Issues in Mining Imbalanced Data Sets - A Review Paper

Sofia Visa

ECECS Department
ML 0030
University of Cincinnati
Cincinnati, OH 45221, USA
svisa@ececs.uc.edu

Anca Ralescu

ECECS Department
ML 0030
University of Cincinnati
Cincinnati, OH 45221, USA
Anca.Ralescu@uc.edu

Abstract

This paper traces some of the recent progress in the field of learning of imbalanced data. It reviews approaches adopted for this problem and it identifies challenges and points out future directions in this relatively new field.

Introduction

Learning with imbalanced class distributions addresses the case when, for a two-class classification problem, the training data for one class (majority) greatly outnumbers the other class (minority). Recently , machine learning community acknowledged that the current learning methods (e.g. C4.5, NN) perform poorly in applications dealing with imbalanced data sets (IDS). On the other hand, it was observed that in many real world domains available data sets are imbalanced. In the literature, the IDS problem is also known as dealing with rare classes, or with skewed data.

The poor performance of the classifiers produced by the standard machine learning algorithms on IDS is mainly due to the following factors:

(F_1) **Accuracy:** The standard algorithms are driven by accuracy (minimization of the overall error) to which the minority class contributes very little;

(F_2) **Class Distribution:** The current classifiers assume that the algorithms will operate on data drawn from the same distribution as the training data;

(F_3) **Error Costs:** The current classifiers assume that the errors coming from different classes have the same costs.

However, upon closer attention to data, it is observed that the real data sets do not respect the above factors, as the following examples show:

Example 1 (Accuracy driven classifiers fail for some data sets) *For a data set consisting of* 98 *majority examples and only* 2 *minority examples, by assigning all data to the majority class, a* 98% *accuracy is achieved. Now, assuming that the minority class represents a rare disease and it is the class of interest for the domain expert (here, the health care provider), the classifier is rather useless for this real world problem.*

Example 2 (Training and testing distribution are rarely the same) *Class distribution is an important issue for learning, in general. The training data might be imbalanced but the testing might not and the other way around. However, experimental studies show that a balanced class distribution is not the best for learning (Weiss & Provost 2003), (Visa & Ralescu 2005) and the open question for further research is:* **What is the best class distribution for learning a given task?**

Example 3 (In applications, error cost are different) *In applications the error cost are different: consider a cancer versus non-cancer, fraud versus valid action, system OK versus system failure situation. If the error costs and class distribution are known the correct threshold can be computed easily. But the difficulty is that error costs are hard to assess even by the human experts in the field, and therefore, these costs are rarely known. Further, it is important to mention that, when the errors coming from different classes have different but unknown cost, classifiers have problems even for the balanced data.*

In order to give a comprehensive view of the state of the art in this field, we review the most commonly used methods for dealing with IDS, present a chronological review of the events dedicated to IDS and the lessons learned so far. In addition we point out future directions in the field.

Methods in Dealing with IDS

Standard classifiers such as neural networks, support vector machines and C4.5 were investigated in many research papers for the imbalance data problem (Fawcett & Provost 1997), (Chan & Stolfo 1998), (Kubat, Holte, & Matwin 1998), (Japkowicz 2000), (Nickerson & Milios 2001), (Japkowicz & Stephen 2002), (Weiss 2003), (Maloof 2003), (Drummond & Holte 2003), (Estabrooks & Japkowicz 2004), (Weiss 2004). It is commonly agreed that they are heavily biased in recognizing mostly the majority class since they are built to achieve overall accuracy to which the minority class contributes very little. Solutions to the class imbalance problem were proposed both at the data and algorithmic levels: different forms of re-sampling address the first type of algorithms and adjusting costs (Pazzani *et al.*

1994), decision thresholds and recognition based classification ((Kubat, Holte, & Matwin 1998), (Japkowicz, Myers, & Gluch 1995)), for the second type of algorithms.

In the up-sampling approach (Ling & Li 1998), data from the minority class are duplicated until the imbalance is eliminated; Likewise, in the down-sampling approach (Kubat & Matwin 1997) data from the majority class are eliminated to rebalance the classes. Combinations of the two methods in an ensemble classifier and an effective ratio of up/down sampling were also investigated in (Estabrooks & Japkowicz 2004) and applied to text categorization.

