Class imbalance in the prediction of dementia from neuropsychological data

Cecília Nunes[1], Dina Silva[2], Manuela Guerreiro[2], Alexandre de Mendonça[2], Alexandra M. Carvalho[3], and Sara C. Madeira[1]

[1] Knowledge Discovery and Bioinformatics (KDBio) Group, INESC-ID and Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal
cnunes@kdbio.inesc-id.pt, smadeira@kdbio.inesc-id.pt

[2] Dementia Clinics, Institute of Molecular Medicine and Faculty of Medicine, University of Lisbon, Lisbon, Portugal
dinasilva@fm.ul.pt, mmgguerreiro@fm.ul.pt, mendonca@fm.ul.pt

[3] Instituto de Telecomunicações (IT) and Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal
alexandra.carvalho@lx.it.pt

Abstract. Class imbalance affects medical diagnosis, as the number of disease cases is often outnumbered. When it is severe, learning algorithms fail to retrieve the rarer classes and common assessment metrics become uninformative. In this work, class imbalance is approached using neuropsychological data, with the aim of differentiating Alzheimer's Disease (AD) from Mild Cognitive Impairment (MCI) and predicting the conversion from MCI to AD. The effect of the imbalance on four learning algorithms is examined through the application of bagging, Bayes risk minimization and MetaCost. Plain decision trees were always outperformed, indicating susceptibility to the imbalance. The naïve Bayes classifier was robust but suffered a bias that was adjusted through risk minimization. This strategy outperformed all other combinations of classifiers and meta-learning/ensemble methods. The tree-augmented naïve Bayes classifier also benefited from an adjustment of the decision threshold. In the nearly balanced datasets, it was improved by bagging, suggesting that the tree structure was too strong for the attribute dependencies. Support vector machines were robust, as their plain version achieved good results and was never outperformed.

Keywords: Alzheimer's disease, dementia, classification, neuropsychological data, class imbalance

1 Introduction

Alzheimer's Disease (AD) is the most common form of dementia among the elderly, affecting 26 million worldwide in 2006 [1]. It remains incurable and its prevalence is estimated to increase given the aging of the world population.

AD progression is categorized in (1) preclinical AD, (2) mild cognitive impairment (MCI) due to AD and (3) dementia due to AD [2]. Characterizing the stage is of the utmost importance for managing the disease, as small delays in AD onset and progression would lead to significant reductions in its global burden [1]. However, the boundaries between the stages are unclear, making this an extremely challenging task [3]. As advanced diagnosis techniques are expensive, invasive and often unavailable, medical doctors rely on neuropsychological assessments. Maximizing the information provided by neuropsychological tests has thus been subject to attention [4, 5]. In this context, given that each MCI patient is subject to cognitive tests several times before progressing to dementia, datasets used for diagnosis contain more MCI evaluations than dementia-labeled evaluations. Furthermore, datasets used for prognosis assimilate the fact that 10-15% of the patients with cognitive complaints progress to dementia each year [2]. Neuropsychological data is hence prone to class imbalance.

Class imbalance is the disparity in the proportions of different classes in datasets used for classification. It affects classification in two ways. First, predictive models neglect the accuracy over the minority. Overcoming this problem involves understanding how classifiers are affected and proposing solutions. Second, the imbalance makes assessment metrics uninformative. For example, if we consider a majority class that corresponds to 90% of the data, an all-majority classifier has no predictive power and yet leads to a 90% accuracy. In this context, the aim of this work is to study the effect of the imbalance of neuropsychological data on four state-of-the-art algorithms. The classification tasks are differentiating MCI from AD (diagnosis) and predicting the conversion from MCI to AD (prognosis) in patients with cognitive complaints. The algorithms are decision trees (DT), the naïve Bayes (NBayes) classifier, the tree-augmented naïve (TAN) Bayes classifier, and Support Vector Machines (SVMs). They are used as base classifiers for the other strategies. Bagging, MetaCost and the minimization of Bayes risk are applied to those classifiers in order to understand their behavior and improve their performance. In addition, assessment metrics for imbalanced data are discussed. Each classifier revealed different behaviors. In particular, DT were unstable and plain SVMs were robust to the imbalance. The best results were achieved when the NBayes was used with risk minimization.

This paper is organized as follows: Section 2 describes the problem of class imbalance and the solutions that have been proposed. Assessment metrics for imbalanced data are briefly discussed. In Section 3, we report the experiment setup, including a description of the data, preprocessing, training and evaluation steps. The results are presented and discussed in Section 4.

