Editorial: Special Issue on Learning from Imbalanced Data Sets

Nitesh V. Chawla
Retail Risk Management,CIBC
21 Melinda Street
Toronto, ON M5L 1A2, Canada
chawla@morden.csee.usf.edu

Nathalie Japkowicz
School of Information
Technology and Engineering
University of Ottawa
ON K1N 6N5, Canada
nat@site.uottawa.ca

Aleksander Kołcz
AOL, Inc.
44900 Prentice Drive
Dulles, VA 20166, USA
a.kolcz@ieee.org

1. INTRODUCTION

The class imbalance problem is one of the (relatively) new problems that emerged when machine learning matured from an embryonic science to an applied technology, amply used in the worlds of business, industry and scientific research.

Although practitioners might already have known about this problem early, it made its appearance in the machine learning/data mining research circles about a decade ago. Its importance grew as more and more researchers realized that their data sets were imbalanced and that this imbalance caused suboptimal classification performance. This increase in interest gave rise to two workshops held in 2000 [1] and 2003 [3] at the AAAI and ICML conferences, respectively.

These workshops and the ensuing e-mail discussions and information seeking requests that followed them allowed us to note two points of importance:

1. The class imbalance problem is pervasive and ubiquitous, causing trouble to a large segment of the data mining community.

2. Despite the fact that two workshops have already been held on the topic, a large number of practitioners plagued by the problem are still working in isolation, not knowing that a large part of the research community is actively looking into ways to alleviate the problem.

The purpose of this special issue is to communicate and present some of the latest research carried out in this area while reviewing other important recent developments in the field. In this Editorial, we begin by reviewing the class imbalance as well as an array of general solutions that were previously proposed to deal with it. We then discuss the progression of ideas starting at the 2000 workshop to today. In order to give a comprehensive picture of the state of the art in the field, we give a short overview of the papers that were presented at the 2003 workshop as well as a short description of the papers contained in this volume. The excellent overview paper by Gary Weiss [55] published in this volume will complete this short picture.

2. THE CLASS IMBALANCE PROBLEM

The class imbalance problem typically occurs when, in a classification problem, there are many more instances of some classes than others. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. In practical applications, the ratio of the small to the large classes can be drastic such as 1 to 100, 1 to 1,000, or 1 to 10,000 (and sometimes even more). (See, for example, [41], [57]). As mentioned earlier this problem is prevalent in many applications, including: fraud/intrusion detection, risk management, text classification, and medical diagnosis/monitoring, but there are many others. It is worth noting that in certain domains (like those just mentioned) the class imbalance is intrinsic to the problem. For example, within a given setting, there are typically very few cases of fraud as compared to the large number of honest use of the offered facilities. However, class imbalances sometimes occur in domains that do not have an intrinsic imbalance. This will happen when the data collection process is limited (e.g., due to economic or privacy reasons), thus creating "artificial" imbalances. Conversely, in certain cases, the data abounds and it is for the scientist to decide which examples to select and in what quantity [56]. In addition, there can also be an imbalance in costs of making different errors, which could vary per case [3].

A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of re-sampling such as random oversampling with replacement, random undersampling, directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed undersampling (where, again, the choice of examples to eliminate is informed), oversampling with informed generation of new samples, and combinations of the above techniques. At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. Many of these solutions are discussed in the papers presented in the workshops [1][3] or are referred to in the active bibliography on the topic[1].

2.1 Recent Trends: 2000–2004

It is interesting to review briefly the types of problems that

[1] http://www.site.uottawa.ca/˜nat/Research/class_imbalance_bibli.html

have been considered by the researchers on the *class imbalance* problem in the past few years.

2.1.1 The AAAI (2000) Workshop

At the first workshop[2], in July 2000, the two issues that received the greatest amount of attention were:

1. How to evaluate learning algorithms in the case of class imbalances?
2. The relationship between class imbalances and cost-sensitive learning.