A guided resampling study is presented in (Nickerson & Milios 2001): the sampling method takes into account *within-class imbalance*, assuming that elements of the imbalanced class are grouped in subclusters. Up-sampling is then guided toward enlarging the under-represented clusters (known also as small disjuncts). Since the up-sampling step assumes that the number of subclusters is known in advance, unsupervised learning methods (e.g. k-means algorithm or self-organizing maps) have to be used prior to up-sampling, in order to obtain such information about the imbalanced class.

In (Zhang & Mani 2003) five different down-sampling algorithms are presented, each being a variation of the k-nearest neighbor technique. (Chawla *et al.* 2002) propose a new up-sampling method(SMOTE): new artificial examples are created for the minority class by interpolating minority-class examples. The method is investigated for C4.5 and gives better results than random up-sampling. By interpolating the minority class examples with new data, the within class imbalance is reduce (fewer small disjuncts) and C4.5 achieves a better generalization of the minority class, opposed to the specialization effect obtained by randomly replicating the minority class examples.

(Chan & Stolfo 1998) investigates the best learning class distribution for fraud detection in a large and imbalanced credit card data set. Multiple data sets with the desired class distribution are generated by joining the minority class with subsets of the majority class. The obtained classifiers are aggregated in an ensemble (composite) classifier. A drawback of this approach is that the best learning class distribution must be investigated prior to the subset generation.

Starting from the original imbalanced training set, (Chan & Stolfo 1998) generate several balanced training sets (each including the minority class and a part of the majority class).

Another common approach to deal with the imbalance aspect is to assign different error penalties: an error on a data point belonging to the small class would have a larger penalty than one for the large class. However, it can be argued that the effect of assessing penalties is equivalent to changing the relative data distribution in the two classes, or, in other words, to re-balancing the data. The experiments carried in (Maloof 2003) suggest that sampling, as well as adjusting the cost matrix, have the same effect as moving the decision threshold.

(Visa & Ralescu 2003) propose a fuzzy based classifier for IDS which means to be less sensitive to the class imbalance by considering a relative frequency to the class size. In the same paper the sensitivity of the fuzzy classifier to the imbalance in the presence of overlap is investigated. The results show that in fact the overlap affects more the classifier than the imbalance. Further in (Visa & Ralescu 2004b) other factors such as complexity and size in combination with imbalance are studied. The results on the artificial data sets suggest that the fuzzy classifier is affected by the imbalance only for a combination of high complexity and small size of the overall data (in this case the major drawback is the lack of information, rather than the majority:minority ratio - the same issue is discussed later in the current paper for C4.5), unlike the neural networks which are consistently biased toward the majority class at any given size (Japkowicz & Stephen 2002). Also, the experiments from (Japkowicz & Stephen 2002) show that C4.5 is affected by high complexity and imbalance at any given size.

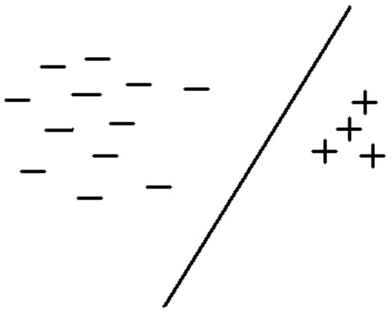

Figure 1: Linear separable data.

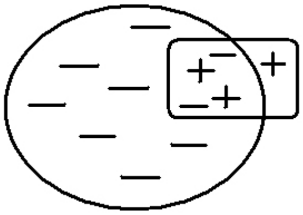

Figure 2: Overlapping data.

Performance Evaluation on IDS

Selecting the best classifier for a given classification problem is the ultimate goal. Accuracy helps in the previous decision only if the error costs are the same, but this is rarely true in practice even for the balanced data. In the same idea, ROC curves help in deciding the best classifier only if one classifier's ROC curve completely dominates the other ROC curves, but this scenario is also rare because

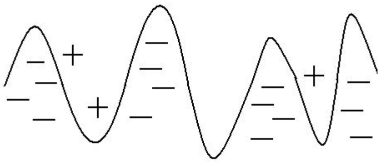

Figure 3: High complexity data.

many times multiple curves dominate in different parts of the ROC space. In this case (Provost 2000) recommends the ROC convex hull as a discriminant. However, as the recent debate within the first workshop dedicated to the ROC analysis held in conjunction with the European Conference on Artificial Intelligence 2004, the ROC curves are difficult and computationally expensive to extend to more than two classes.

