This is a previous version of the paper published in L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255, pp. 335–343, 2017. DOI: 10.1007/978-3-319-58838-4 37

A study of Prototype Selection algorithms for Nearest Neighbour in class-imbalanced problems

Jose J. Valero-Mas, Jorge Calvo-Zaragoza, Juan R. Rico-Juan, and José M. Iñesta

Pattern Recognition and Artificial Intelligence Group,
University of Alicante
{jjvalero,jcalvo,juanra,inesta}@dlsi.ua.es

Abstract. Prototype Selection methods aim at improving the efficiency of the Nearest Neighbour classifier by selecting a set of representative examples of the training set. These techniques have been studied in situations in which the classes at issue are balanced, which is not representative of real-world data. Since class imbalance affects the classification performance, data-level balancing approaches that artificially create or remove data from the set have been proposed. In this work, we study the performance of a set of prototype selection algorithms in imbalanced and algorithmically-balanced contexts using data-driven approaches. Results show that the initial class balance remarkably influences the overall performance of prototype selection, being generally the best performances found when data is algorithmically balanced before the selection stage.

Keywords: kNN, imbalanced data, prototype selection

1 Introduction

The k-Nearest Neighbour (kNN) classifier constitutes one of the most well-known techniques for supervised non-parametric classification, mainly because of its conceptual simplicity and its bounded error rates [1]. Basically, kNN classifies a given input element by assigning the most common label among its k-nearest prototypes of the training set. Such exhaustive search for each element to be classified entails low efficiency figures in both classification time and memory usage, which constitutes the main drawback for this classifier.

Data Reduction (DR) methods are typically considered for tackling this disadvantage [2]. These strategies reduce the training set while trying to keep the classification accuracy of the original data –or even improving it– if noisy elements are removed. Among the different existing possibilities, a relatively straightforward and largely studied methodology known as prototype selection (PS) performs this reduction by selecting a representative subset of the initial training set following a particular heuristic [3].

A large number of works in the classification field assume that the classes of the elements at issue are equally represented. However, this assumption turns out not to be realistic since most data sources do not necessarily exhibit such

equilibrium among the different classes, leading to *class imbalance* problems [4]. In general, the use of such imbalanced data leads to situations in which the performance of the classifier is biased towards the class representing the majority of the elements [5]. In this regard, different strategies have been considered to palliate this issue, being a rather common one the use of data sampling methodologies to artificially equilibrate the class distribution.

In this paper, we aim at studying the behaviour of PS algorithms when dealing with large-scale imbalance datasets in the context of kNN classification. More precisely, the idea is to assess the performance of PS algorithms in class-imbalance situations and compare them to the case in which a sampling-based balancing algorithm is considered as a preprocessing stage before PS.

The rest of the work is structured as follows: Section 2 contextualizes the problem of imbalanced classification; Section 3 presents the experiment proposed as well as the sampling-based balancing and PS techniques considered; Section 4 presents and discusses the results obtained; finally, Section 5 concludes the work and proposes future lines to develop.

2 Classification with imbalanced data

Formally, imbalanced classification tasks refer to the cases in which prior probabilities of the classes at issue significantly differ among them. This particularity generally results in a tendency of the classifier to bias towards the majority class, thus decreasing the overall performance of the system.

Different proposals may be found in the literature to palliate this issue, being typically grouped into three categories [6]: (a) data-level methods that either create artificial data for the minority classes and/or remove elements from the majority one to equilibrate the class representation; (b) algorithmic-level approaches that internally bias the classifier to compensate the skewness in the data; (c) cost-sensitive training methodologies that consider higher penalties for the misclassification of the minority class than for the majority one.

Not all classifiers show this bias towards the majority class. Instance-based algorithms such as kNN report a superior tolerance as they consider all instances during the classification stage. Nevertheless, when this imbalance effect is combined with class overlapping, performance is severely affected [7].

In this work, we study the use of PS algorithms in imbalanced and overlapped scenarios. As aforementioned, PS methods tackle the issues found in kNN related to large and noisy (overlapped) datasets. However, as these processes have not been devised for class-imbalanced sets, it seems necessary to explore their behaviour in such cases and compare the results with the ones obtained if a data-level balancing method is considered as a preprocessing stage.

