ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Comparison of strategies to learn from imbalanced classes for computer aided diagnosis of inborn steroidogenic disorders

Sreejita Ghosh[1], Elizabeth S. 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 - Johann Bernoulli Institute, NL

2- University of Birmingham - Institute of Metabolism and Systems Research, UK

3- Philips Research, and De Montfort University, UK

4- Philips Research - Healthcare, NL

5- University of Birmingham - School of Computer Science, UK

Abstract. In the bio-medical domain, a high detection rate of possibly rare diseases is usually highly desirable while errors in the majority class (e.g. healthy controls) may be more acceptable. Hence, optimizing the overall predictive accuracy is often unsuitable. Here, we analyse a large data set of urine GC/MS measurements from 829 controls and 68 patients suffering from one of three inborn steroidogenic disorders. We use 2 comparable algorithms able to handle large amounts of missing data. Furthermore, we compare a variety of different strategies to deal with the highly imbalanced data, including undersampling, oversampling and the introduction of class-specific costs.

1 Introduction

Some of the challenges in biomedical data are heterogeneous measurements, missingness, and imbalanced classes. The bio-medical data analysed in this paper confronts us with all those problems and this contribution focuses on the third issue. For rare diseases, where the number of patients available for studies is limited, the imbalanced class problem becomes prominent. For such datasets, optimizing the overall class accuracy of the classification technique is not suitable, since high detection rate of the minority classes is particularly desirable. Specific genetic mutations result in inherited or inborn disorders of steroidogenesis, leading to defective production of specific enzymes or a cofactor responsible for catalysing salt and glucose homeostasis, sex differentiation and sex specific development. These disorders need to be diagnosed as early as possible, to avoid delays of lifesaving glucocorticoid therapy for adrenal insufficiency, and to facilitate gender allocation and surgical planning in patients with disordered sex development. In [1] an approach for the computer-aided diagnosis of the most prevalent condition has been introduced. In this paper we provide a thorough comparison of state-of-the-art techniques for learning from imbalanced data by, 1) introducing distinct costs to the training samples [2], or 2) re-sampling the original dataset by either under-sampling the majority class and/or over-sampling the minority classes [3]. Section 2 introduces the dataset, the strategies for imbalanced classes including our adaptation with respect to the missingness, and a short explanation of the classifiers. Section 3 contains the experiments and results and in the last section we present conclusions.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

2 Methods

Here we briefly introduce the data set and the two LVQ variants suitable for classification confronted with missing data. Finally, we explain the strategies to handle imbalanced data, the issue we are focusing on in this contribution.

2.1 Data set

We study a large data set collected at the University of Birmingham, comprising urine GC/MS measurements from 829 healthy controls (305 under 1 year of age) and 118 patients with genetically confirmed steroidogenic disorders. Inborn steroidogenic disorders is primarily present in the paediatric population. Specifically we consider P450 oxidorectase deficiency (PORD), 5α-reductase type 2 deficiency (SRD5A2),21-hydroxylase deficiency (CYP21A2) with 18, 21 and 29 samples each. CYP21A2 and POR deficiency patients get more similar treatment whereas that for SRD5A2 differs. Therefore, we investigate the multi-class classification problem and refrain from comparison with ROC-LVQ, which is a LVQ variant proposed for explicit optimization of the receiver operating characteristics (ROC) in an imbalanced 2-class problem [4]. The data samples are presented as $N = 165$ dimensional ratio vectors which are extracted from the original 34^2 possible ratios (34 =number of distinct steroid metabolite concentrations) using prior knowledge. The measurements in the dataset are very heterogeneous, due to the large variation in subject age groups (ranging from neonates, infants, and adults) and the combination of different disease studies.

2.2 Angle LVQ

Angle Learning Vector Quantization (ALVQ) introduced in [1], is an extension to Generalized Relevance LVQ (GRLVQ) [5, 6]. It focused on solutions for missing and heterogeneous measurements and the influence of the imbalanced classes was not systematically investigated. We assume z-score transformed vectorial measurements (zero mean, unit standard deviation) of N=165 dimensions, accompanied by labels $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, a number of labelled prototypes $\{(\boldsymbol{w}_m, c(\boldsymbol{w}_m))\}_{m=1}^M$ to represent the classes and relevances R =diag($\boldsymbol{r}$) to weight the dimensions. Classification is performed following a Nearest Prototype Classification (NPC) scheme, where a new vector is assigned to the class label of its closest prototype. The dissimilarity of each data sample $\boldsymbol{x}_i$ from the nearest correct prototype with $y_i = c(\boldsymbol{w}_J)$ is denoted 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 are determined by a supervised training procedure minimizing the following cost function [5]:

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

In contrast to GRLVQ, in ALVQ the distances $d_i^{\{J,K\}}$ are replaced by angle-based dissimilarities calculated on the available dimensions of $\boldsymbol{x_i}$.