2 Learning from imbalanced data

Class imbalance can be defined as the proportion of minority instances over the total number of instances. The majority class is hereafter considered the negative class, and the minority the positive. This type of imbalance is in fact between-class imbalance. It can bias the learners towards the overrepresented class, while the minority may go unlearned. As the imbalance grows, we not longer aim at maximizing accuracy. Instead, we want classifiers to pay more attention to the minority class, without jeopardizing the performance over the majority [6, 7]. The specific consequences of the imbalance depend on the algorithm [8–13].

2.1 Solutions to class imbalance

Proposed solutions for class imbalance can be divided into data-level strategies and algorithm-level or cost-sensitive(CS) strategies [14]. Data-level solutions resample the dataset to obtain optimal class proportions [8, 15]. They involve undersampling the majority class or oversampling the minority. Random resampling has some disadvantages [16]. To overcome the overfitting caused by random oversampling, Synthetic Minority Over-sampling Technique (SMOTE) has been developed [17]. This method may lead to overgeneralization, which can be avoided by adaptive synthetic sampling [18, 19].

Instead of manipulating the data, CS solutions draw the attention of the classifiers to the minority by means of a cost. Costs are scores attributed to correct or incorrect classifications, for instance according to the class. The existence of non-uniform unknown misclassification costs is closely related to class imbalance. The relation between both problems is task-specific and method-specific. Nonetheless, while cost-*in*sensitive solutions lead to sub-optimal performances in both cases [20], CS approaches offer a solution to both problems [21]. In effect, there is an equivalence between varying the class proportion, the class prior probabilities or the misclassification costs [9, 22]. Zadrozny *et al.* [23] divided CS approaches in three categories:

1. *CS model inference*: costs are incorporated directly in the classifier induction algorithm [24, 25]. These techniques are out of the scope of this work since they focus on a single learning algorithm.
2. *CS decision making*: as opposed to minimizing the misclassification rate, class predictions are made according to the minimization of the expected loss [9, 26]. This requires class-membership probabilities to be inferred by the classifiers and knowledge of the costs.
3. *CS ensembles*: cost-sensitivity is introduced into ensemble methods [14, 27]. Examples include CS Boosting [14] and MetaCost [27].

In the following paragraphs, we introduce the minimization of Bayes risk, as a way to minimize the expected loss, together with MetaCost.

Minimizing Bayes risk The scores attributed to different predictions are represented in a cost matrix, as depicted in Table 1. $C(k, j)$ is the cost of classifying a class-j instance as k.

Table 1. Cost matrix

predicted negative	predicted positive	
$C(0, 0)$	$C(1, 0)$	real negative
$C(0, 1)$	$C(1, 1)$	real positive

The optimal solution is the one that minimizes the loss function for all instances, over all class hypotheses [26]. Since the predictions made for each instance are independent, this is equivalent to classifying each instance $\mathbf{X}_i$ with the class k that minimizes a quantity known as *conditional risk* [28]:

$$L(\mathbf{X}_i|k) = \sum\nolimits_i P(j|\mathbf{X}_i)C(k,j), \quad (1)$$

where $P(j|\mathbf{X}_i)$ is the probability of class j given $\mathbf{X}_i$, with $1 \leq i \leq |D|$, and D is the dataset. To use this strategy, class-membership probability estimates are required. It follows that accurate classifications depend on accurate estimates. As such, deciding upon a class can be viewed as estimating a score and comparing it to a decision threshold. In binary classification, most algorithms consider a 0.5 threshold. Changing the cost matrix is a way of adjusting the threshold [9].

But how to tackle the problem of unknown costs? Choosing a cost ratio equivalent to inverting the class proportions would not account for the different severity of the errors [29]. The problems related to class proportion and unknown costs can be tackled by searching for the best cost setup [21]. In this work, the optimal cost setup is sought by means of fixed class cost ratios (see Section 3).

MetaCost This approach led to significant cost reductions in many datasets, while dealing with poor or non-existing probability estimates. It uses bagging [6] to train several weak models. Then, the class-membership probabilities of each instance are estimated by averaging the estimates of the weak models, or through voting in case they are not provided. Using those probabilities, each training instance is relabeled with the class that minimizes the total expected cost (1). Finally, the classifier is trained on the relabeled data to build the final model.

