Robustness of learning techniques in handling class noise in imbalanced datasets

D. Anyfantis, M. Karagiannopoulos, S. Kotsiantis and P. Pintelas
Educational Software Development Laboratory
Department of Mathematics, University of Patras, Greece
{dany,mariosk,sotos,pintelas}@math.upatras.gr,
WWW home page: http://www.math.upatras.gr/~esdlab

Abstract. Many real world datasets exhibit skewed class distributions in which almost all instances are allotted to a class and far fewer instances to a smaller, but more interesting class. A classifier induced from an imbalanced dataset has a low error rate for the majority class and an undesirable error rate for the minority class. Many research efforts have been made to deal with class noise but none of them was designed for imbalanced datasets. This paper provides a study on the various methodologies that have tried to handle the imbalanced datasets and examines their robustness in class noise.

1 Introduction

In many applications classifiers are faced with imbalanced data sets, which can cause the classifier to be biased towards one class. This bias is the result of one class being seriously under represented in the training data compared to the other classes. It can be qualified to the way in which classifiers are designed. Inductive classifiers are normally designed to minimize errors over the training examples. Learning algorithms, because of the fact that the cost of performing well on the over-represented class outweighs the cost of poor accuracy on the smaller class, can ignore classes containing few examples [16]. For a number of application domains, a massive disproportion in the number of cases belonging to each class is common. For example, in detection of fraud in telephone calls [9] and credit card transactions the number of legitimate transactions is much higher than the number of fraudulent transactions. Moreover, in direct marketing [19], it is frequent to have a small response rate (about 1%) for most marketing campaigns. Other examples of domains with intrinsic imbalance can be found in the literature such as rare medical diagnoses [22] and oil spills in satellite images [18].

The machine learning community has mostly addressed the issue of class imbalance in two ways. One is to give distinct costs to training instances [8]. The other is to re-sample the original dataset, either by oversampling the minority class and/or under-sampling the majority class [17], [12]. Thus, existing research endeavors have made significant progress in exploring techniques for handling imbalanced datasets with assumptions that the input data are noise-free or noise in the data sets is not significant. However, real-world data are rarely perfect and can often suffer from corruptions that may impact interpretations of the data, models created from the data, and decisions made on the data.

Many research efforts have been made to deal with class noise [14], [23], [4], [11], [24], and have suggested that in many situations, eliminating instances that contain class noise will improve the classification accuracy. Although, many research efforts have focused on noise identification and data cleansing, none of them was originally designed for imbalanced datasets. In this study, the effectiveness of techniques for handling imbalanced datasets in class noise is evaluated over 7 imbalanced datasets using the C4.5 [20], Naive Bayes [6] and 5NN [1] as classifiers and the geometric mean of accuracies as performance measure [17].

Section 2 reviews the attempts for handling imbalanced datasets, while section 3 presents experimental results of the techniques for handling imbalanced datasets in class noise. Finally, section 4 discusses the results and suggests directions.

2 Review of existing techniques for handling imbalanced datasets

A simple method that can be used to imbalanced data sets is to reweigh training examples according to the total cost assigned to each class [6]. The idea is to change the class distributions in the training set towards the most costly class. The effect of imbalance in a dataset is also discussed in [12]. Japkowicz mainly evaluated two strategies: under-sampling and resampling. She noted that both the sampling approaches were helpful. In [17] the researchers selectively under-sampled the majority class while keeping the original population of the minority class with satisfied results. Batista et al. [2] used a more sophisticated under-sampling technique in order to reduce the amount of potentially useful data. Another approach is that of [19]. They combined over-sampling of the minority class with under-sampling of the majority class. However, the over-sampling and under-sampling combination did not provide significant improvement. In [5] they recommend an over-sampling approach in which the minority class is over-sampled by creating "synthetic" instances rather than by over-sampling with replacement with better results.