Issues on IDS

In this section we analyze the imbalance factor from various directions and we will focus on answering questions, such as:

(Q_1) When does the imbalance matter?

(Q_2) Is data re-balancing necessary? That is, "$50 : 50$" is the desired (best) distribution?

(Q_3) Are there some learning algorithms insensitive to the imbalance?

First question is worth to investigate since some research papers claim that other factors in combination with the imbalance lead to poor performance (Jo & Japkowicz 2004), (Visa & Ralescu 2003), (Visa & Ralescu 2004b). For the second question, initial results are reported in (Weiss & Provost 2003) and (Visa & Ralescu 2005), whereas for the third question there is little work reported (Visa & Ralescu 2004b).

The Nature of IDS

From the application point of view, the nature of the imbalance falls in two cases:

- The data are naturally imbalanced (e.g. credit card frauds and rare disease) or,
- The data are not naturally imbalanced but it is too expensive to obtain data for learning the minority class (e.g. shuttle failure).

An open question remains as whether these two types of imbalance should/can be differentiated, and whether they should be addressed differently: for instance, deciding against (in the first case) or in favor (in the second case) of rebalancing.

Class Distribution in IDS

Another important aspect in learning IDS is the class distribution, that is the ratio minority/majority examples in the training set. (Weiss & Provost 2003) investigate, on several data sets from the UCI Repository, the effect of various class distribution for learning, when testing with the natural class distribution. The experimental results show that, for C4.5, neither a $50 : 50$(that is balanced) learning distribution, nor the natural distribution are the best for the learning task. Similar results are reported in (Visa & Ralescu 2004b) for a fuzzy classifier where, the fuzzy classifier proves to be less sensitive to the learning class distribution (its output was less variant than C4.5's output).

Imbalance Ratio versus Lack of Information

We believe that a major issue in the IDS problem is to distinguish between two components of the IDS:

(IR) the imbalance as the ratio $\frac{NumberOfMinority}{NumberOfMajority}$;

(LI) the lack of information for the minority class.

Both the above components are present in an IDS learning problem but each, in combination with other factors (such as overlap, complexity of the function to be learned, size of the data sets and small disjuncts), affects a particular learning algorithm differently. Of course, all the algorithms suffer from lack of representation (what is not present in training, cannot be learned), but it is crucial to determine which ones do not suffer from the imbalance ratio (Visa & Ralescu 2004b).

Example 4 *For a data set consisting of* $5 : 95$ *minority:majority examples the imbalance factor* IR *is the same as in a data set of* $50 : 950$*. Though, in the first case the minority class is poorly represented and suffers more from the* LI *factor than in the second case.*

Other Factors - Size, Complexity, Overlap and Small Disjuncts

Example 4 shows that the overall size of the data set matters to the extend to which adds more information and this is crucial when the data to be learned have high complexity (as in Figure 3). If the curve in the Figure 3 is the real boundary between classes, some additional data for the minority class might come from the area where currently there is no data for learning (the valley in the middle) and the recognition of the minority class will improve, even though the imbalance ratio IR is the same.

Clearly, for simple data sets (e.g. linearly separable) any classifier produces a good discrimination, regardless of the amount of imbalance IR (Figure 1). So, for low complexity the imbalance seems not to matter.

For overlapping classes (see Figure 2), that is, when the some data points belong to both classes, it is harder to analyze the IR and LI components. However, it is obvious that for the accuracy-driven algorithms, the tendency is to reduce the overall error by assigning the overlapped area to the majority class (since from the overlapping area there are more majority class data than minority ones). For example, neural networks are biased toward the majority class: the overlapping examples for the minority class are considered noise and are discharged.