With regard to the first issue, it was emphasized that the use of common evaluation measures such as accuracy can yield misleading conclusions. It was suggested that more accurate measures such as ROC curves and Cost Curves be used. Some of the major papers on this topic that came out at the time of or since the first workshop are: [44][18][13][20]. In fact, a workshop on ROC Analysis in AI is being held in August 2004 [4]. In addition, an evaluation measure was proposed for the case where only data from one class is available.

With regard to the second issue, a close connection was recognized between re-sampling approaches and cost-sensitive approaches [53][15][12][52]. Cost-sensitive learning or measures assume that a cost-matrix is known for different types of errors or even examples, which can be used at classification time [53][15]. However, we often do not know the cost matrix. Furthermore, cost-sensitive learning does not modify the class distribution of the data the way re-sampling does. Finally, cost-sensitive learning is not encumbered by large sets of duplicated examples, which, as discussed in [34] can also be a drawback. Practically, it is often reported that cost-sensitive learning outperforms random re-sampling (e.g., [29][12]).

Other issues discussed at that workshop were the facts that:

- A distinction should be drawn between the small sample and the imbalance problem;
- Although smart sampling can, sometimes, help, it is not always possible;
- One-class learning can be useful in class imbalanced problems, where either the majority or the minority class (or both) can be learned separately.
- Creating a classifier that performs well across a range of cost/priors is a desirable goal.

2.1.2 The Inter-Years: 2000–2003

The three years separating the two workshops saw a great amount of activity. The impact of the discussions held at the 2000 workshop was clearly seen on the first issue, in the fact that most of the research conducted in the class imbalance area after that first workshop made use of ROC curves for evaluating the results. In addition, there was a bias towards various over and under-sampling techniques. Decision trees remained a popular classifier for research.

With regard to the relationship between cost-sensitive learning and re-sampling, the impact was not as direct or clear, but it remained present in the form of two emerging ideas:

- Clever re-sampling and combination methods can do quite more than cost-sensitive learning as they can provide new information or eliminate redundant information for the learning algorithm, as shown by [10][11][35][21][6].
- Class Imbalances are not the only problem to contend with: the distribution of the data within each class is also relevant (between-class versus within-class imbalance) [25][58]. This issue is closely related to the 2000 workshop's issue of creating classifiers that performs well across a range of costs/options.

Regarding the idea of clever re-sampling and combination methods, [10] showed that creating synthetic examples of the minority class that spread the decision regions, thus making them larger and less specific while undersampling the majority class to various degrees gave very effective results.

Regarding the second idea, the problem was considered in parallel in both the context of cost-sensitive learning and that of re-sampling. We would like to point the reader to the ICML workshop on cost-sensitive learning [2] and an online bibliography on cost-sensitive learning[3]. Within cost-sensitive learning, various ways of adjusting the probabilistic estimate at the tree leaf (when working with decision trees) were explored so as to give a more direct and more flexible approach to the treatment of different parts of the decision space [58]. A similar idea was used in the context of re-sampling, with recourse to unsupervised learning [25]. This work also gave rise to the question of how related the class imbalance problem is to the problem of small disjuncts, previously discussed in the literature. This last issue was also linked to the small sample versus imbalance problem discussed at the last workshop.

The inter-years also saw some research on one-class learning as well as on a new twist of one-class learning that can be thought of as extreme semi-supervised learning. Such techniques can be used in the context where unlabeled data are readily available, while labeled data for at least one of the classes are missing. In such cases, iterative techniques such as Expectation Minimization (EM), have been used to assign/estimate the missing-class labels [33][36]. This essentially manufactures the missing portion of the training data, which can then be used in the standard induction process. The important differentiator from other semi-supervised problems (e.g., [40]) is that there are no labeled seed data to initialize the estimation model for the missing class.

2.1.3 The ICML (2003) Workshop

First, a hint at the fact that research on the class imbalance problem is starting to mature is the fact that a big proportion of the 2003 workshop was occupied by the comparison of previously proposed schemes for dealing with the class imbalanced problem [9][14][37]. These papers looked at random oversampling, oversampling with artificially generated samples, random undersampling, and directed undersampling. In conjunction with sampling, these papers also considered probabilistic estimates, pruning, threshold adjusting and cost-matrix adjusting. In addition, the paper by [60] proposed several new directed undersampling

[2] These remarks have been adapted from [28].