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

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

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

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. This variant of LVQ is particularly suitable for heterogeneous data with missingness. More details and the derivatives can be found in [1].

2.3 NaNLVQ

Like ALVQ, NaNLVQ is apt for data with missingness. During training, NaNLVQ updates only those dimensions of the prototypes which are available in the presented sample [7]. NaNLVQ uses the Partial Distance Strategy [8] to accommodate for incomplete samples by taking into account only the available dimensions in the presented sample, for calculating the distance between that sample and the prototype. Conceptually it is equivalent to imputing the missing value with the current value of the corresponding prototype component [7].

2.4 Undersampling

One of the most popular means to address imbalance is undersampling. Undersampling artificially reduces the majority class by randomly selecting t samples used for training to reduce the difference in comparison to the number of samples available for the minority class. This method exhibits the risk of discarding useful data by reducing the size of training set.

2.5 Oversampling with SMOTE

Oversampling artificially increases the minority class by synthesizing new training samples. In [3] the Synthetic Minority Over-sampling Technique (SMOTE) has been proposed. Dependent on the desired number of synthetic samples ψ out of k nearest neighbours are chosen for each sample in each of the classes. Between these neighbouring samples from the same class a synthetic sample is generated $\boldsymbol{s} = \boldsymbol{x} + \alpha \cdot (\boldsymbol{x}_\psi - \boldsymbol{x})$ with $\alpha \in [0, 1]$ and $\boldsymbol{x}_\psi \in \mathcal{N}_{\boldsymbol{x}}$. Since our data contains missing values we use the NaNLVQ and Angle LVQ dissimilarities for the computation of the nearest neighbours with SMOTE.

2.6 Oversampling on the Hypersphere

Angle LVQ classifies on the unit hypersphere, therefore we propose here a geodesic SMOTE variant generating synthetic samples on the hypershere. We use an important tool of Riemannian geometry, which is the exponential map [9, 10]. The exponential map has an origin M which defines the point for the construction of the tangent space T_M of the manifold. Let P be a point on the manifold and $\hat{P}$ a point on the tangent space than $\hat{P} = \mathrm{Log}_M P$, $P = \mathrm{Exp}_M \hat{P}$ and $d_g(P, M) = d_e(\hat{P}, M)$ with d_g being the geodesic distance between the points on the manifold and d_e being the Euclidean distance on the tangent space. The Log and Exp notations denote a mapping of points from the manifold to the tangent space and vice versa. In our case we present a point $\boldsymbol{x}$ from class c on the unit sphere with fixed length $|\boldsymbol{x}| = 1$, which becomes the origin of the map and the tangent space (the centre of the hypersphere is the origin). We find k nearest neighbours $\boldsymbol{x}_\psi \in \mathcal{N}_{\boldsymbol{x}}$ of the same class as selected sample $\boldsymbol{x}$ using the angle between the vectors $\theta = \cos^{-1}(\boldsymbol{x}^\top \boldsymbol{x}_\psi)$. Each random neighbour $\boldsymbol{x}_\psi$ is now

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

projected onto that tangent space using only the present features and the Log_M transformation for spherical manifolds:

$$\hat{\boldsymbol{x}}_\psi = \frac{\theta}{\sin\theta}(\boldsymbol{x}_\psi - \boldsymbol{x}\cos\theta) \tag{4}$$

Next, a synthetic sample is produced on the tangent space as before $\hat{\boldsymbol{s}} = \boldsymbol{x} + \alpha \cdot (\hat{\boldsymbol{x}}_\psi - \boldsymbol{x})$. The new angle $\hat{\theta} = |\hat{\boldsymbol{s}}|$ is then used to project the new sample back to the unit hypersphere by the Exp_M transformation:

$$\boldsymbol{s} = \boldsymbol{x}\cos\hat{\theta} + \frac{\sin\hat{\theta}}{\hat{\theta}}\hat{\boldsymbol{s}} \tag{5}$$

This procedure is repeated with another sample from the class until the desired number of training samples is reached for that class.