To address the question (Q_3), we point out that the fuzzy classifier proposed in (Visa & Ralescu 2003) learns *independently* each class, and represents them as fuzzy sets such that each example has a *membership degree to its class computed relatively to the class size* thus giving chances to the minority class too, even when the classes overlap. The importance of this feature was also pointed out in (Chawla, Japkowicz, & Kolcz 2004).

It is also expected that SVM classifiers are less affected by the imbalance, since only support vectors are considered in classification. Basically, the idea is that the support vectors are selected only from the points around the boundaries so the imbalance should affect less (or not at all). As in the case of the fuzzy classifier, the imbalance factor affects SVM classifiers only by the lack of information (LI) and not by biasing it toward the majority class (RI).

(Weiss 2003) and (Jo & Japkowicz 2004) identify small disjuncts (sub-clusters in the minority class, very poorly represented in the training phase) as another factor when dealing with IDS. They claim that up-sampling must be guided such that more artificial up-sampled data must come from the small disjuncts in order to overcome the lack of representation of the minority class.

Issues on the Current Approaches

Next we point out problems of standard learning methods when applied to IDS.

Rebalancing - Loss or Gain in Information

The sampling method has known drawbacks: under-sampling involves a loss of information (by discarding potentially useful data) and over-sampling increases the training size without any gain in information (Provost 2000) or, perhaps even worse, by replicating the same examples leads to over-fitting. Considering this fact, the best research strategy is to concentrate on how machine learning algorithms can deal most effectively with whatever data they are given. Thus, the research must focus in developing new classifiers/learning methods.

NN and C4.5

Neural networks have a large range of applications in inductive learning, but in the case of imbalanced data sets, neural networks are prone to treat the minority class as noise and therefore to disregard it. Therefore, in this context, most often they are used in conjunction with up/down-sampling of the training data as shown in several research papers. It is observed that random down-sampling outperformed random up-sampling method (Japkowicz, Myers, & Gluch 1995); up-sampling of the minority class leads to a slower convergence of the network and does not bring new information. (Japkowicz, Myers, & Gluch 1995) shows experimentally (on three data sets) that a neural network recognition-based system performs better or equally well to a discriminating NN approach.

Decision trees algorithms (C4.5) otherwise very popular choice for learning classification rules, are also limited in the presence of imbalance: a common problem is that they might loose big parts of the minority class at the pruning phase, and therefore, for such problems, C4.5 without pruning performs better than C4.5 with pruning (Weiss 2003). Further, random up-sampling of the minority class leads to trees of larger size and over-fitting of the minority class (e.g. in the extreme case, each example of the minority class will correspond to a leaf in the tree, in which case there is no rule ”learned” for the minority class).

Historical Threat of the IDS Problem

The issue of learning of imbalanced classes has recently received increased attention, as it can be seen from the recent increase of scientific events dedicated to this topic. Conferences gather researchers to discuss new ideas and the results of their work. Usually held in conjunction with conferences, workshops can be seen as offsprings of conferences that contribute to the field by guiding research into even newer areas of interest.

Several workshops were dedicated specifically to the IDS problem as follows.

AAAI’2000 - Workshop on Learning from Imbalanced Data Sets I

The first workshop dedicated to the IDS problem was held in conjunction with the American Association for Artificial Intelligence Conference 2000. The main observation at the time was that there are many domains dealing with imbalanced data sets, such as:

- medical diagnosis (e.g. rare disease and rare genes mutations);
- network monitoring and behavior profiling - intrusion;
- fraud detection (e.g. credit card, phone calls, insurance)(Fawcett & Provost 1997);
- risk management;
- helicopter gear-box fault monitoring;
- shuttle system failure;
- earthquakes and nuclear explosions;
- text classification;
- oil spills detection.

The issues debated at this workshop can be summarized as follows (Japkowicz & Holte 2001):

1. How to evaluate learning algorithms;
2. Evaluation measures: it was commonly agreed that accuracy yields misleading conclusions. Instead, ROC and cost curves were proposed;
3. One class learning versus discriminating methods;
4. Discussions over various data resampling;
5. Discussion of the relation between class imbalance problem and cost-sensitive learning;
6. The goal of creating classifiers that performs well across a range of costs.

