(\mathbf{x}_j - \mathbf{w}_j^J)^2$ to $d(\mathbf{x}, \mathbf{w}^J)$ is relatively large. Conversely the weight of a feature with relatively small $(\mathbf{x}_j - \mathbf{w}_j^J)^2$ is increased in such a case. The learning rates η_w and η_λ control the magnitude of the prototype and relevance factor updates at each step.

4 Causal RLVQ

CRLVQ is our extention of the RLVQ scheme where updates favor features that are causally related to the target feature. Our assessment of causal relevance is based on identifying V-structures with respect to the target. RLVQ gives a profile that represents how strongly each single dimension of the data is predictive of the target. In CRLVQ instead of looking at a single dimension, we look for evidence that there are two dimensions x_i and x_k that are predictive of the target. To ascertain that x_i and x_k are in a V-structure with the target we also check that they are independent of each other.

Figure 2(a) illustrates this salient distiction between RLVQ and CRLVQ. We conceive of features $(x_1, \ldots, x_P)$ as being characterisable along two dimensions: their predictiveness of the target variable, and their causal effect on the target variable. Standard relevance learning aims to give higher weight to a set of variables (1) which are highly predictive; causal structure learning aims to find causes (2) or effects of a target; our work here identifies variables which are predictive causes (3). Figure 2(b) shows the difference in which RLVQ considers each feature separately; CRLVQ takes features pairwise in order to identify V-structures with the target.

CRLVQ extends the RLVQ update by adding two extra evaluation criteria to equation 2. The three criteria in total are in a sense a distance-based formulation of the V-structure condition. For every example presented to the CRLVQ classification scheme, each component of λ in CRLVQ is hence updated for every



(a) (b)

Fig. 2: RLVQ and CRLVQ formulations. Panel (a) is an illustration of placement of variables across the predictive-causal space. Panel (b) shows how RLVQ is extended to CRLVQ

dimension x_j for each introduction of a data example as follows

$$\lambda_j(t) = \lambda_j(t-1) - \eta_\lambda \phi \cdot (x_j - w_j^J)^2 - \alpha\eta_\lambda \cdot \left(\min_{k \neq j} \left(\phi \cdot (x_k - w_k^J)^2 - (x_j - x_k)^2 \right) \right) \tag{3}$$

The parameter α is a parameter that weights the two new criteria. Standard RLVQ is hence a special case of CRLVQ when $\alpha = 0$. We evaluate the independence of the different data dimensions by looking at their absolute difference. In Z-score transformed data, this will be a small quantity for pairs of dimensions that are positively correlated. The update hence rewards any feature x_j if it has a strong correlation (small difference) with the target/label vector as represented by a correct prototype $(x_j - w_j^J)^2$, and also if there is strong evidence of another feature with a strong correlation to the target as well $(x_k - w_k^J)^2$, and if feature x_j and x_k are weakly correlated (large difference apart).

Note however that there might be other relationships between the dimensions that make them dependent on each other, but that this update would not capture – for example negative correlation.

5 Experiments

Simulated causal networks were used to demonstrate the effect of the causal λ update. One causal network was formulated as a 5-feature linear Gaussian network with 2 causes, 2 effects and 1 irrelevant feature, shown in Figure 3(a). In Figure 3(b) we can see that as α is increased, the relevance weights increase for the features that are causally relevant.

For experiments on manipulated test sets we used a network simulated from a Bayes net and used for several causal competitions as a trial set for tuning causal algorithms. It is commonly called the Lucas dataset[6, 7] and tries to predict lung cancer based on different features. Three versions of the dataset were used: Lucas0 - the natural network (unmanipulated) from which the training set was

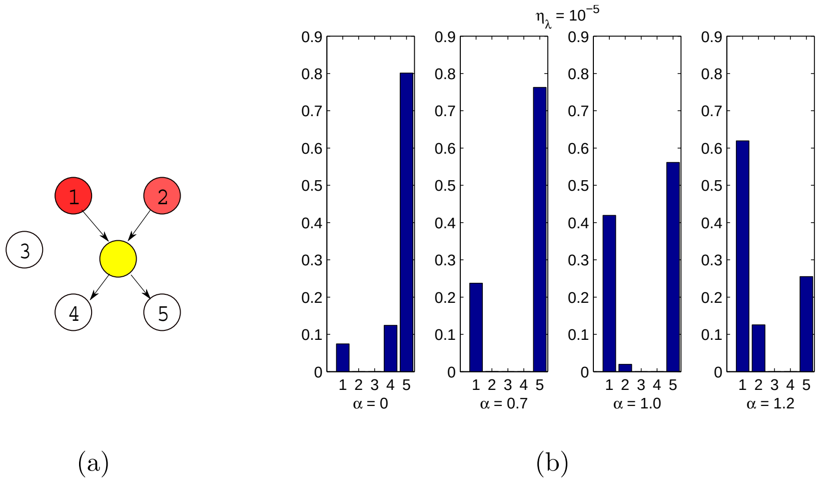


Fig. 3: CRLVQ relevance weights for simulated network data. Panel (a) shows the simulated network with 2 causes, 2 effects and 1 irrelevant feature. Panel (b) shows the relevance profile under RLVQ and under CRLVQ for different parameter settings.

obtained, and manipulated datasets Lucas1 and Lucas2 (manipulated variables shown in Figure 4).