[3] http://purl.org/peter.turney/bibliographies/cost-sensitive.html

schemes which were compared to each other and to random undersampling. [47] consider two different methods for balancing the class distribution in the data set, and then extend that scenario to only one-class learning. They show that one-class SVM learning can be beneficial for certain domains (an extended version of their paper appears in this Volume). Though all these papers shed some light on the way various methods compare, there is no single final word on the question. In other words, a number of techniques were shown to be effective if applied in a certain context, where the breadth of the context may vary. It is worth noting that there was much more discussion at ICML 2003 on sampling methods as compared to the AAAI workshop. That brings us to question: *Is sampling becoming a de facto standard for countering imbalance?*

In addition to sampling, an important question to consider, particularly when applying machine learning to a real-world problem, is the cost associated in acquiring the data. Given these costs, a "budgeted" sampling approach is required. In the opening talk, Foster Provost discussed different costs in procuring the data, and learning from it [43]. His presentation suggested that when using ROC as the performance criteria, a balanced distribution is mostly the preferred choice. He also addressed the question: *Given a budget for data procurement, what class distribution should be used?* He proposed a novel budget-sensitive progressive sampling approach, which is not worse than choosing a balanced or natural distribution.

Although the issue of how to evaluate classifiers in cases of class imbalances seemed settled by the adoption of ROC curves and, to a lesser extent, Cost Curves, the question resurfaced at the 2003 workshop when Charles Elkan [16] pointed out that ROC curves are unable to deal with within-class imbalances and different within-class misclassification costs. However, if there are well defined subcategories of one class, the evaluation set can be resampled in proportion to true sub-class proportions and their costs. Elkan suggested that the issue of how to evaluate classifiers in cases of class-imbalances should be revisited in accordance with this question. Another issue with regard to the evaluation of classifiers is concerned with the distribution that should be used in the testing set. It is often assumed that the target distribution should be used, but the issue with this solution is the fact that the target distribution is usually unknown.

One interesting criticism raised in conjunction with this series of papers, however, is worth noting here: too much reliance of the class-imbalance research community on C4.5 [45]. It was argued, in particular, that C4.5 is not the best classifier for dealing with class imbalances and that the community should focus on it less. Two ensuing questions related to that issue were whether some classifiers are insensitive to class imbalances [24]; and whether classification really is the task to focus on, or whether it would be best to perform ranking (through probability estimation) [16].

In another invited talk, Naoki Abe presented various strategies for selective sampling based on query learning [5]. This aids in selecting data near the decision boundary. He proposed cost-proportionate sampling methodologies. He also discussed a cost-sensitive ensemble learning method called *Costing* [59], which achieved significant improvements over random resampling methods. [32] also discussed using selective sampling as a part of active sampling before learning a one-class classifier. In that work, it was shown that data selection driven by the uncertainty of a classifier and/or distance from the target class provide viable (although somewhat classifier dependent) approaches.

A number of papers discussed interaction between the class imbalance and other issues such as the small disjunct [27] and the rare cases [23] problems, data duplication [34], and overlapping classes [54]. It was found that in certain cases, addressing the small disjunct problem with no regard for the class imbalance problem was sufficient to increase performance[4]. Though in other cases, [41] found that handling the small disjuncts was not sufficient. The method for handling rare case disjuncts was found to be similar to the m-estimation Laplace smoothing, but it requires less tuning. It was also found that data duplication is generally harmful, although for classifiers such as Naive Bayes and Perceptrons with Margins, high degrees of duplication are necessary to harm classification [34]. It was argued that the reason why class imbalances and overlapping classes are related is that misclassification often occurs near class boundaries where overlap usually occurs as well.

Two of the workshop papers also presented novel approaches such as [61] who proposed a feature selection approach specifically tuned to the class imbalance problem (an expanded version of their work appears in this volume and will be discussed in more detail below) and [57] who propose to modify the Kernel function or matrix of an SVM by adapting it locally based on the data distribution.