ICML'2003 - Workshop on Learning from Imbalanced Data Sets II

The main issues discussed on the 2000 workshop guided the research on IDS for the second workshop held as part of the International Conference on Machine Learning 2003. For example, ROC or cost curves were used as method of evaluation, rather than accuracy. The workshop was followed by an interesting and vivid panel discussion.

(Japkowicz 2003) questioned the fact that the within class imbalance is responsible for the IDS mall-learning. The idea is that the within class imbalance (the small disjuncts) leads to a sever lack of representation of some important aspects of the minority class. (Zhang & Mani 2003) investigate various selection methods for down-sampling the majority class based on k Nearest Neighbors algorithms and (Nickerson & Milios 2001) investigate the relative advantage, if any, of down/up-sampling techniques.

Besides NN and C4.5, other classification methods, such as SVM ((Wu & Chang 2003) and (Raskutti & Kowalczyk 2003)) and, for the first time, a fuzzy classifier (Visa & Ralescu 2003), were investigated for IDS. (Wu & Chang 2003) point out potential problems such as skewed boundary toward minority class due to its lack of representation of SVM when applied to IDS. Despite the fact that one of the conclusions of the first workshops was that there was a need for *new* classifiers for IDS, the papers presented at the second workshop aimed mainly to tune existing classifiers, e.g. C4.5, to perform better on IDS. (Visa & Ralescu 2003) proposed a fuzzy classifier based on relative class size, that proved to be less sensitive to the class imbalance. In the same paper, the effect of the imbalance combined with various degrees of overlap was studied and it was concluded that the overlap affects more the classification task than the imbalance.

We now identify two major (and what we think wrong) directions present in the research papers of the workshop:

1. Many papers reported various tuning methods applied to decision trees in order to perform better on IDS, even though presentations in the previous workshop showed their shortcoming, and it was commonly agreed that new methods/classifiers are needed for IDS;
2. Sampling, under various aspects, was present in half of the papers and was the most debated issue, even though (Maloof 2003) shows that sampling has the same result as moving the decision threshold or adjusting the cost matrix (a result known since 1984 (Breiman & Stone 1984)).

ACM SIGKDD Exploration 2004 - Special Issue on Learning from Imbalanced Data Sets

The sixth issue of SIGKDD Exploration was dedicated entirely to the imbalance problem. (Weiss 2004) presents a very good review of the current research on IDS. The other papers in the volume address mainly the issue of sampling, feature selection and one class learning.

(Guo & Herna 2004) investigate a boosting method combined with various up-sampling techniques of the hard to classify examples. The method improves the prediction accuracy for both the classes and does not sacrifice one class for the other. Results on 17 data sets are presented.

(Batista & Monard 2004) suggest that in fact, it is not imbalance, but other factors such as the overlap between classes which hinder the learning system. This fact was pointed out firstly in (Visa & Ralescu 2003). (Jo & Japkowicz 2004) suggest that the small disjuncts are responsible for the imbalance problem and investigates the idea on artificial and real domains.

Imbalanced Data in Midwest Artificial Intelligence and Cognitive Science Conference (MAICS)

Learning in various forms is the major theme for MAICS conference and IDS is identified as a learning drawback in some MAICS papers too. For example, (Lei & Cheh 2003) investigate the performance of a rule-based classifier for bankruptcy data. They observe that decision trees learners learn well the non-bankruptcy data but consistently learns poorly the bankruptcy data. We suspect that, since their data sets are highly imbalanced (e.g. in one of their data sets the minority class accounts for 0.25% of all data: 12 bankruptcy and 4771 non-bankruptcy data), the decision trees over-fit the non-bankruptcy class, which is the majority class here.

(Kamei & Ralescu 2003) present a piecewise linear separable SVM classification. As discussed earlier, the SVM are better suited to IDS since only support vectors are considered in classification (only the LI component affects SVM). The SVM described in this paper can be applied for IDS: each iteration considers a subset of all data (containing data from both classes) and the optimum hyperplane using SVM is selected. The subsets can be formed such that there is no imbalance in the training. Network intrusion is an example of highly imbalanced domain, and (Miller & Inoue 2003) introduces a framework for intrusion detection - SPIDER.

