Computer aided diagnosis under the influence of heterogeneous data and imbalanced classes

Sreejita Ghosh[1], Elizabeth Sarah Baranowski[2], Rick van Veen[3]
Gert-Jan de Vries[4], Michael Biehl[1], Wiebke Arlt[2], Peter Tino[5], Kerstin Bunte[1]

1- University of Groningen - JBI of Mathematics and Computer Science, NL
2- University of Birmingham - IMSR, UK
3- Philips Research, UK
4- Philips Research - Department of Chronic Disease Management, NL
5- University of Birmingham - School of Computer Science, UK

Introduction Some of the challenges in biomedical data are heterogenous measures, missingness, and imbalanced classes. For rare diseases where the number of patients available for studies is limited and the imbalanced class problem becomes prominent. For such datasets, optimizing the overall class accuracy of the classification technique is not suitable, and high detection rate of the minority classes is particularly desirable. We investigated two main strategies for learning from imbalanced data: 1) introduce distinct costs to the training samples [1], or 2) re-sample the original dataset by either under-sampling the majority class and/or over-sampling the minority classes [2]. Specific genetic mutation result in inherited or inborn disorders of steroidogenesis, and lead to defective production of any of the enzymes or a cofactor responsible for catalysing salt and glucose homeostasis, sex differentiation and sex specific development. Inborn steroidogenic disorders need to be diagnosed as early as possible, to avoid delaying of lifesaving glucocorticoid therapy for adrenal insufficiency, and to facilitate gender allocation and surgical planning in patients with disordered sex development. Our data set, collected at the University of Birmingham, consist of urine GC/MS measurements from 829 healthy controls (305 under 1 year of age) and 118 genetically confirmed patients with steroidogenic disorders. Data samples are presented as 165 dimensional ratio vectors of 34 distinct steroid metabolite concentrations constructed using domain knowledge. An approach for the computer-aided diagnosis of the most prevalent condition, 21-hydroxylase deficiency (CYP21A2), and two other representative, 5α-reductase type 2 deficiency (SRD5A2) and P450 oxidoreductase deficiency (PORD), and simultaneously handling missing and heterogenous measurements in the urine data has been introduced in [3]. In this abstract we present an overview of the state-of-the-art techniques to deal with imbalanced classes for training two classifiers suitable for learning with missing values. We submitted this work to the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) 2017.

Angle LVQ In angle LVQ we assume z-transformed vectorial measurements (zero mean, unit standard deviation) accompanied by labels $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, a number of labelled prototypes $\{(\boldsymbol{w}_m, c(\boldsymbol{w}_m))\}_{m=1}^{M}$ representing the classes and relevances $R = \text{diag}(\boldsymbol{r})$ to weight the dimensions. Classification is performed following a Nearest Prototype Classification (NPC) scheme, where a new vector is assigned the class label of its closest prototype. The dissimilarity of each data sample $\boldsymbol{x}_i$ with its nearest correct prototype with $y_i = c(\boldsymbol{w}_J)$ is defined by d_i^J and by d_i^K for the closest wrong prototype ($y_i \neq c(\boldsymbol{w}_K)$). Both prototypes and relevances $R = \text{diag}(\boldsymbol{r})$ are determined by a supervised training procedure minimizing the following cost function [4] calculated on the observed dimensions,

$$E = \sum_{i=1}^{N} \mu(s) \quad \text{with} \quad \mu(s) = \frac{d_i^J - d_i^K}{d_i^J + d_i^K} \ . \tag{1}$$

In contrast to Generalized Relevance LVQ [4, 5], in angle LVQ the distances $d_i^{\{J,K\}}$ are replaced by angle-based dissimilarities,

$$d_i^L = \Phi\left(\frac{\boldsymbol{x}_i^\top R \boldsymbol{w}_L}{\sqrt{\boldsymbol{x}_i^\top R \boldsymbol{x}_i}\sqrt{\boldsymbol{w}_L^\top R \boldsymbol{w}_L}}\right) \text{ with } L \in \{J, K\} \tag{2}$$

$$\text{and } \Phi(b) = \ g_\beta(b) = \frac{\exp\{-\beta(b+1)\} - 1}{\exp(2\beta) - 1} \quad \text{or} \quad \Phi(b) = \frac{1}{2} - \frac{b}{2} \ . \tag{3}$$

The function Φ transforms the weighted dot product $b = \cos \Theta_R \in [-1, 1]$ to a dissimilarity $\in [0, 1]$ either linearly or using an exponential function g_β with slope β.

Undersampling Undersampling artificially reduces the majority class by randomly selecting k samples used for training to reduce the difference in comparison to the number of samples available for the minority class. Undersampling can improve the sensitivity of the models but exhibits the risk of discarding useful data by reducing the size of training samples.

Oversampling With oversampling, new training samples are artificially synthesized to increase the minority class. We have applied Synthetic Minority Over-sampling Technique (SMOTE), proposed in [2]. Additionally we proposed a variant of SMOTE which generates samples on a hypersphere since the decision of our classifier ALVQ takes place on it.

Costfunction Weighting Another strategy for imbalanced classes we analysed is the introduction of explicit misclassification costs [1]. We introduce a hypothetical cost matrix $\Gamma = \gamma_{cp}$, with $\sum_{cp}^{C} \gamma_{cp} = 1$. The rows correspond to the actual classes c and columns denote the predicted classes p. We include those costs in our cost function Eq. (1),

$$\hat{E} = \sum_{i=1}^{N} \frac{\gamma_{cp(\boldsymbol{x}_i)}}{n_c} \mu(\boldsymbol{x}_i) \quad , \tag{4}$$

where $c = y_i$ is the class label of sample $\boldsymbol{x}_i$, n_c defines the number of samples within that class and p being the predicted label (label of the nearest prototype).

Experiments We performed 5 fold crossvalidation combined with several random initializations in each fold. ALVQ was trained using comparable settings in each of the experiments and one prototype per class. The models are trained on the 4 classes mentioned before, however for simplicity we report the performance on the main imbalance between the negative class (healthy) and the positive class (combination of the patients). Therefore, we investigate Matthews correlation coefficient (MCC) calculated from the confusion matrix, as proposed for imbalanced classes in [6], and the AUC from the ROC curve.

Results Undersampling suffered from omitting training samples. The AUC when the minority class was oversampled by 200% and 400%, on the hypersphere were 98.6(0.4) and 98(0.3) respectively. The AUC from the cost function approach was 99.3(0.2), but this approach may lead to a lot of false positives. However, with an adaptation of the threshold in the the ROC curve this problem can be solved as depicted by the excellent AUC values. The results indicate that the strategies mentioned in this abstract can be efficient in learning from heterogeneous and imbalanced data.

References

[1] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing misclassification costs. In *Proc. of the 11th ICML*, San Francisco, 1994. Morgan Kauffmann.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] K. Bunte, E. S. Baranowski, W. Arlt, and P. Tino. Relevance learning vector quantization in variable dimensional spaces. Workshop of the GI-Fachgruppe Neuronale Netze and the German Neural Networks Society in connection to GCPR 2016, pages 20–23, Hannover, Germany, August 2016. LNCS.

[4] A. S. Sato and K. Yamada. Generalized learning vector quantization. In *Advances in Neural Information Processing Systems*, volume 8, pages 423–429, 1996.

[5] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8–9):1059 – 1068, 2002.

[6] Gary M. Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *J. Artif. Int. Res.*, 19(1):315–354, October 2003.