Changing the class distribution is not the only technique to improve classifier performance when learning from imbalanced data sets. A different approach to incorporating costs in decision-making is to define fixed and unequal misclassification costs between classes. Cost model takes the form of a cost matrix, where the cost of classifying a sample from a true class j to class i corresponds to the matrix entry λ_{ij}. This matrix is usually expressed in terms of average misclassification costs for the problem. The diagonal elements are usually set to zero, meaning correct classification has no cost. We define conditional risk for making a decision α_i as: $R(a_i \mid x) = \sum_i \lambda_{ij} P(v_j \mid x)$. The equation states that the risk of choosing class i is defined by fixed misclassification costs and the uncertainty of our knowledge about the true class of x expressed by the posterior probabilities. The goal in cost-sensitive classification is to minimize the cost of misclassification, which can be realized by choosing the class (v_j) with the minimum conditional risk.

An alternative to balancing the classes is to develop a learning algorithm that is intrinsically insensitive to class distribution in the training set. An example of this kind of algorithm is the SHRINK algorithm [17] that finds only rules that best summarize the positive instances (of the small class), but makes use of the information from the negative instances. MetaCost [6] is another method for making a classifier cost-sensitive. The procedure begins to learn an internal cost-sensitive model by applying a cost-sensitive procedure, which employs a base learning algorithm. Then, MetaCost procedure estimates class probabilities using bagging and then re-labels the training instances with their minimum expected cost classes, and finally relearns a model using the modified training set.

3 Experiments

For the aim of our study the most well-known decision tree algorithm - C4.5 [20] – was used. One of the latest researches that compare decision trees and other learning algorithms is made in [21] and shows that the mean error rates of most algorithms are similar and that their differences are statistically insignificant. But, unlike error rates, there are huge differences between the training times of the algorithms. C4.5 has one of the best combinations of error rate and speed. Decision tree classifiers, regularly, employ post-pruning techniques that evaluate the performance of decision trees as they are pruned using a validation set. Any node can be removed and assigned the most common class of the training examples that are sorted to the node in question. As a result, if a class is rare, decision tree algorithms often prune the tree down to a single node that classifies all instances as members of the common class leading to poor accuracy on the examples of minority class. The extreme skewness in class distribution is problematic for Naïve Bayes [7]. The prior probability of the majority class overshadows the differences in the attribute conditional probability terms. Instance-based learning algorithms belong to the category of lazy-learning algorithms, as they delay the induction until classification is performed. One of the most straightforward instance-based learning algorithms is the nearest neighbour algorithm [1]. In our study, we made use of the commonly used 5-NN algorithm. In imbalanced data sets as the number of the instances of the majority class grows, so does the likelihood that the nearest neighbour of any instance will belong to the majority class. This leads to the problem that many instances of the minority class will be misclassified.

In Table 1, there is a brief description of the data sets that we used for our experiments. Except for the “eap” data set, all were drawn from the UC Irvine Repository [3]. Eap data is from Hellenic Open University and was used in order to determine whether a student is about to drop-out or not [15].

Table 1. Description of the data sets

Data sets	Instances	Categorical Features	Numerical Features	Instances of minority class	Classes
breast-cancer	286	9	0	85	2
credit-g	1000	13	7	300	2
Diabetes	768	0	8	268	2
Haberman	306	0	3	81	2
Hepatitis	155	13	6	32	2
Ionosphere	351	34	0	126	2
Eap	344	11	0	122	2

For most of the datasets we used, they don't actually contain noise, so we use manual mechanisms to add class noise. For class noise, we adopt a pairwise scheme [24]: given a pair of classes (X, Y) and a noise level x, an instance with its label X has an x*100% chance to be corrupted and mislabeled as Y, so does an instance of class Y. Meanwhile, we only report the value x of class noise (which is not the actual class noise level in the dataset) in all tables below.

