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

Prototype based vector quantization schemes essentially operate by defining representatives (prototypes) in the data space. A dissimilarity measure, most commonly a distance based measure, is used to determine the dissimilarity between a data point and the prototype and hence perform the classification (or clustering) task. In this sense, Learning Vector Quantization (LVQ) and all its derivatives that are prototype based tend to be relatively easy to implement and provide classifiers that are intuitive to understand.

For our purposes here we look at Relevance Learning Vector Quantization (RLVQ)[1] which introduces adaptive versions of the dissimilarity measure based on how relevant the individual features in the data are for the classification task at hand. This has a two-fold advantage;(1) scaling the metric to fit the specific data hence improving classification and (2) introducing a feature selection or pruning algorithm. Like in most classification schemes, it is assumed the classifier will be used on a new dataset (test set) that has the same distribution as the dataset used for the training. Many real problems tend to violate this assumption because usually someone or some external factor has intervened on the new dataset. External factors could be artifacts due to the data collection process, or direct interventions for example one can imagine an economic status classification task where several interventions could significantly change the test set.

In this paper we introduce a scheme that extends RLVQ to produce a causally relevant profile that will tend to offer more robust classification under such cases

where the new data is suspected to have been intervened upon. We do this by trying to identify V-structures in the data and updating the relevance profile based on evidence we receive in the training of such a configuration amongst some of the features. This applies to the supervised learning LVQ schemes.

The structure of the remainder of the paper is as follows; initially we explain the problem of identifying causes from observational data, then give a small background on the basic RLVQ scheme that we extend in the following section. We then follow on with a description of how the causal relevance scheme works and we conclude with some experiments on different datasets.

2 Identifying causes in observational data

Causation denotes that relationship between any two variables that entails that a change in one will influence a change in the other positively or negatively. Causal discovery by this definition hence necessitates an active process of intervening on one variable and determining if there is a corresponding change in the other one, an example being Randomized Control Experiments (RCE). This is however for most cases impractical or unethical and as such recent research in causality has focused on causal discovery purely from observational data. This is generally a daunting task and one has to be careful not to confuse correlation with causation, this leads to what is termed the classical paradoxes [2].

To discover causes from observational data, many of the methods todate use conditional independence(dependence) tests. X is conditionally independent of Y given Z, written $X \perp\!\!\!\perp Y|Z$, if $P(X|Y,Z) = P(X|Z)$.

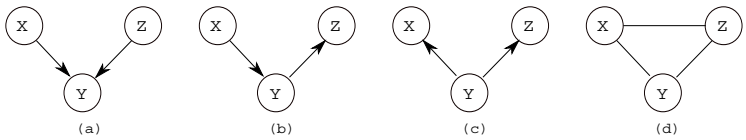


Fig. 1: Conditional independence configurations of 3 variables

Figure 1 shows the several conditional independence configurations of three variables X, Y and Z commonly identified as; (a) Collider: $X \perp\!\!\!\perp Z$ but $X \not\perp\!\!\!\perp Z|Y$, (b) Chain : $X \perp\!\!\!\perp Z|Y$, (c) Fork : $X \perp\!\!\!\perp Z|Y$, and (d) the fully connected configuration with indistinguishable conditional independence properties. Of these, the collider that provides the so-called 'V-structure' is of most critical importance in causal discovery because it is the only configuration that provides unique causal characteristics of the three.

It suffices to note that most algorithms in causal discovery aim at discovering V-structures in the data as a way of determining causal relationships in the data. Causal discovery methods have also been extended in several ways including; using bayesian networks to model causal relationships using conditional independencies, search-and-score methods of causal discovery, causal discovery in multivariate relatinoships between continous variables using Structural Equa-

tion Models (SEMs), causal discovery using Independent Component Analysis and more recently progress has also been in discovering causality from cause-effect pairs [3, 4].

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.

In the literature, 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. A good background to these modifications is given in the literature[6, 7, 8]. 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;

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

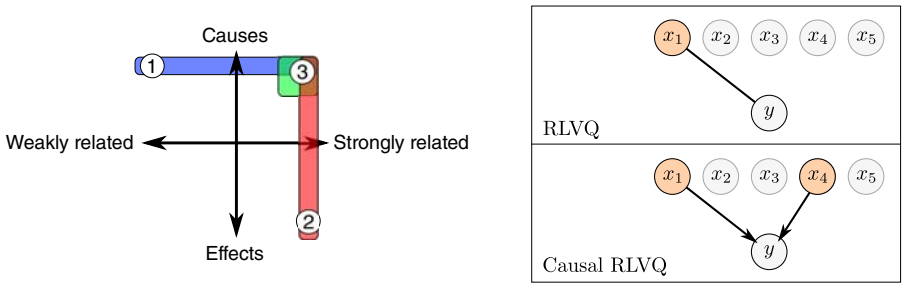
where $\lambda_j(t)$ is restricted to non-negative values and obeys the normalization $\sum_{j=1}^{N} \lambda_j = 1$.