3. SUMMARY OF THIS VOLUME'S CONTRIBUTIONS (2004)

In this section, we summarize the most recent developments in the area of class imbalances by describing briefly the contributions to this volume along with the context in which they fall. Gary Weiss [55] presents an overview of the field of learning from imbalanced data. He pays particular attention to differences and similarities between the problems of rare classes and rare cases. He then discusses some of the common issues and their range of solutions in mining imbalanced datasets. The rest of the contributions are made to three subareas of the class imbalance problem: *Sampling*, *One Class Learning*, and *Feature Selection*.

3.1 Sampling

The compelling question, given the different class distributions, is: *What is the correct distribution for a learning algorithm?* It has been observed that naturally occurring distribution is not always the optimal distribution [56]. In addition, the imbalance in the data can be more characteristic of the "sparseness" in feature space than the class imbalance [10].

Random undersampling can potentially remove certain important examples, and random oversampling can lead to overfitting. In addition, oversampling can introduce an additional computational task if the data set is already fairly large but imbalanced. How much to oversample or undersample is usually empirically detected. There has been a progression in sampling methods to focus on particular majority or minority class samples. Another interesting paradigm of research utilizing (adaptive or random or fo-

[4] An expanded version of [27] appears in this volume and will be discussed further below.

cused) sampling is evolving under the multiple classifier systems or ensembles domain [8][12][11][46][51][17][31][59].

Various papers in this special issue focus on utilizing sampling methods directly or as a part of ensemble learning. Batista et. al [6] present a comparison (and combination) of various sampling strategies. They note that combining focused over and undersampling, such as SMOTE+Tomek or SMOTE+ENN is applicable when the data sets are highly imbalanced or there are very few instances of the minority class. Guo and Viktor [21] propose another technique that modifies the boosting procedure — DataBoost. As compared to SMOTEBoost, which only focuses on the hard minority class cases, this technique employs a synthetic data generation process for both minority and majority class cases. Phua et. al [42] combine bagging and stacking to identify the best mix of classifiers. In their insurance fraud detection domain, they note that stacking-bagging achieves the best cost-savings. Jo and Japkowicz [30] shed some new and different light to the problem of class imbalance in a data set. They suggest that small disjuncts (due to class imbalance) in C4.5 decision trees and backpropagation neural networks are responsible for performance degradation. The (often) negative impact of class imbalance is compounded by the problem of small disjuncts, particularly in small and complex data sets. They propose use of cluster-based oversampling to counter the effect of class imbalance and small disjuncts.

3.2 One-class Learning

When negative examples greatly outnumber the positive ones, certain discriminative learners have a tendency to overfit. A recognition-based approach provides an alternative to discrimination where the model is created based on the examples of the target class alone. Here, one attempts to measure (either implicitly or explicitly) the amount of similarity between a query object and the target class, where classification is accomplished by imposing a threshold on the similarity value [26].

Mainly, two classes of learners were previously studied in the context of the recognition-based one-class approach—SVMs [50][49] and autoencoders [26][38]—and were found to be competitive [38].

An interesting aspect of one-class (recognition-based) learning is that, under certain conditions such as multi-modality of the domain space, one class approaches to solving the classification problem may in fact be superior to discriminative (two-class) approaches (such as decision trees or Neural Networks) [26]. This is supported in the current volume by [48], who demonstrate the optimality of one-class SVMs over two-class ones in certain important imbalanced-data domains, including genomic data. In particular, [48] shows that one class learning is particularly useful when used on extremely unbalanced data sets composed of a high dimensional noisy feature space. They argue that the one-class approach is related to aggressive feature selection methods, but is more practical since feature selection can often be too expensive to apply.