When comparing the performance of different classifiers in imbalanced data sets, accuracy as a measure is not enough. A classifier's performance of two class problems can be separately calculated for its performance over the positive instances (denoted as α^+) and over the negative instances (denoted as α^-). The true positive rate (α^+) or sensitivity is the fraction of positive instances predicted correctly by the model. Similarly, the true negative rate (α^-) or specificity is the fraction of negative instances predicted correctly by the classifier. In [17] they propose the geometric mean of the accuracies: $g = \sqrt{a^+ \times a^-}$ for imbalanced data sets. The basic idea behind this measure is to maximize the accuracy on both classes. Classification ability of the learning methods in our experiments was measured with geometric mean of the accuracies. For the examined models, the relationship between false negative and false positive costs was chosen to be the inverse of the assumed prior to compensate for the imbalanced priors.

In Table 2, one can see the comparisons with class noise of the attempts that have tried to obtain the best performance of a given imbalance data set using Naive Bayes (NB) as base classifier. Three well-known algorithms were used for the comparison: Reweighing and Cost Sensitive method [6] and Metacost algorithm [8]. We also present the accuracy of the simple Bayes algorithm as borderline. It must be mentioned that we used the free available source code for these methods [22] for our experiments. In Table 2 and Table 3 except for geometric mean we also present the true-positive rate, and true-negative rate. It must be mentioned that positive class is the majority class for our experiments. In the last row of Table 2, the average value of the geometric means is also calculated in all data sets. It must be mentioned that for Naïve Bayes classifier, modifying the decision boundary (Cost Sensitive method) is equivalent to reweighing training instances so as the relationship between false negative and false positive costs to be the inverse of the imbalanced priors. All the tested techniques give better results than the single Naive Bayes in class noise. The Reweighing and Cost Sensitive method gave better results with little class noise; however Metacost can handle better more class noise.

Table 2. Accuracy on majority class (α+), accuracy on minority class (α-) and geometric mean (g) with NB as base classifier

Datasets		ReWNB or CostNB Without Noise	ReWNB or CostNB With 10% Noise	ReWNB or CostNB With 20% Noise	Meta-cost NB Without Noise	Meta-cost NB With 10% Noise	Meta-cost NB With 20% Noise	NB Without Noise	NB With 10% Noise	NB With 20% Noise
breast-cancer	g	0.66	0.65	0.54	0.65	0.67	0.63	0.6	0.62	0.61
	α+	0.74	0.65	0.43	0.79	0.72	0.59	0.85	0.84	0.8
	α-	0.58	0.66	0.69	0.54	0.62	0.67	0.43	0.46	0.46
credit-g	g	0.72	0.72	0.7	0.66	0.69	0.7	0.65	0.68	0.68
	α+	0.75	0.68	0.59	0.77	0.73	0.65	0.86	0.85	0.84
	α-	0.69	0.77	0.82	0.57	0.65	0.75	0.49	0.54	0.55
diabetes	g	0.73	0.74	0.71	0.70	0.71	0.7	0.71	0.72	0.72
	α+	0.78	0.76	0.68	0.75	0.72	0.66	0.84	0.84	0.85
	α-	0.68	0.72	0.75	0.66	0.71	0.74	0.6	0.62	0.61
Haberman	g	0.56	0.58	0.46	0.57	0.59	0.5	0.44	0.45	0.39
	α+	0.89	0.83	0.26	0.87	0.84	0.3	0.94	0.94	0.95
	α-	0.35	0.4	0.83	0.38	0.42	0.82	0.21	0.22	0.16
Heapatitis	g	0.8	0.79	0.78	0.81	0.8	0.8	0.78	0.79	0.8
	α+	0.83	0.81	0.67	0.79	0.76	0.73	0.87	0.83	0.82
	α-	0.78	0.78	0.91	0.84	0.84	0.88	0.7	0.75	0.78
Ionosphere	g	0.82	0.83	0.81	0.77	0.77	0.76	0.83	0.82	0.81
	α+	0.78	0.8	0.81	0.68	0.68	0.69	0.8	0.83	0.84
	α-	0.87	0.86	0.81	0.88	0.88	0.84	0.86	0.81	0.79
eap	g	0.85	0.82	0.78	0.85	0.84	0.8	0.84	0.82	0.82
	α+	0.87	0.79	0.68	0.88	0.85	0.72	0.9	0.88	0.87
	α-	0.83	0.85	0.89	0.83	0.83	0.88	0.78	0.76	0.78
Average	g	0.73	0.73	0.68	0.72	0.72	0.7	0.7	0.7	0.69