2.7 Costfunction Weighting

The last strategy for imbalanced classes we analyse here is the introduction of explicit misclassification costs [2]. In case the classifiers confuse one condition for another, the patient still receives treatment. However if a healthy person is classified as a patient then the hormone therapy on him/her will have antagonistic effect on his/her health. Similarly, if a patient is misclassified as healthy then he/she does not receive the life-saving treatment. Therefore, we introduce a hypothetical cost matrix $\Gamma = \gamma_{cp}$ with rows corresponding to the actual classes c and columns denoting the predicted classes p. We assume $\sum_{cp}^{C} \gamma_{cp} = 1$ and include γ_{cp} as weighting factors in the cost function:

$$\hat{E} = \sum_{i=1}^{n} \frac{\gamma_{cp(\boldsymbol{x}_i)}}{n_c}\mu(\boldsymbol{x}_i) \ , \tag{6}$$

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).

3 Experiments

We performed 5 fold cross-validation combined with several random initializations of prototypes in each fold. The ALVQ and NaNLVQ were trained using comparable settings without regularization and one prototype per class. We ran 2 variants of ALVQ using the linear and exponential dissimilarities (eq. (3)) referred to as ALVQ and ALVQ$_\beta$ respectively. The models were trained on all 4 classes, however for brevity we report how well the positive class (combination of the 3 disease classes) can be distinguished from the negative class (healthy). Note that the full confusion matrix is still available for analysis[1]. We investigate Matthews correlation coefficient (MCC) calculated from the confusion matrix, as proposed for imbalanced classes in [11], and the area under curve (AUC) from the ROC curve. The experimental settings for comparison are as follows:

- E1: **Baseline:** training on the 5 folds containing the original imbalanced classes with 5 random initializations in each fold for the LVQ variants.
- E2: **Undersampling:** each of the 5 folds has about 165 healthy controls, and in each of the random initializations of each fold, 70 or 140 samples are randomly selected to form the training set.

[1] supplementary material www.cs.rug.nl/∼kbunte/material/ConfusionMatrices.pdf

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

E3: **Oversampling:** each class of patients is oversampled with SMOTE by 5 times producing $q\%$ ($q \in \{200, 400\}$) synthetic points in each fold.

E4: **Geodesic Oversampling:** each class of patients is oversampled with geodesic SMOTE by 5 times synthesizing $q\%$ ($q \in \{200, 400\}$) new data points from the training set in each of the folds.

E5: **Cost weighting:** we train ALVQ on the imbalanced 5 folds with cost function as described in section 2.7. Here, we use the cost matrix $\gamma_{cp} = \frac{10}{16}$ $\forall(c = 1 \wedge p \neq 1)$, $\gamma_{cp} = \frac{10}{16}$ $\forall(p = 1 \wedge c \neq 1)$ and $\gamma_{cp} = \frac{1}{16}$ for the rest.

The results are summarized in Table 1.

Table 1: Mean performance (and std) of the methods in the experiments on the test set and full data. We report AUC and MCC combining the prediction of the diseases as positive and the healthy controls as negative class.

Experiment		**Test set** (mean (std))				**All data** (mean (std))			
		MCC		AUC		MCC		AUC	
Baseline:	ALVQ	0.228	(.054)	0.763	(.048)	0.227	(.030)	0.769	(.030)
	$\text{ALVQ}_{\beta=1}$	0.799	(.069)	0.981	(.008)	**0.794**	(.021)	**0.981**	(.003)
	NaNLVQ	0.708	(.152)	0.933	(.046)	0.753	(.111)	0.937	(.048)
E2(140)**:**	ALVQ	0.804	(.058)	0.983	(.016)	0.293	(.025)	0.869	(.014)
	$\text{ALVQ}_{\beta=1}$	0.879	(.031)	0.995	(.006)	0.320	(.030)	0.891	(.017)
	NaNLVQ	0.885	(.028)	0.994	(.004)	**0.627**	(.038)	**0.929**	(.009)
E2(70)**:**	ALVQ	0.825	(.049)	0.968	(.014)	0.437	(.055)	0.925	(.015)
	$\text{ALVQ}_{\beta=1}$	0.913	(.044)	0.987	(.009)	**0.421**	(.042)	**0.938**	(.012)
	NaNLVQ	0.861	(.059)	0.991	(.011)	0.538	(.061)	0.913	(.015)
E3(200)**:**	ALVQ	0.416	(.174)	0.889	(.066)	0.429	(.194)	0.885	(.063)
	$\text{ALVQ}_{\beta=1}$	0.760	(.100)	0.983	(.010)	**0.801**	(.022)	**0.987**	(.003)
	NaNLVQ	0.800	(.188)	0.955	(.043)	0.844	(.099)	0.963	(.048)
E3(400)**:**	ALVQ	0.352	(.082)	0.871	(.077)	0.369	(.054)	0.872	(.056)
	$\text{ALVQ}_{\beta=1}$	0.738	(.068)	0.979	(.008)	**0.763**	(.001)	**0.983**	(.002)
	NaNLVQ	0.802	(.185)	0.959	(.048)	0.850	(.108)	0.963	(.050)
E4(200)**:**	ALVQ	0.298	(.074)	0.823	(.063)	0.304	(.042)	0.830	(.040)
	$\text{ALVQ}_{\beta=1}$	0.765	(.080)	0.981	(.012)	**0.797**	(.025)	**0.986**	(.004)
E4(400)**:**	ALVQ	0.302	(.069)	0.828	(.050)	0.306	(.031)	0.832	(.029)
	$\text{ALVQ}_{\beta=1}$	0.798	(.072)	0.987	(.007)	**0.796**	(.020)	**0.986**	(.003)
E5:	ALVQ	0.519	(.090)	0.968	(.021)	0.508	(.087)	0.964	(.014)
	$\text{ALVQ}_{\beta=1}$	0.754	(.086)	0.992	(.006)	**0.761**	(.047)	**0.993**	(.002)