2.2 Assessment metrics for imbalanced learning

The accuracy and the error rate involve ratios between sums of instances of different classes. As such, they are uninformative in imbalanced data [30]. A high accuracy can correspond to a correctly classified majority, and hide a misclassified minority. Alternatively, authors have turned their attention to other metrics. Besides the area under the Receiver Operating Characteristic (ROC) curve, two other single assessment metrics are frequently used in imbalanced learning: the geometric mean (G-mean) [8, 11, 29] and the F-measure [14]. These metrics are defined from the ratios computed from the confusion matrix, such as the True Positive Rate (TPR), also known as recall, the True Negative Rate (TNR), the False Positive Rate(FPR), and the Precision. The G-mean is defined as [31]:

$$G - mean = \sqrt{TPR \times TNR}.$$

The shortcoming of this metric is that it can be optimistic in large imbalances. Given the learning bias, the TNR can be very high regardless of the actual learning ability, and the effect spreads to the G-mean[4]. This shortcoming is largely avoided by the F-measure [32]:

$$F - measure = \frac{(1+\beta^2) \cdot Recall \cdot Precision}{\beta^2 Recall + Precision},$$

[4] This also occurs in ROC space metrics. The learning bias can cause the FPR to be low even if a large number of false positives occurs.

where typically $\beta = 1$, leading to the harmonic mean of recall and precision. Unlike the FPR, which compares the false positives with the total number of negatives, precision compares the false positives with the true positives [32]. In an imbalanced set, the number of true positives is smaller than the total number of negatives and thus negative misclassifications are better captured. In addition, the harmonic mean of two values is closer the lowest of them than the arithmetic mean. As such, a high F-measure assures both a high precision and recall [10]. For these reasons, it is our metric of choice. A limitation of the F-measure is the fact that it disregards the performance of the negative class.

3 Methods

3.1 Data description

The Cognitive Complaints Cohort (CCC) [4] is a study conducted at Instituto de Medicina Molecular (IMM) to investigate dementia on subjects with cognitive complaints. The CCC database comprises the results of neuropsychological tests applied to subjects respecting the inclusion criteria specified by Silva *et al.* [33]. The tests correspond to the Battery of Lisbon for the Assessment of Dementia (BLAD), proposed by Garcia [34]. The battery is validated for the Portuguese population and comprises tests targeting different cognitive domains. The test results are mapped to the stage of dementia provided by medical doctors in the categories: normal, pre-MCI, MCI and dementia due to AD. The latter is simply denoted as AD. The database contains 1642 evaluations of 950 subjects and 162 attributes. Each evaluation is an instance and the attributes are the neuropsychological tests. The original classes are the aforementioned stages.

3.2 Data preprocessing

The first step was the removal of normal and pre-MCI instances. This was followed by the elimination of non-informative attributes, as well as instances of patients evaluated only once, given their uselessness in prognosis. At last, removing instances with more than 90% missing values yielded 677 instances of 336 patients and 157 attributes, with yet nearly 50% of missing values.

The diagnosis of dementia was done considering each evaluation as an independent instance. Dementia prognosis required relabeling the MCI evaluations according to the progression to dementia of the corresponding patient, withing a given time frame. The prognosis classes are evolution (Evol) and non-evolution (noEvol) to dementia, and 2, 3 and 4-year time frames are considered. The datasets were divided into training and validation data. The training data was used to build and test the models through cross-validation (CV). The validation data was used in a final assessment of the models built with the CV data, as in a hold-out (HO) test. The HO data contains 25% of the original data, sampled in a stratified way. No different evaluations of the same patient are contained in both CV and HO datasets. The final datasets are summarized in Table 2.

Preprocessing involved two final procedures: attribute selection and missing value imputation. Correlation-based feature subset selection [35] was performed on the training data and extrapolated to the validation data. Mean-mode missing value imputation was used on the training data. A more sophisticated technique was also tested [36] but it introduced a learning bias. Mean-mode imputation avoided this bias, providing a straightforward solution to the problem.

Table 2. Summary of the datasets, with their learning aim and imbalances.

Learning task	Abbreviation	Imbalance	Minority class
distinguish MCI from AD	CV_Diag	15.7%	AD
predict MCI-to-AD conversion in 2 years	CV_2Y	36.5%	Evol
predict MCI-to-AD conversion in 3 years	CV_3Y	47.6%	noEvol
predict MCI-to-AD conversion in 4 years	CV_4Y	32.4%	noEvol
distinguish MCI from AD	HO_Diag	8.3%	AD
predict MCI-to-AD conversion in 2 years	HO_2Y	22.2%	Evol
predict MCI-to-AD conversion in 3 years	HO_3Y	41.1%	Evol
predict MCI-to-AD conversion in 4 years	HO_4Y	41.9%	noEvol

3.3 Classification

Bagging, risk minimization, and MetaCost were applied to the NBayes classifier, TAN Bayes classifier, DT and SVMs. The strategies were chosen since they can be applied to any classifier and give insight about its behavior. The NBayes classifier was used with kernel density estimation since it showed superior results compared to using a normal distribution, in all datasets. The TAN Bayes classifier was used due to its efficiency and efficacy [37].