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 CRLVQ

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.



(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

Figure 2 illustrates this salient distiction between RLVQ and CRLVQ. Panel (a) shows a placement grid for variables $(x_1, \ldots, x_P)$ categorised according to causal relevance or simply predictive relevance to the target variable y. 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). Panel (b) shows the criteria used for evaluating relevance scores in RLVQ and CRLVQ.

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 dimension x_j for each introduction of a data example as follows

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

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 give a low number in the 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 validate 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..

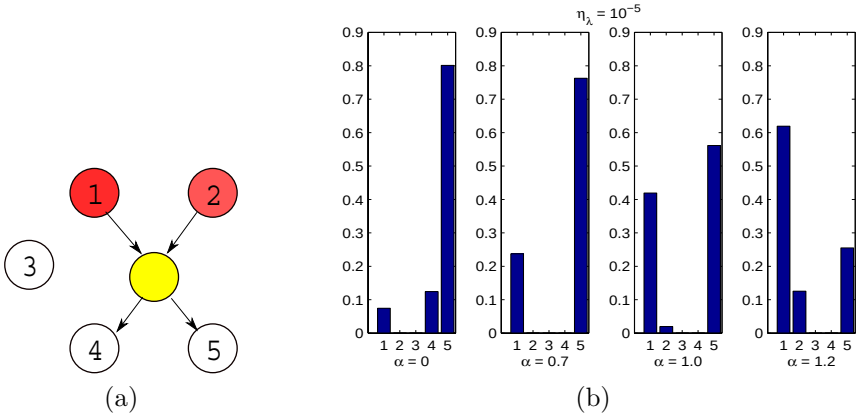


Fig. 3: RLVQ and CRLVQ for Simulated Network. Panel (a) shows the simulated network with 2 causes, 2 effects and 1 irrelevant feature. Panel (b) shows the relevance profile under RLVQ and under CRLVQ for different parameter settings.

From Figure 3, it is evident to see that with introduction of the causal update, the learning favours the features that are causally relevant over the 'just' correlated features and the irrelevant or non-correlated features. Making the learning rate for the causal update bigger results in bigger updates in the learning as well.

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[9] and tries to predict lung cancer based on different features. A full description of the dataset and its previous use can be found on the causality workbench[10]. 3 versions of the dataset were used; Lucas0 - the natural network(unmanipulated) from which the training set was obtained, Lucas1 and Lucas2 - manipulated datasets of the same network from which the test data was obtained.

Figure 4 illustrates the three networks from which the datasets were sampled. Networks for Lucas1 and Lucas2 show the features that were manipulated by setting them to arbitrary values. The effect of manipulations on a feature is

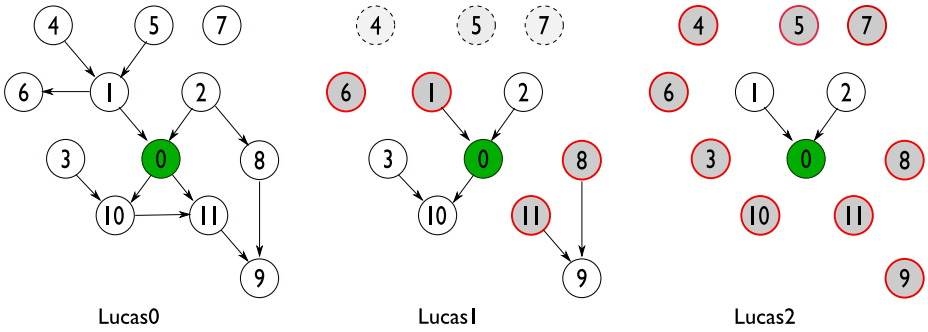


Fig. 4: Lucas graphs showing the network structures for the unmanipulated set, Lucas0 and the manipulated sets, Lucas1 and Lucas2. The different features in the graphs are labelled as follows; 0 - Lungcancer, 1 - Smoking, 2 - Yellow Fingers, 3 - Anxiety, 4 - Peer Pressure, 5 - Genetics, 6 - Attention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

to separate it from its parent features. Lucas1 had a few features manipulated while Lucas2 had all the features manipulated. The target was not manipulated. The network in Lucas2 represents an extreme case where all the features are manipulated, while this is unnatural, it illustrates the role of causes in classification when the data has been intervened upon because causes remain predictive and hence relevant for the classification.