3 Experimentation

Figure 1 shows the scheme implemented for the experiments. As it can be checked, the *train set* may undergo a class-balancing process and/or a PS method

before getting to the kNN classifier, which are the situations to be compared. For our experiments, we fixed $k = 1$ for the classifier as well as considering Euclidean distance for the dissimilarity measure.

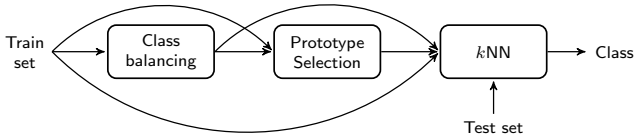

Fig. 1. Scheme proposed for the experiments.

The following sections introduce both the class-balancing techniques considered as well as the PS strategies tested. Also, one last section introducing the evaluation methodology is included.

3.1 Prototype Selection techniques

In terms of PS methods, we have contemplated a comprehensive set of techniques from the ones in the literature.

As examples of the most classic approaches for PS, we have considered the Edited NN (ENN), the Condensed NN (CNN), and the Fast CNN (FCNN). Additionally, we considered EFCNN, which consists of an FCNN process with a previous ENN stage to remove noisy elements.

In terms of more recent approaches, we have considered the use of the Decremental Reduction Optimization Procedure (DROP3) as an example of hybrid approach between the ENN and CNN families. Also, we have tested the performance of the Cross-generational elitist selection, Heterogeneous recombination and Cataclysmic mutation search (CHC), which constitutes a very successful example of genetic algorithm applied to PS.

Finally, we also also studied the use of the Farthest Neighbour (FN) and Nearest to Enemy (NE) algorithms as they constitute representative examples of the so-called *rank methods*. These methods give each prototype a score indicating its relevance with respect to classification accuracy, so that they can be ranked to eventually select a subset of them.

For a comprehensive explanation of the methods the reader is referred to [3] except for the rank ones for which reader is addressed to [8]. For our experiments we set a value of $k = 5$ for all PS schemes.

3.2 Data sampling class balancing

As commented, data-level class balancing techniques equilibrate the classes in the training set by *oversampling* the minority class and/or *undersampling* the

majority one. To assess their relevance in the context of this experiment, we considered a set of examples of each family as well as combinations of them.

Regarding oversampling, we considered the Synthetic Minority Over-sampling Technique (SMOTE, SMT in this work) [9] as well as two existing extensions (B1 and B2 for SMOTE Borderline 1 and 2, respectively) that focus on detecting and remarking transition zones between classes [10].

So as to undersampling, we included Condensing-based undersampling (CNN), Neighborhood cleaning rule (NCL), and Tomek links (TL). Due to space issues, the reader is referred to [11] for a thorough explanation of these methods.

Finally, the combinations of techniques considered comprise all undersampling methods followed by oversampling. We set a value of $k = 5$ for all cases.

3.3 Evaluation

For the experimentation we have considered five datasets from the UJI[1] and the KEEL[2] collections. Additionally, we have considered the music dataset Prosemus[3] that is meant for *onset detection*, ie. the detection of the beginnings of music note events in audio streams, and whose features have been extracted with the methodology in [12]. All these datasets only contain two classes as it constitutes a common practice in studies about imbalanced classification. Also, these sets contain more than 1500 instances so that PS can be reasonably applied. A 5-fold cross-validation scheme has been considered for the experimentation. Table 1 describes these datasets.

Table 1. Description of the datasets considered in terms of the amount of instances of the majority (Maj.) and minority (Min.) classes. Symbols † and ‡ depict sets obtained from the UJI or the KEEL collections, respectively.