When dealing with imbalanced datasets and DT, either no pruning or pruning preceded by Laplace smoothing is advised [38, 39]. Hence, both methods were tested and the best was used. The pruning confidence factor was subject to grid-search in order to maximize F-measure for each dataset. SVMs were used with the polynomial kernel and the Gaussian kernel function. Grid-search was also performed for the SVM parameters. Regarding bagging and MetaCost, results for 10 iterations were considered. Attempts with less iterations led to worse results, while using more than 10 iterations does not provide significant bagging improvements [6]. Given the moderate dataset size, a 100% bag size was used. Implementations were provided by WEKA 3.6 [40].

Regarding the cost-setup, correct predictions were defined to have zero cost, that is $C(0,0) = 0$ and $C(1,1) = 0$. Since the cost matrix is invariant to multiplication by a positive factor, the majority class error cost was kept equal to one, and the minority error cost was varied. The goal was then to find the optimal misclassification cost ratio (MCR):

$$MCR = \frac{C(0,1)}{C(1,0)} = C(0,1),$$

which corresponds to the value that maximizes F-measure. Empirical tests showed that the optimal MCRs were never superior to 14 (obtained applying risk minimization to DTs). The second highest MCR was 8 (obtained using MetaCost with the NBayes). Since the difference in F-measure between the two cases was only 0.02, the MCR search interval was restricted to $[1, 8]$ with a step of 0.25.

The models were built using 10-fold CV, performed 10 times with different random seeds. The CV partitions were the same for all methods. A Friedman ranks test and its post-hoc Nemenyi pairwise comparisons were applied, as advised when testing more than two algorithms over multiple datasets [41]. Rejecting the null hypothesis of the Friedman test means that at least two of the results of applying the base classifier, bagging, risk minimization and MetaCost come from populations with different medians [42], that is significant differences in the performances were found. A significance level of 0.05 was considered.

4 Results and discussion

In this section, we first present and discuss the results obtained for each classifier. The average values of F-measure can be observed in Table 3 and the results of the Friedman test and the pairwise comparisons are depicted in Table 4. For the CS methods, the MCR that maximized the F-measure was selected (Table 6 in Appendix A). Finally, we compare the best strategies for each classifier (Table 5).

Table 3. F-measure averaged over 10 CV experiments for each dataset, base classifier and meta-learning/ensemble method. The base classifier is the classifier without any method. For the CS methods, the MCR that maximized F-measure was selected.

classifier	dataset	Base classifier	Bagging	Risk min	MetaCost
DT	CV_Diag	0.464	0.492	0.504	**0.531**
	CV_2Y	0.514	0.551	0.578	**0.621**
	CV_3Y	0.679	**0.729**	0.692	0.695
	CV_4Y	0.579	0.607	0.604	**0.630**
	Average	0.559	0.595	0.594	**0.619**
NBayes	CV_Diag	**0.598**	0.596	0.597	0.583
	CV_2Y	0.636	0.640	**0.693**	0.655
	CV_3Y	0.745	0.746	**0.784**	0.737
	CV_4Y	0.675	0.679	**0.693**	0.661
	Average	0.664	0.665	**0.692**	0.659
TAN	CV_Diag	0.548	0.557	0.551	**0.583**
	CV_2Y	0.623	0.631	0.653	**0.655**
	CV_3Y	0.744	**0.769**	0.748	0.737
	CV_4Y	0.625	**0.668**	0.660	0.661
	Average	0.635	0.656	0.653	**0.659**
Polynomial-kernel SVM	CV_Diag	0.550	0.561	**0.569**	0.563
	CV_2Y	0.649	0.643	**0.681**	0.659
	CV_3Y	**0.74**	0.740	0.763	0.770
	CV_4Y	**0.722**	0.713	0.714	0.715
	Average	0.667	0.664	**0.682**	0.677
Gaussian-kernel SVM	CV_Diag	0.556	**0.579**	**0.579**	0.576
	CV_2Y	0.651	0.644	**0.685**	0.655
	CV_3Y	0.733	0.727	0.760	**0.771**
	CV_4Y	**0.712**	0.703	0.695	0.697
	Average	0.663	0.663	**0.680**	0.675