(Visa & Ralescu 2004a) present initial results of up-sampling methods based on various approaches to aggregation of class information such as spread, imbalance factor and the distance between the classes. Artificially generated data are used for experiments. The performance of each up-sampling method is evaluated with respect to how well the resulting data set reflects the underlying original class distribution, in terms of mean, standard deviation and (empirical) distribution function.

Some Lessons Learned

Some lessons can be drawn from reviewing the emergence of IDS as a separate and important issue in machine learning:

- Many papers report on improvements of one method over another for some real data sets, e.g. UCI Repository. On the other hand, the current learning algorithms might ”overfitt” the UCI Repository (Lavrac & Fawcett 2004). Therefore, papers that show limitations of current learning methods are of interest as well.
- People involved in applications of machine learning have known for some time that IDS create problems. However, the machine learning community started to pay more

attention to this problem only recently, when it was observed that the current standard learning methods behaved poorly in IDS. The major lesson learned here is that the scientific research must be guided more closely by the industry since it is designed for it. The industry must set the environment and the requirements for the problems to be solved by future research.

- The gap between the scientific research and the applications must diminish, if research is to fulfill its ultimate goal of being applied to specific domains. Therefore, experts in domains must guide the researchers by specifying the expectations. For example, data mining and knowledge discovery conferences host workshops on new challenging topics triggered by the market: IDS, bioinformatics, text and web mining, multimedia mining, etc. With a better setup of what is needed in a particular field, the researchers may reformulate the problem and design the solutions accordingly to the external context.

- Future research on learning methods must be focused on new areas. For example, since each learning algorithm has its unique strength, investigations of how learning algorithm are different and further, what types of learning problems (data sets) trigger which specific method are to be considered. The evaluation must be performed accordingly to the expectations of the experts in the field of the problem (not necessary the accuracy).

- Future research must focus on better data understanding. For example, for the the KDD COIL Cup 2000, researchers were challenged with a specific data mining/classification task and the participants were evaluated by how well their entries satisfied the domain expert. The question was: *"Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?"* The data set given to the participants was noisy, imbalanced, correlated and high dimensional with a weak relation between input and target. The winner's (Elkan 2000) entry based on the naive Bayesian learning satisfied the domain expert by identifying the two most important features that predict who would buy a caravan insurance. The judging panel of the contest acknowledged that the winner's clearly explained solution helped in their decisions and was preferred over the complicated performance analysis submitted by the other participants.

- Feature selection remains an important field for machine learning research: (Van Der Putten & Van Someren 2004) analyzed the COIL 2000 data sets using the bias-variance decomposition and they reported that the key issue for this particular data set was avoiding overfitting. They conclude that feature selection in such domains is even more important than the choice o the learning method.

Conclusions

We attempted to convey a global picture of the research in IDS as it has emerged from events (publications and meetings) in machine learning and data mining dedicated to it. Some lessons and new research directions are pointed out.

Acknowledgments

This work was partially supported by the Ohio Board of Regents, Grant N000140310706 from the Department of the Navy, and the Rindsberg Graduate Fellowship of the College of Engineering, University of Cincinnati.

References

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

Breiman, L.;Friedeman, J. R., and Stone, C. 1984. Chapman and Hall/CRC Press.

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

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, W. 2002. Smote: Synthetic minority over-sampling technique. *Artificial Intelligence Research* 16:321–357.

Chawla, N.; Japkowicz, N.; and Kolcz, A. 2004. Editorial: Special issues on learning from imbalanced data sets. *SIGKDD Explorations* 6(1):1–6.

Drummond, C., and Holte, C. 2003. C4.5, class imbalance, and cost sensitivity: Why undersampling beats oversampling. In *Proc. of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II*, 1–8.

Elkan, C. 2000. Coil challenge 2000 entry. http://www.cse.ucsd.edu/users/elkan/papers.

Estabrooks, A.; Jo, T., and Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* 20(1):18–36.

Fawcett, T., and Provost, F. 1997. Adaptive fraud detection. *Data Mining and Knowledge Discovery* 1(3):291–316.

Guo, H., and Herna, V. 2004. Learning from imbalanced data sets with boosting and data generation: The databoost-im approach. *SIGKDD Explorations* 6(1):30–39.