Dataset	**Min.**	**Maj.**	**Dataset**	**Min.**	**Maj.**	**Dataset**	**Min.**	**Maj.**
Prosemus	1041	4045	phoneme†	3673	5170	spam†	1813	2788
scrapie†	531	2582	segment0‡	329	1979	yeast3‡	163	1321

Regarding figures of merit, we considered the F-measure (F_1) as it constitutes a typical measure in the context of imbalanced classification. Focusing on the minority class, this metric summarizes the correctly classified elements (True Positives, TP), the misclassified elements from the majority class as minority ones (False Positives, FP), and the misclassified elements from the minority class as majority class (False Negative, FN) in a single value as follows:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{1}$$

[1] http://www.vision.uji.es/∼sanchez/Databases/
[2] http://sci2s.ugr.es/keel/datasets.php
[3] http://grfia.dlsi.ua.es/cm/projects/prosemus/database.php

Note that for the case of the Prosemus set, a tolerance window of 50 ms is given, following the common evaluation procedure for onset detection [12].

Additionally, as pointed out in [13], PS evaluation may be seen as a multi-objective problem with two opposed objectives to be optimized, accuracy and set size. Thus, we shall analyse the results in terms of the non-dominance concept: one solution is said to dominate another when it is better or equal in each of the two objectives considered and, at least, strictly better in one of them; the best solutions, as there may be more than one, are those that are non-dominated.

4 Results

The results obtained are shown in Table 2. These figures depict the average F_1 score and reduction rate (in percentage) obtained for the considered datasets in terms of the balancing techniques and PS strategy used.

According to the results, the use of PS on the initial imbalance situation implies a decrease in the F_1 measure for all cases. For instance, CHC lowers performance in more than 0.15 points in the F_1 measure. However, in this context, the results achieved by FCNN are particularly interesting since, although there is a decrease in performance as in the other cases, F_1 is just slightly lower than the original case (0.2 points) but with less than a third of its set size.

When an oversampling technique is considered, the results show a slight improvement but also implies an increase in the set size. Nevertheless, given that some cases retrieve competitive F_1 results but still with a large reduction rate (for instance, FCNN and EFCNN when considering SMT), this balancing scheme seems appropriate as a preprocessing stage.

Regarding the undersampling schemes, it can be checked that, in general, this balancing process results in slightly worse scores than when oversampling the set. Particularly, the use of the CNN balancing method implies a general decrease in the F_1 results when PS is applied. However, when this CNN method is used without any PS, results are remarkably good as it achieves the same F_1 as in the initial set but with roughly half of its set size. NCL and TL schemes show better performance when coupled with PS as F_1 results get to improve when compared to their corresponding PS schemes in the initial imbalanced situation.

In terms of the combined balancing strategies, it can be checked that, in general, they obtain intermediate figures between the solely use of oversampling or undersampling. For instance, for the ENN selection method scheme, CNN-B1 achieves an $F_1 = 0.49$ with a 49.1 % of set size while the oversampling B1 retrieves an $F_1 = 0.68$ with a set size of 132.2 % and undersampling CNN gets to an $F_1 = 0.60$ with 34.2 % of the initial prototypes. Thus, these solutions may suit cases with medium reduction requirements, being undersampling techniques the ones indicated for drastic size reductions.

Figure 2 graphically shows these results and allows their analysis in terms of the non-dominance criterion. As a first point, most of the non-dominance set comprises cases in which balancing is considered before PS. While all these solutions entail a (sometimes slight) decrease in the F_1 score when compared to the

Table 2. Results obtained in terms of the F_1 and reduction rate (in percentage referred to the initial case without PS) figures of merit for each combination of PS and balancing method. Bold results remark the elements that belong to the non-dominated set.