4.1 Decision Trees

DT usually lack robustness to the imbalance. They tend to grow mixed leaves with few minority instances that get disregarded. In addition, minority instances may end up isolated in single leaves, leading to overfitting [8]. Accordingly, in Table 4 it is possible to observe that plain DT were outperformed by all other methods. Both bagging and MetaCost improved the performance of DT, which meets the expectations given their instability [6]. Although DTs can suffer from a learning bias, they provide inaccurate class-membership probabilities, and are therefore bad candidates for risk minimization. Indeed, the bias was preferably tackled through MetaCost, since it led to the highest values of F-measure in most datasets. In the most balanced dataset, the CV_3Y dataset, MetaCost was outperformed by bagging indicating the absence of the bias.

Table 4. Results of the Friedman tests and post-hoc pairwise comparisons over all datasets. For each classifier, rejecting the null hypothesis corresponds to finding significantly different performances among all methods. Each entry indicates if the F-measure obtained with the method in the corresponding column was significantly greater or smaller than the F-measure obtained with the method in the row. The entry is blank in case the comparison revealed no statistically significant difference.

	Base classifier	Bagging	Risk min	MetaCost	Best strategy
Decision Trees	p-value=1.37E-10				Bagging, Risk min and MetaCost
Base classifier	-	greater	greater	greater	
Bagging	smaller	-			
Risk min	smaller		-		
MetaCost	smaller			-	
NBayes	p-value=3.34E-12				Risk min
Base classifier	-		greater		
Bagging		-	greater	smaller	
Risk min	smaller	smaller	-	smaller	
MetaCost		greater	greater	-	
TAN	p-value=5.96E-5				Risk min and MetaCost
Base classifier	-		greater	greater	
Bagging		-			
Risk min	smaller		-		
MetaCost	smaller			-	
Polynomial-kernel SVM	p-value=1.96E-3				Base classifier
Base classifier	-				
Bagging		-	greater	greater	
Risk min		smaller	-		
MetaCost		smaller		-	
Gaussian-kernel SVM	p-value=3.64E-3				Base classifier
Base classifier	-				
Bagging		-			
Risk min			-		
MetaCost				-	

4.2 Naïve Bayes classifier

In the NBayes classifier, computing the *Maximum a Posteriori Hypothesis* for the class involves estimating the class prior probabilities and the conditional probabilities. The imbalance mainly affects the prior probabilities, while the conditional probabilities are calculated for each class [9]. Therefore, although naïve Bayesian class-membership probability estimates are inaccurate [43], when test instances are ranked according to them, they tend to be ordered according to the class [9]. The decision threshold may thus benefit from an adjustment, which seems to have been the case in our results. Applying risk minimization to the NBayes classifier was statistically superior to all the approaches (Table 4).

On the other hand, the fact that conditional probabilities are skew-independent makes the NBayes robust to the imbalance. Indeed, the plain NBayes led to the greatest F-measure in the most imbalanced dataset (Table 3).

Since the NBayes is a stable algorithm, strategies involving bagging are typically not suitable [6]. This is clear in the results, given that bagging had no benefit compared to the other methods, including the plain classifier. Moreover, MetaCost was outperformed by risk minimization.

4.3 Tree-Augmented Naïve Bayes classifier

Table 4 shows that the best methods for the TAN Bayes classifier were risk minimization and MetaCost, while bagging provided no benefit. As the NBayes, the TAN Bays classifier may benefit from an adjustment of the decision threshold, given the potential bias in prior probabilities. The effect of using risk minimization on this classifier has not been extensively studied in the literature. One study showed that the TAN Bayes was improved compared to the NBayes, when risk minimization or resampling followed by F-measure threshold optimization were employed [44]. However, if minority class dependencies are incorrectly modeled by the TAN Bayes network [10], shifting the threshold does not seem to be adequate. The effect of risk minimization on this learner is thus unpredictable. In our results, although risk minimization outperformed the plain classifier, it did not maximize F-measure in any dataset.

TAN Bayes classifiers are not good candidates for bagging [45]. However, imposing a tree structure to rare data can be too strong and lead to overfitting. In this case, bagging-based methods can be useful. This seems to have been the case, as MetaCost maximized the F-measure in the most imbalanced datasets. In the nearly-balanced datasets, CV_3Y and CV_4Y, the best results were obtained with bagging, possibly given the absence of a bias in the prior probabilities.