Japkowicz, N., and Holte, R. 2001. ”workshop report: Aaai-2000 workshop on learning from imbalanced data sets”. *AI Magazine* 22(1):127–136.

Japkowicz, N., and Stephen, S. 2002. The class imbalance problem: A systematic study. *Intelligent data Analysis* 6(5):429–450.

Japkowicz, N.; Myers, C.; and Gluch, M. 1995. A novelty detection approach to classification. In *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*, 518–523.

Japkowicz, N. 2000. Learning from imbalanced data sets: A comparison of various strategies. In *Proceedings of Learning from Imbalanced Data*, 10–15.

Japkowicz, N. 2003. Class imbalances: Are we focusing on the right issue? In *Proc. of the ICML-2003 Workshop:Learning with Imbalanced Data Sets II*, 17–23.

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

Kamei, R., and Ralescu, A. 2003. Piecewise linear separanility using support vector machines. In *Proc. of the MAICS Conference,* 52–53.

Kubat, M., and Matwin, S. 1997. Addressing the curse of imbalanced data set: One sided sampling. In *Proc. of the Fourteenth International Conference on Machine Learning,* 179–186.

Kubat, M.; Holte, R.; and Matwin, S. 1998. Machine learning for the detection of oil splis in radar images. *Machine Learning* 30:195–215.

Lavrac, N.; Motoda, H., and Fawcett, T. 2004. Editorial: Data mining lessons learned. *Machine Learning* 57(1-2):5–11.

Lei, H.;Chan, C., and Cheh, J. 2003. Rule-based classifier for bankruptcy prediction. In *Proc. of the MAICS Conference*, 74–81.

Ling, C., and Li, C. 1998. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth ACM SIGKDD,* 73–79.

Maloof, M. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proc. of the ICML-2003 Workshop:Learning with Imbalanced Data Sets II,* 73–80.

Miller, P.;Mill, J., and Inoue, A. 2003. Synergistic and perceptual intrusion detection with reinforcement learning (spider). In *Proc. of the MAICS Conference*, 102–108.

Nickerson, A.S.; Japkowicz, N., and Milios, E. 2001. Using unsupervised learning to guide resampling in imbalanced data sets. In *Proc. of the Eighth International Workshop on AI and Statistics*, 261–265.

Pazzani, M.; Marz, C.; Murphi, P.; Ali, K.; Hume, T.; and Brunk, C. 1994. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, 217–225.

Provost, F. 2000. Machine learning from imbalanced data sets. Proc. of the AAAI'2000 Workshop on Imbalanced Data Sets.

Raskutti, B., and Kowalczyk, A. 2003. Extreme rebalancing for svm's: a case study. In *Proc. of the ICML-2003 Workshop:Learning with Imbalanced Data Sets II*, 57–64.

Van Der Putten, P., and Van Someren, M. 2004. A bias-varianve analysis of a real world learning problem: the coil challenge 2000. *Machine Learning* 57(1-2):177–195.

Visa, S., and Ralescu, A. 2003. Learning imbalanced and overlapping classes using fuzzy sets. In *Proc. of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II,* 97–104.

Visa, S., and Ralescu, A. 2004a. Experiments in guided class rebalance based on class structure. In *Proc. of the MAICS Conference*, 8–14.

Visa, S., and Ralescu, A. 2004b. Fuzzy classifiers for imbalanced, complex classes of varying size. In *Proc. of the IPMU Conference, Perugia,* 393–400.

Visa, S., and Ralescu, A. 2005. The effect of imbalanced data class distribution on fuzzy classifiers - experimental study. In *Proc. of the FUZZ-IEEE Conference*.

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

Weiss, G. M. 2003. The effect of samall disjuncts and class distribution on decision tree learning. *PhD Thesis, Rutgers University*.

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

Wu, G., and Chang, E. 2003. Class-boundary alignment for imbalanced dataset learning. In *Proc. of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II*, 49–56.

Zhang, J., and Mani, I. 2003. knn approach to unbalanced data distributions: A case study involving information extraction. In *Proc. of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II, 42-48,* 42–48.