In Table 3, one can see the comparisons with class noise of the attempts that have tried to obtain the best performance of a given imbalance data set using C4.5 as base classifier. The same three well-known techniques for handling imbalanced data sets were also used for this comparison. In general, all the tested techniques give better results than the single C4.5 in class noise. The Reweighing method gave better results with little class noise, however Metacost can handle better more class noise.

In Table 4, one can see the comparisons of the proposed technique with other attempts that have tried to obtain the best performance of a given imbalance data sets using 5NN as base classifier. The same three well-known techniques for handling imbalanced data sets were also used for this comparison. It must be mentioned that for 5NN classifier, modifying the decision boundary (Cost Sensitive method) is equivalent to reweighing training instances so as the relationship between false negative and false positive costs to be the inverse of the imbalanced priors. In general, all the tested techniques give similar better results than the single 5NN in class noise and there was no difference between them.

Table 3. Accuracy on majority class (α+), accuracy on minority class (α-) and geometric mean (g) with C4.5 as base classifier

Data-sets		ReW-C4.5 Without Noise	ReW-C4.5 With 10% Noise	ReW-C4.5 With 20% Noise	Cost-C4.5 Without Noise	Cost-C4.5 With 10% Noise	Cost-C4.5 With 20% Noise	Meta-cost C4.5 Without Noise	Meta-cost C4.5 With 10% Noise	Meta-cost C4.5 With 20% Noise	C4.5 Without Noise	C4.5 With 10% Noise	C4.5 With 20% Noise
breast-cancer	g	0.57	0.58	0.47	0.5	0.52	0.44	0.55	0.61	0.53	0.5	0.46	0.46
	α+	0.72	0.56	0.29	0.85	0.84	0.73	0.84	0.78	0.39	0.95	0.9	0.85
	α-	0.45	0.6	0.77	0.3	0.32	0.27	0.36	0.48	0.72	0.26	0.24	0.25
credit-g	g	0.66	0.63	0.6	0.61	0.63	0.64	0.64	0.65	0.66	0.58	0.6	0.61
	α+	0.67	0.57	0.47	0.82	0.76	0.68	0.76	0.71	0.65	0.85	0.83	0.79
	α-	0.65	0.69	0.76	0.46	0.52	0.6	0.54	0.6	0.68	0.4	0.44	0.47
Diabetes	g	0.72	0.69	0.63	0.72	0.66	0.65	0.73	0.7	0.65	0.7	0.67	0.65
	α+	0.72	0.66	0.47	0.78	0.79	0.63	0.78	0.73	0.59	0.82	0.8	0.87
	α-	0.73	0.72	0.84	0.67	0.56	0.68	0.67	0.68	0.71	0.6	0.56	0.49
Haberman	g	0.63	0.59	0.42	0.58	0.58	0.4	0.62	0.59	0.38	0.52	0.56	0.43
	α+	0.68	0.56	0.19	0.66	0.81	0.19	0.76	0.61	0.16	0.85	0.83	0.9
	α-	0.58	0.62	0.95	0.51	0.41	0.83	0.52	0.58	0.91	0.32	0.38	0.21
Heapatitis	g	0.73	0.72	0.59	0.64	0.54	0.51	0.68	0.67	0.71	0.58	0.52	0.51
	α+	0.62	0.69	0.55	0.86	0.77	0.47	0.83	0.76	0.63	0.9	0.87	0.84
	α-	0.85	0.75	0.63	0.48	0.38	0.56	0.56	0.59	0.81	0.37	0.31	0.31
Ionosphere	g	0.89	0.83	0.8	0.88	0.82	0.77	0.9	0.85	0.78	0.88	0.82	0.77
	α+	0.94	0.88	0.91	0.94	0.94	0.92	0.98	0.92	0.86	0.94	0.94	0.92
	α-	0.85	0.79	0.7	0.82	0.71	0.64	0.82	0.78	0.71	0.82	0.71	0.64
eap	g	0.81	0.78	0.71	0.83	0.8	0.79	0.82	0.79	0.76	0.83	0.83	0.86
	α+	0.86	0.76	0.57	0.94	0.84	0.75	0.89	0.78	0.69	0.94	0.94	0.92
	α-	0.77	0.8	0.89	0.74	0.76	0.84	0.76	0.8	0.84	0.74	0.74	0.8
Average	g	0.72	0.69	0.6	0.68	0.65	0.6	0.71	0.69	0.64	0.66	0.64	0.61