4.4 Support Vector Machines

Two behaviors were described for SVMs in imbalanced data [11]. If the imbalance is moderate, they perform well, while in severe imbalances, SVMs are likely to classify everything as majority. Plain SVMs were never outperformed by the other methods, indicating their robustness (Table 4). Indeed, the greatest CV imbalance is 15.7%, which is a moderate imbalance. In Gaussian-kernel SVMs, different datasets were better modeled by different methods. No method was statistically superior to the others.

SVMs do not predict class-membership probabilities with high accuracy. Furthermore, in case the data is non-separable, biasing the output of the model does not provide any benefit. Nonetheless, the highest values of F-measure for the most imbalanced datasets were obtained through risk minimization with polynomial-kernel SVMs, revealing that a learning bias was present. In spite of SVM stability, if the dataset is small or the minority is rare, SVMs can overfit the data and thus benefit from bagging [13]. A fact that may corrupt the success of bagging is that SVM parameters are optimized for one of the 10 rounds of CV. This optimization is lost when the training data is changed. This appears to have been the case of polynomial-kernel SVMs, as bagging did not improve the base classifier and was outperformed by the CS methods.

A final note goes to the validation results. They were consistent to the CV results for the DT and for the NBayes classifier. DT were always outperformed and the NBayes was preferentially improved by risk minimization. The results obtained for the TAN Bayes classifier and SVMs were not very consistent with the CV results. The TAN Bayes not improved by MetaCost, as was the case in CV. Possibly, the increase in the size of the training data reduced th instability that rendered this classifier suitable for ensembling. The SVMs benefited from risk minimization in the CV test and from MetaCost in the HO test.

4.5 Comparison of the best strategies for each classifier

The best strategies for each classifier were also compared through a Friedman test and Nemenyi pairwise comparisons (Table 5). All strategies outperformed strategies with DTs as base learner, which led to the lowest values of F-measure. Combining risk minimization with the NBayes classifier achieved greater values of F-measure than all other strategies except polynomial-kernel SVMs. Therefore, and given the efficiency and simplicity of the combination the NBayes with risk minimization, it is the preferred strategy.

Table 5. Results of the Friedman tests and post-hoc pairwise comparisons between the best strategies for each classifier. The p-value is 2.41E-41.

	DT+ Bag	DT+ Risk	DT+ MetaCost	NBayes+ Risk	TAN+ Risk	TAN+ MetaCost	Poly SVM	Gaussian SVM
DT+Bag	-			greater	greater	greater	greater	greater
DT+Risk		-		greater	greater	greater	greater	greater
DT+MetaCost			-	greater	greater	greater	greater	greater
NBayes+Risk	smaller	smaller	smaller	-	smaller	smaller		smaller
TAN+Risk	smaller	smaller	smaller	greater	-			
TAN+MetaCost	smaller	smaller	smaller	greater		-		
Polynomial SVM	smaller	smaller	smaller				-	
Gaussian SVM	smaller	smaller	smaller	greater				-

Dataset	F-measure							
CV_Diag	0.49	0.50	0.53	0.60	0.55	0.58	0.55	0.56
CV_2Y	0.55	0.58	0.62	0.69	0.65	0.65	0.65	0.65
CV_3Y	0.73	0.69	0.70	0.78	0.75	0.74	0.75	0.73
CV_4Y	0.61	0.60	0.63	0.69	0.66	0.66	0.71	0.71
Average	**0.59**	**0.59**	**0.62**	**0.69**	**0.65**	**0.66**	**0.67**	**0.66**

5 Conclusions

In this work, we examined the effect of the imbalance of neuropsychological data on DT, NBayes, TAN Bayes and SVMs, in the diagnosis and prognosis of dementia in patients with cognitive complaints. The most consistent results were obtained for DT and NBayes. The first learner benefited from any meta-learning/ensemble strategy, namely MetaCost, while the second is clearly improved by the risk minimization. As a consequence and given the good performances obtained by the NBayes classifier combined with risk minimization, this is our method of choice for reliable and predictable results in neuropsychological data. SVMs were robust to the imbalances, but it was not possible to conclude which method is the best match for it.

Directions for future work include the study of an assessment strategy that can avoid the optimism of the G-mean and complement the F-measure by focusing on the majority class. Moreover, it would be relevant to compare the presented results with a resampling method, such as SMOTE, and observe its effect on each classifier. A final comment goes to the other challenges posed by the data, such as the high attribute dimensionality, and the high percentage of missing values. A deeper study of these problems could reduce the effect of the class skew and make the neuropsychological tests more informative.