3.3 Feature Selection

Feature selection is an important and relevant step for mining various data sets [22]. Learning from high dimensional spaces can be very expensive and usually not very accurate. It is particularly relevant to various real-world problems such as bioinformatics, image processing, text classification, Web categorization, etc. High dimensional real-world data sets are often accompanied by another problem: high skew in the class distribution, with the class of interest being relatively rare. This makes it particularly important to select features that lead to a higher separability between the two classes. It is important to select features that can capture the high skew in the class distribution. The majority of work in feature selection for imbalanced data sets has focused on text classification or Web categorization domain [39][19].

A couple of papers in this issue look at feature selection in the realm of imbalanced data sets, albeit in text classification or Web categorization. Zheng and Srihari [62] suggest that existing measures used for feature selection are not very appropriate for imbalanced data sets. They propose a feature selection framework, which selects features for positive and negative classes separately and then explicitly combines them. The authors show simple ways of converting existing measures so that they separately consider features for negative and positive classes. Castillo and Serrano [7] do not particularly focus on feature selection, but make it a part of their complete framework. They use a multi-strategy classifier system to construct multiple learners, each doing its own feature selection based on genetic algorithm. Their proposed system also combines the predictions of each learner using genetic algorithms.

4. SUMMARY

To summarize the Editorial, we attempted to (briefly) chart out the progress in related areas of learning from imbalanced data sets by outlining some of the trends since the AAAI 2000 workshop. The problem of class or cost imbalance is prevalent in various real world scenarios. As this field slowly matures, novel questions and problems stem requiring equally novel solutions. We hope that this Issue stimulates new directions and solutions that can lead to both theoretical insight and practical applications.

5. ACKNOWLEDGEMENTS

We thank the reviewers for their useful and timely comments on the papers submitted to this Issue. We would also like to thank the participants and attendees of the previous workshops for the enlightening presentations and discussions.

6. REFERENCES

[1] In N. Japkowicz, editor, *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets , AAAI Tech Report WS-00-05*. AAAI, 2000.

[2] In T. Dietterich, D. Margineantu, F. Provost, and P. Turney, editors, *Proceedings of the ICML'2000 Workshop on Cost-sensitive Learning*. 2000.

[3] In N. V. Chawla, N. Japkowicz, and A. Kołcz, editors, *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets*. 2003.

[4] In C. Ferri, P. Flach, J. Orallo, and N. Lachice, editors, *ECAI' 2004 First Workshop on ROC Analysis in AI*. ECAI, 2004.

[5] N. Abe. Invited talk: Sampling approaches to learning from imbalanced datasets: active learning, cost sensitive learning and beyond. http://www.site.uottawa.ca/~nat/Workshop2003/ ICML03Workshop_Abe.ppt, 2003.

[6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29, 2004.

[7] M. Castillo and J. Serrano. A multistrategy approach for digital text categorization from imbalanced documents. *SIGKDD Explorations*, 6(1):70–79, 2004.

[8] P. K. Chan and S. J. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceedings of Knowledge Discovery and Data Mining*, pages 164–168, 1998.

[9] N. V. Chawla. C4.5 and imbalanced datasets: Investigating the effect of ampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML'03 Workshipkshop on Class Imbalances*, 2003.

[10] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling TEchnique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 107–119, Dubrovnik, Croatia, 2003.

[12] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, San Diego, CA, 1999. ACM Press.

[13] C. Drummond and R. Holte. Explicitly representing expected cost: An alternative to ROC representation. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 198–207, 2001.

[14] C. Drummond and R. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[15] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[16] C. Elkan. Invited talk: The real challenges in data mining: A contrarian view. http://www.site.uottawa.ca/~nat/Workshop2003/ realchallenges2.ppt, 2003.

[17] W. Fan, S. Stolfo, J. Zhang, and P. Chan. Adacost: Misclassification cost-sensitive boosting. In *Proceedings of Sixteenth International Conference on Machine Learning*, pages 983–990, Slovenia, 1999.

[18] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. http://www.hpl.hp.com/personal/Tom_Fawcett/papers/ index.html, 2003.

[19] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[20] J. Furnkranz and P. Flach. An analysis of rule evaluation metrics. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 202–209, 2003.

[21] H. Guo and H. L. Viktor. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *SIGKDD Explorations*, 6(1):30–39, 2004.