As a general conclusion, the Reweighing method is a more appropriate technique in the presence of little class noise, however Metacost can handle better more class noise.

Conclusion

Existing research endeavors have made significant progress in exploring techniques for handling imbalanced datasets with assumptions that the input data are noise-free or noise in the data sets is not significant. However, real-world data are rarely perfect and can often suffer from corruptions that may impact interpretations of the data, models created from the data, and decisions made on the data. In this study, the effectiveness of techniques for handling imbalanced datasets in class noise is evaluated over 7 imbalanced datasets. Metacost seems to be more robust as the class noise increased. In a following study, we will examine multi-class imbalanced datasets and will propose a more robust technique in the class noise.

Table 4. Accuracy on majority class (α+), accuracy on minority class (α-) and geometric mean (g) with 5NN as base classifier

Data sets		ReW5NN Or Cost5NN Without Noise	ReW5NN Or Cost5NN With 10% Noise	ReW5NN Or Cost5NN With 20% Noise	Metacost 5NN Without Noise	Metacost 5NN With 10% Noise	Metacost 5NN With 20% Noise	5NN Without Noise	5NN With 10% Noise	5NN With 20% Noise
breast-cancer	g	0.62	0.6	0.59	0.51	0.59	0.58	0.45	0.44	0.47
	α+	0.73	0.6	0.47	0.86	0.67	0.51	0.96	0.95	0.92
	α-	0.52	0.61	0.73	0.3	0.52	0.67	0.21	0.2	0.24
credit-g	g	0.66	0.63	0.58	0.63	0.66	0.59	0.57	0.58	0.59
	α+	0.69	0.58	0.44	0.73	0.64	0.45	0.89	0.85	0.76
	α-	0.63	0.69	0.77	0.55	0.67	0.78	0.37	0.39	0.46
diabetes	g	0.71	0.67	0.62	0.71	0.69	0.64	0.68	0.65	0.59
	α+	0.69	0.61	0.51	0.75	0.69	0.58	0.83	0.84	0.78
	α-	0.74	0.74	0.75	0.68	0.7	0.71	0.56	0.5	0.45
haberman	g	0.57	0.54	0.5	0.59	0.53	0.49	0.39	0.41	0.44
	α+	0.68	0.55	0.41	0.66	0.53	0.62	0.9	0.84	0.76
	α-	0.47	0.53	0.61	0.52	0.53	0.39	0.17	0.2	0.25
hepatitis	g	0.69	0.68	0.6	0.8	0.7	0.6	0.66	0.6	0.64
	α+	0.79	0.73	0.55	0.84	0.62	0.41	0.94	0.93	0.83
	α-	0.6	0.63	0.66	0.76	0.78	0.88	0.46	0.41	0.5
ionosphere	g	0.83	0.83	0.76	0.79	0.78	0.75	0.78	0.76	0.73
	α+	0.97	0.88	0.7	0.98	0.94	0.85	0.98	0.95	0.9
	α-	0.71	0.78	0.83	0.63	0.64	0.67	0.62	0.61	0.6
eap	g	0.8	0.75	0.62	0.77	0.75	0.59	0.78	0.76	0.73
	α+	0.84	0.64	0.44	0.87	0.7	0.4	0.9	0.89	0.88
	α-	0.76	0.87	0.88	0.69	0.8	0.88	0.68	0.65	0.61
Average	g	0.7	0.67	0.61	0.69	0.67	0.61	0.62	0.6	0.6