Acknowledgments

This work was partially supported by FCT (Fundação para a Ciência e a Tecnologia) under projects: PTDC/EIA-EIA/111239/2009 (NEUROCLINOMICS - Understanding NEUROdegenerative diseases through CLINical and OMICS data integration) and PEst-OE/EEI/LA0021/2011 (INESC-ID multiannual funding). The work of AMC was partially supported by FCT under grant PEst-OE/EEI/LA0008/2011 (IT multiannual funding).

References

1. Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M.: Forecasting the global burden of Alzheimer's disease. Alzheimers dementia the journal of the Alzheimers Association **3**(3) (2007) 186–191
2. Alzheimer's Association: 2012 Alzheimer's Disease Facts and Figures. Technical report, Alzheimer's Association
3. Yesavage, J.A., O'Hara, R., Kraemer, H., Noda, A., Taylor, J.L., Rosen, A., Friedman, L., Sheikh, J., Derouesné, C.: Modeling the prevalence and incidence of Alzheimers disease and mild cognitive impairment. Journal of Psychiatric Research **36** (2002) 281–286
4. Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., Mendonça, A.D.: Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes **4:299** (2011)
5. Lemos, L.: A data mining approach to predict conversion from mild cognitive impairment to Alzheimers Disease. Master's thesis, IST (2012)
6. Breiman, L.E.O.: Bagging Predictors. Machine Learning **24**(2) (1996) 123–140
7. Kearns, M., Valiant, L.: Cryptographic limitations on learning Boolean formulae and finite automata. Journal of the Association for Computing Machinery **41**(1) (1994) 67–95
8. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. Training (1997) 179–186

9. Elkan, C.: The Foundations of Cost-Sensitive Learning. Int. Joint Conf. on Artificial Intelligence **17**(1) (2001) 973–978
10. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of Imbalanced Data: a Review. Int. Journ. of Pattern Recognition and Artificial Intelligence **23**(04) (June 2009) 687–719
11. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. Machine Learning **3201** (2004) 39–50
12. Chang, E.Y., Ece, E., Edu, U.: Class-Boundary Alignment for Imbalanced Dataset Learning. In: ICML 2003 Workshop on Learning from Imbalanced Data Sets. (2003)
13. Tao, D., Tang, X.: Assymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Transactions on Pattern Analysis and Machine IntelligenceAnalysis and Machine Intelligence **28**(7) (2006) 1088–1099
14. Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition **40** (2007) 3358–3378
15. Japkowicz, N.: The Class Imbalance Problem: Significance and Strategies. Complexity **1** (2000) 111–117
16. McCarthy, K., Zabar, B., Weiss, G.: Does cost-sensitive learning beat sampling for classifying rare classes? In: Proceedings of the 1st Int. Work. on Utilitybased Data Mining, ACM Press New York, NY, USA (2005) 69–77
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research **16**(1) (2002) 321–357
18. Han, H., Wang, W., Mao, B.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Learning **3644** (2005) 878–887
19. Garcia, E.a.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence) (3) (June 2008) 1322–1328
20. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter **6**(1) (2004) 40–49
21. Maloof, M.A.: Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. Analysis **21**(9) (2003) 1263–1284
22. Breiman, L., Friedman, J.H., Stone, C.j., Olshen, R.A.: Classification and Regression Trees. (1984)
23. Zadrozny, B., Langford, J., Abe, N.: Cost-Sensitive Learning by Cost-Proportionate Example Weighting. Third IEEE Int. Conf. on Data Mining (2003) 435–442
24. Ting, K.M.T.K.M.: An instance-weighting method to induce cost-sensitive trees (2002)
25. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the Sensitivity of Support Vector Machines. Heart Disease (1999) 55–60
26. Bishop, C.M.: Pattern Recognition and Machine Learning. Volume 4 of Information science and statistics. Springer (2006)
27. Domingos, P.: MetaCost: A General Method for Making Classifiers Cost-Sensitive. Proceedings of the Fifth Int. Conf. on Knowledge Discovery (1999) 155–164
28. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2. edn. Wiley (2001)
29. Thai-nghe, N., Gantner, Z., Schmidt-thieme, L.: Cost-Sensitive Learning Methods for Imbalanced Data. In: The 2010 Int Joint Conf. on Neural Networks. (2010) 1–8
30. Lawrence, S., Burns, I., Back, A., Tsoi, A.C., Giles, C.L.: Neural network classification and prior class probabilities. LNCS **1524** (1998) 299–314