Figure 5 shows the test error scores for the three datasets Lucas0, Lucas1 and Lucas2 for CRLVQ with varying α parameters. The first row with $\alpha = 0$ represents RLVQ. Lucas0 represents the unmanipulated(natural) dataset while Lucas1 and Lucas2 represent manipulated/intervened upon datasets. Test results are obtained after 50 epochs through the data with $\eta_\lambda = 10^{-5}$.

For Lucas0, the unmanipulated dataset we notice a better test error for RLVQ ($\alpha = 0$) than for any other value of α (CRLVQ). This is plausible because for good classification performance effects (and possibly other features) are just as relevant as causes. For the manipulated datasets, we notice that test error increases for RLVQ as expected because there are fewer relevant features. We however notice an increase in performance as we tune up α because then only possible causes are identified which remain relevant even under manipulations. Figure 5 (b) illustrates the relevance profile for the different values of α.

6 Conclusion

This paper applies concepts from causal discovery to relevance learning for prototype-based classification. The goal is to assure robust classification when it is not certain that the test set has the same distribution as the training set. We have given a proof of possibility on simulated data, though our formulation still has some limitations, particularly in our assessment of independence amongst features in the data. An interesting direction to address this could be to extend

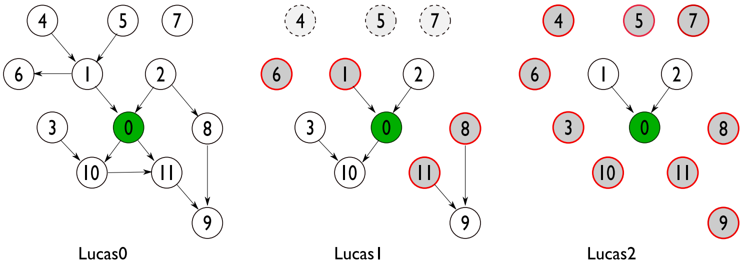
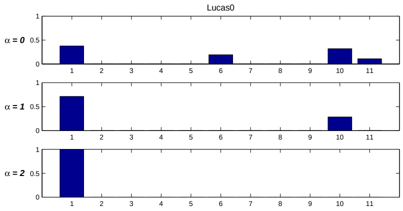


Fig. 4: Lucas graphs showing the network structures for the unmanipulated set, Lucas0 and the manipulated sets, Lucas1 and Lucas2. The different features in the graphs are labelled as follows; 0 - Lungcancer, 1 - Smoking, 2 - Genetics, 3 - Allergy, 4 - Anxiety, 5 - Peer Pressure, 6 - Yellow Fingers, 7 - Born an Even Day, 8 - Attention Disorder, 9 - Car Accident, 10 - Coughing, 11 - Fatigue.

	Lucas Dataset		
α	Lucas0	Lucas1	Lucas2
0	0.1970	0.2370	0.2620
1.0	0.2040	0.2040	0.2030
2.0	0.2040	0.2040	0.2030

(a) Results



(b) RLVQ as λ changes

Fig. 5: Test error results for RLVQ and CRLVQ for the different datasets Lucas0, Lucas1 and Lucas2.

our current work to matrix relevance LQV [8].

Acknowledgement: We would like to acknowledge funding from NUFFIC Project NPT-UGA-238: Strengthening ICT Training and Research Capacity in Uganda, and from Google, for this work.

References

[1] Judea Pearl. *Causality : Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2000.

[2] T. Bojer, B. Hammer, D. Schunk, and Tluk von Toschanowitz K. Relevance determination in learning vector quantization. In *9th European Symposium on Artificial Neural Networks. ESANN'2001. Proceedings. D-Facto, Evere, Belgium*, pages 271–276, 2001.

[3] D.M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.

[4] J. Peters, D. Janzing, and B. Schölkopf. Identifying cause and effect on discrete data using additive noise models. *AISTATS*, 2010.

[5] T. Kohonen. *Self Organizing Maps*, volume 2nd Edition. Springer, Berlin, 1997.

[6] Isabelle Guyon. Lung cancer simple model, 10 2009.

[7] Causality Workbench. Causality workbench data repository, November 2010.

[8] P. Schneider., M. Biehl., and B. Hammer. Adaptive revelance matrices in learning vector quantization. *Neural Computing*, 21:3535–3561, 2009.