6 Results

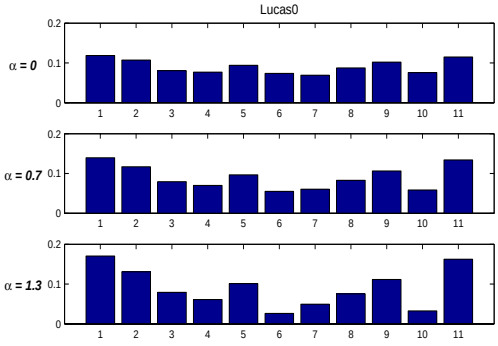
Table 1 shows the test error scores for the three datasets Lucas0, Lucas1 and Lucas2 for CRLVQ with varying α parameters. The first row with $\alpha = 0$ represents RLVQ. Lucas0 represents the unmanipulated(natural) dataset while Lucas1 and Lucas2 represent manipulated/intervened upon datasets. Test results are obtained after 50 epochs through the data with $\eta_\lambda = 10^{-6}$.

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, so as we increase α we seive out most of the other features and keep the causally relevant features which may generally lead to poorer performance as is evidenced in the table.

For the manipulated datasets, Lucas1 and Lucas2 we notice the opposite effect. There is better accuracy as α is increased as expected because the causes remain predictive after the data is manipulated. A higher α biases the relevance profile towards causally relevant features hence the increase in performance. While generally this is true, the performance tends to be sensitive to the values of α and the RLVQ learning rate η_α.

	Lucas Dataset		
α	Lucas0	Lucas1	Lucas2
0	0.1820	0.1700	0.1800
0.3	0.1860	0.1680	0.1780
0.7	0.1860	0.1690	0.1780
1.0	0.1840	0.1680	0.1740
1.3	0.1850	0.1660	0.1740

(a) Results



(b) RLVQ as λ changes

Table 1: Test Error results for RLVQ and CRLVQ for the different datasets Lucas0, Lucas1 and Lucas2.

7 Conclusion

This paper attempts to leverage techniques in causal learning and discovery and apply them 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, which is a common scenario in real applications. While our scheme is theoretically plausible, its actual formulation is still not very specific. This paper attempts a possible proof of concept and as such has several limitations in its present formulation; for example the scheme will only work if there are more than one cause and the causes are not related to each other. Also if causative cycles are present in the data, unreliable performance will result.

The introduction of α is interesting because it can be tuned across a scale of RLVQ to CRLVQ and in a sense can depict the degree of certainty of manipulation of the test set. Presently this parameter is set arbitrarily, future work will look into how to learn this parameter as well. We also used a basic derivative of LVQ which has inherent problems in convergence and generalization of the classifier, however since CRLVQ mainly concerns how to update the relevance vector, we believe it can be easily extended to the different derivatives of LVQ including GRLVQ[7], GMLVQ[6] and others. In future we will attempt to extend the update to these different schemes as well. Our formulation of the V-structures is not very specific and we have tried to highlight the short comings in different sections of this paper, in future providing a proper formulation will be one of our key concerns.

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

References

[1] 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.

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

[3] Patrik Hoyer, Shohei Shimizu, Aapo Hyvärinen, Yutaka Kano, and Antti Kerminen. New permutation algorithms for causal discovery using ica. In Justinian Rosca, Deniz Erdogmus, José Príncipe, and Simon Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 115–122. Springer Berlin / Heidelberg, 2006.

[4] Alexander Medvedovsky, Vineet Bafna, Uri Zwick, and Roded Sharan. An algorithm for orienting graphs based on cause-effect pairs and its applications to orienting protein networks. In Keith Crandall and Jens Lagergren, editors, *Algorithms in Bioinformatics*, volume 5251 of *Lecture Notes in Computer Science*, pages 222–232. Springer Berlin / Heidelberg, 2008.

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

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

[7] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.

[8] B. Hammer, F.M. Schleif, and T. Villmann. On the generalization ability of prototype based classifiers with local relevance determination. Technical Report Ifi-05-14, Clausthal University of Technology, 2005.

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

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