References

1. Aha, D. (1997). Lazy Learning. Dordrecht: Kluwer Academic Publishers.
2. Batista G., Carvalho A., Monard M. C. (2000), Applying One-sided Selection to Unbalanced Datasets. In O. Cairo, L. E. Sucar, and F. J. Cantu, editors, Proceedings of the Mexican International Conference on Artificial Intelligence – MICAI 2000, pages 315–325. Springer-Verlag.
3. Blake, C., Keogh, E. & Merz, C.J. (1998). UCI Repository of machine learning databases [http:// www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California.
4. Brodley, C. E. & Friedl, M. A. (1999). Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research 11: 131–167.

5. Chawla N., Bowyer K., Hall L., Kegelmeyer W. (2002), SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16, 321 - 357.
6. Domingos P. (1998), How to get a free lunch: A simple cost model for machine learning applications. Proc. AAAI-98/ICML98, Workshop on the Methodology of Applying Machine Learning, pp1-7.
7. Domingos P. & Pazzani M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, 103-130.
8. Domingos, P. (1999). MetaCost: A General Method for Making Classifiers Cost-Sensitive. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, 155-164. ACM Press.
9. Fawcett T. and Provost F. (1997), Adaptive Fraud Detection. Data Mining and Knowledge Discovery, 1(3):291–316.
10. Friedman J. H. (1997), On bias, variance, 0/1-loss and curse-of-dimensionality. Data Mining and Knowledge Discovery, 1: 55-77.
11. Gamberger, D., Lavrac, N. & Dzeroski, S. (2000). Noise Detection and Elimination in Data Preprocessing: experiments in medical domains. Applied Artificial Intelligence 14, 205-223.
12. Japkowicz N. (2000), The class imbalance problem: Significance and strategies. In Proceedings of the International Conference on Artificial Intelligence, Las Vegas.
13. Japkowicz N. and Stephen, S. (2002), The Class Imbalance Problem: A Systematic Study Intelligent Data Analysis, Volume 6, Number 5.
14. John, G. H. (1995). Robust Decision Trees: Removing Outliers from Databases. Proc. of the First International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp. 174–179.
15. Kotsiantis, S., Pierrakeas, C., Pintelas, P., Preventing student dropout in distance learning systems using machine learning techniques, Lecture Notes in Artificial Intelligence, KES 2003, Springer-Verlag Vol 2774, pp 267-274, 2003.
16. Kotsiantis S., Kanellopoulos, D. Pintelas, P. (2006), Handling imbalanced datasets: A review, GESTS International Transactions on Computer Science and Engineering, Vol.30 (1), pp. 25-36.
17. Kubat, M. and Matwin, S. (1997), 'Addressing the Curse of Imbalanced Data Sets: One Sided Sampling', in the Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179-186.
18. Kubat, M., Holte, R. and Matwin, S. (1998), 'Machine Learning for the Detection of Oil Spills in Radar Images', Machine Learning, 30:195-215.
19. Ling, C., & Li, C. (1998). Data Mining for Direct Marketing Problems and Solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) New York, NY. AAAI Press.
20. Quinlan J.R. (1993), C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.
21. Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih (2000), A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning, 40, 203–228, 2000, Kluwer Academic Publishers.
22. Witten Ian H. and Frank Eibe (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
23. Zhao, Q. & Nishida, T. (1995). Using Qualitative Hypotheses to Identify Inaccurate Data. Journal of Artificial Intelligence Research 3, pp.119–145.
24. Zhu, X., Wu, X. & Yang, Y. (2004). Error Detection and Impact-sensitive Instance Ranking in Noisy Datasets. In Proceedings of 19th National conference on Artificial Intelligence (AAAI-2004), San Jose, CA.