Balancing	Metric	PS method								
		ALL	CNN	FCNN	ENN	EFCNN	DROP3	CHC	$EN_{0.1}$	$FN_{0.1}$
Original	F_1	0.69	0.64	0.67	0.64	0.63	0.58	**0.52**	0.52	0.58
	Size (%)	100.0	30.6	28.2	85.6	7.2	8.0	**0.7**	1.6	3.3
SMT	F_1	0.70	0.64	0.68	0.68	0.67	0.60	**0.62**	0.53	0.59
	Size (%)	150.3	40.4	34.0	132.0	13.3	14.7	**2.1**	2.3	4.9
B1	F_1	**0.70**	0.64	0.68	0.68	**0.67**	0.60	0.59	0.48	0.55
	Size (%)	**150.3**	37.9	32.8	132.2	**12.5**	13.9	2.5	2.8	4.9
B2	F_1	0.69	0.64	0.67	0.67	0.66	0.60	0.58	0.49	0.54
	Size (%)	150.3	40.7	35.9	129.5	12.7	14.9	2.2	3.1	5.0
CNN	F_1	**0.69**	0.62	0.66	0.60	0.60	0.56	**0.47**	0.52	0.54
	Size (%)	**49.2**	27.1	26.4	34.2	4.8	5.3	**0.4**	2.5	2.7
NCL	F_1	0.69	0.64	0.67	0.66	0.65	0.56	**0.59**	0.53	0.59
	Size (%)	78.2	18.6	16.7	71.6	5.9	6.2	**0.8**	1.0	2.5
TL	F_1	0.69	0.63	0.67	0.66	**0.66**	0.58	**0.54**	0.52	0.59
	Size (%)	92.3	25.6	23.2	80.8	**7.1**	7.3	**0.7**	1.4	3.0
CNN-SMT	F_1	0.69	0.63	0.66	0.66	0.64	0.60	0.56	0.49	0.52
	Size (%)	65.8	32.4	30.0	49.1	8.4	9.4	1.0	2.9	3.2
CNN-B1	F_1	**0.69**	0.63	0.67	0.66	0.64	0.57	0.54	0.47	0.53
	Size (%)	**65.9**	31.3	29.3	49.1	8.5	9.4	1.1	3.0	3.2
CNN-B2	F_1	0.69	0.62	0.66	0.65	0.63	0.56	0.55	0.47	0.50
	Size (%)	65.9	32.3	31.1	47.9	8.9	9.3	1.0	3.1	3.3
NCL-SMT	F_1	0.69	0.65	0.67	0.67	0.67	0.61	0.59	0.52	0.58
	Size (%)	109.7	22.6	18.3	101.6	9.6	10.5	1.7	1.2	3.5
NCL-B1	F_1	0.69	0.65	**0.68**	0.68	**0.67**	0.59	0.58	0.49	0.54
	Size (%)	109.5	21.9	**18.4**	101.7	**9.4**	10.3	2.0	1.7	3.4
NCL-B2	F_1	0.69	0.64	0.67	0.67	0.66	0.60	0.58	0.49	0.53
	Size (%)	109.7	23.5	20.1	100.3	9.8	11.7	1.9	1.8	3.5
TL-SMT	F_1	0.69	0.65	0.67	0.68	0.67	0.59	**0.61**	0.52	0.59
	Size (%)	134.9	32.9	27.8	120.6	11.5	11.3	**1.7**	1.9	4.3
TL-B1	F_1	0.69	0.64	**0.68**	0.67	0.66	0.59	0.59	0.48	0.54
	Size (%)	134.9	31.2	**26.7**	120.3	10.9	11.4	2.3	2.3	4.3
TL-B2	F_1	0.69	0.64	0.67	0.67	0.65	0.59	0.58	0.48	0.53
	Size (%)	134.9	33.6	29.5	117.9	11.2	12.6	2.1	2.6	4.4

initial case, the resulting set is remarkably more compact than the original situation. For instance, the NCL-B1 balancing method coupled with FCNN achieves an $F_1 = 0.68$ with less than a fifth of the total number of prototypes.

Regarding PS without balancing, CHC algorithm is the only case among the non-dominated solutions. Thus, according to this criterion, solutions involving PS without a balancing stage may not be considered as optimal, in general.

Finally, the cases that only consider the balancing scheme and avoid the PS stage are also present among the non-dominant solutions. Particularly, the non-dominated solutions by the CNN and CNN-B1 balancing cases achieve the same F_1 scores than the original case with a remarkable set reduction. Also in this regard, it must pointed out the B1 oversampling algorithm that, despite