31. Kubat, M., Holte, R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning **30** (1998) 195–215
32. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06 (2006) 233–240
33. Silva, D., Guerreiro, M., Maroco, J.a., Santana, I., Rodrigues, A., Bravo Marques, J., de Mendonça, A.: Comparison of Four Verbal Memory Tests for the Diagnosis and Predictive Value of Mild Cognitive Impairment. Dementia and Geriatric Cognitive Disorders Extra **2**(1) (2012) 120–131
34. Garcia, C.: A Doença de Alzheimer, problemas do diagnóstico clínico. Phd, Universidade de Medicina de Lisboa (1984)
35. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. Methodology **21i195-i20**(April) (1999) 1–5
36. Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., Yumei, C.: A SVM Regression Based Approach to Filling in Missing Values. LNCS **3683** (2005) 581–587
37. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning **29**(1) (1997) 131–163
38. Bradford, J., Kunz, C., Kohavi, R., Brunk, C.: Pruning Decision Trees with Misclassification Costs. In: Eur. Conf. on Machine Learning. (1998) 131–136
39. Provost, F., Domingos, P.: Well-Trained PETs: Improving Probability Estimation Trees (2000)
40. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
41. Demsar, J.: Statistical Comparison of Classifiers over Multiple Data Sets. Journal of Machine Learning Research **7**(7) (2006) 1–30
42. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Volume 51. CRC Press (1997)
43. Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classi er. Machine Learning **29**(2/3) (1997) 105–112
44. Thai-nghe, N., Schmidt-thieme, L., Techniques, A.M.: Learning Optimal Threshold on Resampling Data to Deal with Class Imbalance. In: 8th IEEE Int. Conf. on Computing and Communication Technologies: Research, Innovation, and Vision for the Future. (2010)
45. Quinn, C.J., Coleman, T.P., Kiyavash, N.: Approximating discrete probability distributions with causal dependence trees (2010)

6 Appendix A. Misclassification Cost Ratios

classifier	dataset	Risk min	MetaCost
DT	CV_Diag	7.75	5
	CV_2Y	6.	4
	CV_3Y	5.25	5
	CV_4Y	4.5	3.5
	Average	5.88	4.38
NBayes	CV_Diag	1.75	1
	CV_2Y	7.5	4
	CV_3Y	7.75	6.75
	CV_4Y	1.5	4.75
	Average	4.63	4.13
TAN	CV_Diag	3.5	1
	CV_2Y	2.75	4
	CV_3Y	1.25	6.75
	CV_4Y	3	4.75
	Average	2.63	4.13
Polynomial-kernel SVM	CV_Diag	1.5	1.25
	CV_2Y	2.5	2.5
	CV_3Y	2	1.5
	CV_4Y	1.25	2.5
	Average	1.82	1.94
Gaussian-kernel SVM	CV_Diag	4.5	2.5
	CV_2Y	2.75	2.5
	CV_3Y	2.5	2.25
	CV_4Y	1.25	1.5
	Average	2.75	2.19

Table 6. MCRs that maximized the value of F-measure in the CS methods.

7 Appendix B. Validation results using optimal MCRs

classifier	dataset	Base classifier	Bagging	Risk min	MetaCost
DT	HO_Diag	0.444	**0.5**	0.367	0.391
	HO_2Y	0.267	0.364	**0.4**	0.36
	HO_3Y	0.7	0.776	0.795	**0.815**
	HO_4Y	0.595	**0.629**	0.618	0.6
NBayes	HO_Diag	**0.516**	0.5	**0.516**	0.457
	HO_2Y	0.429	0.429	**0.556**	0.529
	HO_3Y	0.818	0.818	**0.824**	**0.824**
	HO_4Y	0.647	0.686	**0.722**	0.718
TAN	HO_Diag	**0.6**	0.522	0.48	0.5
	HO_2Y	0.581	0.417	**0.6**	0.565
	HO_3Y	0.818	0.769	**0.836**	0.753
	HO_4Y	0.667	**0.686**	0.651	0.619
Polynomial-kernel SVM	HO_Diag	0.5	0.5	0.467	**0.552**
	HO_2Y	0.435	0.381	**0.632**	0.615
	HO_3Y	0.721	0.781	**0.806**	**0.806**
	HO_4Y	0.667	0.667	0.667	**0.684**
Gaussian-kernel SVM	HO_Diag	0.48	0.48	0.488	**0.533**
	HO_2Y	0.435	0.455	0.6	**0.615**
	HO_3Y	0.733	0.762	0.806	**0.841**
	HO_4Y	0.647	0.667	0.647	**0.706**

Table 7. Values of F-measure obtained by training the models on the CV datasets and testing them on the corresponding HO datasets. MCRs in Table 6 were used.