First we notice that ALVQ with linear distance conversion performs consistently worse for this data set. NaNLVQ works quite well but cannot reach the performance of ALVQ with exponential conversion and $\beta = 1$. Even in the baseline setting ignoring the imbalance of the classes the ALVQ has an excellent AUC, but the model with threshold 0 (NPC) in the ROC curve exhibits a

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

large number of false negatives (FNs) (low MCC). The NPC models using the strategies to handle the imbalance (sections 2.4 to 2.7) on the other hand cause a larger number of false positives (FPs) (again, low MCC). MCC is computed from the NPC, so if the working point on the ROC is very much biased towards one of the classes it leads to a large FP or FN rate, both decrease the MCC. But AUC takes all possible working points on the ROC into account and is therefore more robust measure. We also find that undersampling suffers from omitting training samples while oversampling performs quite well when the number of samples in the minority classes is tripled. The cost function based approach may lead to a lot of FPs, but with an adaptation of the threshold in the ROC curve this problem can be solved as confirmed by the excellent AUC values.

4 Conclusion

In this contribution we investigated several state-of-the-art strategies to deal with imbalanced classes. Our bio-medical data set consists of urine metabolome profiles of healthy controls and several very rare steroidogenic disorders. The training on the imbalanced data exhibits a lot of FNs, i.e., it misses a lot of patients, although $\text{ALVQ}_{\beta=1}$ performed surprisingly well also on the imbalanced data as shown by the AUC value. We investigated if this could be improved by undersampling, oversampling, a geodesic variant of oversampling as well as cost weighting for the imbalanced classes. These approaches led to a decrease in FNs and increase in FPs, so the MCC remained low, but the AUC improved. From Table 1, cost weighting method seems to be a promising approach for learning from imbalanced data. In the future we would like to see the effect of varying β-value on the performance of ALVQ, and the effect of assigning a misclassification cost to SRD5A2 which is different from that of PORD and CYP21A2.

[1] K. Bunte, E. S. Baranowski, W. Arlt, and P. Tino. Relevance learning vector quantization in variable dimensional spaces. In Barbara Hammer, Thomas Martinetz, and Thomas Villmann, editors, *New Challenges in Neural Computation (NC^2)*, 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.

[2] 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.

[3] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16:321–357, 2002.

[4] T. Villmann, M. Kaden, W. Hermann, and M. Biehl. Learning vector quantization classifiers for ROC-optimization. *Computational Statistics*, pages 1–22, 2016.

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

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

[7] Rick van Veen. Analysis of missing data imputation applied to heart failure data. Masters thesis, University of Groningen, 2016.

[8] E. Eirola, G. Doquire, M. Verleysen, and A. Lendasse. Distance estimation in numerical data sets with missing values. *Information Sciences*, 240:115–128, 2013.

[9] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. on Medical Imaging*, 23(8):995–1005, 2004.

[10] R. C. Wilson, E. R. Hancock, E. Pekalska, and R. P. W. Duin. Spherical and hyperbolic embeddings of data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(11):2255–2269, 2014.

[11] 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, 2003.