[22] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[23] R. Hickey. Learning rare class footprints: the reflex algorithm. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[24] R. Holte. Summary of the workshop. http://www.site.uottawa.ca/~nat/Workshop2003/ workshop2003.html, 2003.

[25] N. Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. In *Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence*, pages 67–77, 2001.

[26] N. Japkowicz. Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, 42(1/2):97–122, 2001.

[27] N. Japkowicz. Class imbalance: Are we focusing on the right issue? In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[28] N. Japkowicz and R. Holte. Workshop report: Aaai-2000 workshop on learning from imbalanced data sets. *AI Magazine*, 22(1), 2001.

[29] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):203–231, 2002.

[30] T. Jo and N. Japkowicz. Class imbalances versus small disjuncts. *SIGKDD Explorations*, 6(1):40–49, 2004.

[31] M. Joshi, V. Kumar, and R. Agarwal. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In *Proceedings of the First IEEE International Conference on Data Mining*, pages 257–264, San Jose, CA, 2001.

[32] P. Juszczak and R. P. W. Duin. Uncertainty sampling methods for one-class classifiers. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[33] A. Kołcz and J. Alspector. Asymmetric missing-data problems: overcoming the lack of negative data in preference ranking. *Information Retrieval*, 5(1):5–40, 2002.

[34] A. Kołcz, A. Chowdhury, and J. Alspector. Data duplication: An imbalance problem ? In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*, 2003.

[35] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186, Nashville, Tennesse, 1997. Morgan Kaufmann.

[36] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining*, pages 19–22, 2003.

[37] M. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[38] L. M. Manevitz and M. Yousef. One-class SVMs for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001.

[39] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In *Proceedings of the 16th International Conference on Machine Learning*, pages 258–267, 1999.

[40] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[41] R. Pearson, G. Goney, and J. Shwaber. Imbalanced clustering for microarray time-series. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[42] C. Phua and D. Alahakoon. Minority report in fraud detection: Classification of skewed data. *SIGKDD Explorations*, 6(1):50–59, 2004.

[43] F. Provost. Invited talk: Choosing a marginal class distribution for classifier induction. http://www.site.uottawa.ca/~nat/Workshop2003/ provost.html, 2003.

[44] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.

[45] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[46] P. Radivojac, N. V. Chawla, K. Dunker, and Z. Obradovic. Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics*, 2004. Accepted.

[47] B. Raskutti and A. Kowalczyk. Extreme re-balancing for SVM's: a case study. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[48] B. Raskutti and A. Kowalczyk. Extreme rebalancing for svms: a case study. *SIGKDD Explorations*, 6(1):60–69, 2004.

[49] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1472, 2001.

[50] D. Tax. *One-class classification*. PhD thesis, Delft University of Technology, 2001.

[51] K. M. Ting. A comparative study of cost-sensitive boosting algorithms. In *Proceedings of Seventeenth International Conference on Machine Learning*, pages 983–990, Stanford, CA, 2000.

[52] K. M. Ting. An instance-weighting method to induce cost-sensitive trees. *IEEE Transaction on Knowledge and Data Engineering*, 14:659–665, 2002.

[53] P. Turney. Types of cost in inductive concept learning. In *Proceedings of the ICML'2000 Workshop on Cost-Sensitive Learning*, pages 15–21, 2000.

[54] S. Visa and A. Ralescu. Learning imbalanced and overlapping classes using fuzzy sets. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[55] G. Weiss. Mining with rarity: A unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.

[56] G. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.

[57] G. Wu and E. Y. Chang. Class-boundary alignment for imbalanced dataset learning. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[58] B. Zadrozny and C. Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 204–213, 2001.

[59] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the Third IEEE International Conference on Data Mining*, pages 435–442, Melbourne, FL, 2003.

[60] J. Zhang and I. Mani. knn approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*, 2003.

[61] Z. Zheng and R. Srihari. Optimally combining positive and negative features for text categorization. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003.

[62] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explorations*, 6(1):80–89, 2004.