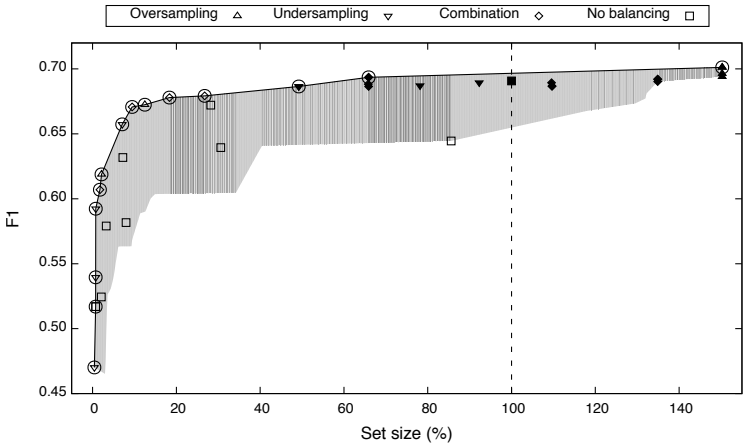

Fig. 2. Graphical representation of the results obtained. Balancing paradigms are represented by the symbols in the legend. The use or not of PS is shown by being these symbols either empty or filled, respectively. Circled symbols remark the elements belonging to the non-dominance set whereas the vertical dashed line refers to the original set size. To avoid graph overload, the grey region depicts the space occupied by all results obtained in this work from the combinations of balancing techniques (oversampling, undersampling, and combination) and PS strategies studied.

achieving the best F_1 score, the set size is remarkably higher than the initial one, being thus an option to discard as our premise is to reduce our initial set.

5 Conclusions and future work

Imbalance in class distributions typically affects the performance in classification schemes as it biases the response of the system towards the majority class. To tackle it, data-level approaches that artificially equilibrate the class distribution have been proposed and studied. As a particular process found in instance-based classification schemes, prototype selection (PS) schemes are typically designed for balanced data distributions, but this is not realistic as real-life data sources do not exhibit such ideal distribution.

In this context, we performed a study comparing the performance of PS schemes on imbalanced collections and the same sets after being balanced with data-driven approaches for Nearest Neighbour classification. Results obtained considering six datasets and a comprehensive collection of PS schemes and balancing techniques suggest that general PS techniques achieve better performances when data is balanced and that some balancing techniques based on undersampling the majority class do not require of a PS stage as by themselves achieve good reduction rates while keeping fairly accurate classification figures.

Future work considers the development of PS strategies at an algorithmic level, that is, biasing their internal figure of merit so that the selection additio-

nally considers the class imbalance present in the data.

Acknowledgements. Work partially supported by the Spanish Ministerio de Economía y Competitividad through Project TIMuL (No. TIN2013-48152-C2-1-R supported by EU FEDER funds), the Spanish Ministerio de Educación, Cultura y Deporte through FPU program (AP2012–0939) and the Vicerrectorado de Investigación, Desarrollo e Innovación de la Universidad de Alicante through FPU program (UAFPU2014–5883).

References

1. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley & Sons, 2001.
2. S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, ser. Intelligent Systems Reference Library. Springer, 2015, vol. 72.
3. S. García, J. Derrac, J. Cano, and F. Herrera, "Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 417–435, Mar. 2012.
4. V. García, J. Sánchez, and R. Mollineda, "An empirical study of the behavior of classifiers on imbalanced and overlapped data sets," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2007, pp. 397–406.
5. V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.
6. V. García, J. S. Salvador, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 13–21, 2012.
7. A. Fernández, S. García, and F. Herrera, "Addressing the classification with imbalanced data: open problems and new challenges on class distribution," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2011, pp. 1–10.
8. J. R. Rico-Juan and J. M. Iñesta, "New rank methods for reducing the size of the training set using the nearest neighbor rule," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 654–660, 2012.
9. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
10. H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International Conference on Intelligent Computing*. Springer, 2005, pp. 878–887.
11. R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "Data mining with imbalanced class distributions: concepts and methods," in *Proceedings of the 4th Indian International Conference on Artificial Intelligence*, India, 2009, pp. 359–376.
12. J. J. Valero-Mas, J. M. Iñesta, and C. Pérez-Sancho, "Onset detection with the user in the learning loop," in *Proceedings of the 7th International Workshop on Music and Machine Learning (MML)*, Barcelona, Spain, 2014.
13. J. Calvo-Zaragoza, J. J. Valero-Mas, and J. R. Rico-Juan, "Improving kNN multi-label classification in prototype selection scenarios using class proposals," *Pattern Recognition*, vol. 48, no. 5, pp. 1608–1622, 2015.