Dipartimento di scienze economiche, aziendali, matematiche e statistiche “Bruno de Finetti”

Working Paper Series, N. 1, 2010

Effect of training set selection when predicting defaulter SMEs with unbalanced data

GIOVANNA MENARDI
Department of Statistical Sciences
University of Padova

NICOLA TORELLI
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti"
University of Trieste

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Working paper series

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
"Bruno de Finetti"
Piazzale Europa 1
34127, Trieste

EUT Edizioni Università di Trieste
Via E.Weiss, 21 - 34128 Trieste
Tel. ++40 558 6183
Fax ++40 558 6185
http://eut.units.it
eut@units.it

ISBN: 978-88-8303-320-9

Working Paper Series, N. 1, 2010

Effect of training set selection when predicting defaulter SMEs with unbalanced data

GIOVANNA MENARDI
Department of Statistical Sciences
University of Padova

NICOLA TORELLI
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti"
University of Trieste

ABSTRACT

We focus on credit scoring methods to separate defaulter small and medium enterprises from non-defaulter ones. In this framework, a typical problem occurs because the proportion of defaulter firms is very close to zero, leading to a class imbalance problem. Moreover, a form of bias may affect the classification. In fact, classification models are usually based on balance sheet items of large corporations which are not randomly selected. We investigate how different criteria of sample selection may affect the accuracy of the classification and how this problem is strongly related to the imbalance of the classes.

KEYWORDS: defaulter business, training set selection, unbalanced data.

1. Introduction

The second Basel capital accord (Basel II, 2004) and its recent and updated version Basel III (2010) define recommendations and directives with the purpose of creating a set of international standard rules on banking laws. The accords delineate the consistency of capital regulations, aim at guaranteeing banking credit policies more risk sensitive, and suggest the required prerequisites for an internal rating based approach. Since their introduction, the internal banking processes of modeling and measuring the credit risk have been strongly affected, *e.g.* in terms of lower capital requirements in comparison with external rating information (Altman and Sabato, 2005).

Credit risks models and methods are meant to find rules for measuring the risk associated with credit applications or separating defaulter credit applicants from non-defaulter ones. These methods follow both non statistical approaches based on linear programming, genetic algorithms or neural network and statistical approaches as classification algorithms or ensemble techniques (see, for example, Thomas et al. (2002)). Usually classification techniques aim at finding what would have been the best rule to apply on a sample of previous applicants (the so called *training set*). The advantage of this procedures is that the subsequent behavior of the training applicants is known.

In this work we focus on credit scoring methods to separate defaulter small and medium enterprises (SMEs) from non-defaulter ones. In this framework, a typical problem occurs because the proportion of defaulter firms is very close to zero, leading to an imbalance class situation. A form of bias may affect the classification. In fact, the classification models are usually based on balance sheet items of large corporations which are not randomly selected (see, for example, Dietsch and Petey (2004)). This problem has been largely ignored by the literature about credit risk analysis. We investigate how different criteria of sample selection may affect the accuracy of the classification and how this problem is strongly related to the class imbalance problem.

Section 2 describes briefly the effects of the use of unbalanced data and the existing remedies. In Section 3, the problem and the severity of sample selection bias is investigated through a simulation study. Results from an application to real data show that the class imbalance hinders the sample selection bias. Section 4 introduces a method for generating new data and then facing the issue of sample selection without the intrusion of the imbalance effect. Some final considerations and recommendations conclude the paper.

2. The class imbalance problem

Most of the classification methods base their theoretical efficiency on the assumption that the distribution of the classes is well balanced over the training set. However, there exist many contexts where this assumption is not met because the large majority of instances are concentrated only in some classes while one class (usually the most interesting) is rare. Examples include identifying fraudulent credit card transactions (Chan and Stolfo, 2001), defaulter credit applicants (Stanghellini, 2006.), cancerous cells from radiographies (Woods et al., 1993), learning word pronunciations (Van den Bosch et al., 1997.), detecting oil spills from satellite images (Kubat, 1998). In certain domains (like those just mentioned) the class imbalance is intrinsic to the problem. However, class imbalance sometimes occurs when the

data collection process is limited (e.g., due to economic or privacy reasons), thus creating an artificial imbalance.

In the recent years, the issue of class imbalance has been widely investigated by the statistical and machine learning communities (see for a review, Japkowicz and Stephen (2002) or Weiss (2004)). Two main sorts of consequences arise in learning with rare classes: the standard measures of accuracy of the model are not appropriate and, even worse, classifiers tend to learn from the large classes and ignore the rare events.

a. Measuring the accuracy of the models

In the context of rare events, the use of common evaluation measures such as overall error (proportion of misclassified units), can lead to misleading conclusions. For example, in a problem where the rare class is represented in only 1% of the data, the naive strategy of allocating each data to the prevalent class would achieve a good level of accuracy, presenting an overall error equal to 1%. For this reason, alternative measures of accuracy should be used, which consider separately the accuracy in predicting the prevalent class and the accuracy in predicting the rare class. In a binary decision problem, the classified data are labeled as either positive or negative examples. The classifier outcome can be represented in the so-called confusion matrix, a contingency table representing the occurrence of four categories: (i) true positives (TP), examples correctly labeled as positives, (ii) false positives (FP), negative examples incorrectly labeled as positive, (iii) true negatives (TN), negative examples correctly labeled as negative., (iv) false negatives (FN), positive examples incorrectly labeled as negative.

In class imbalance problems, the evaluation measures take into account the different propensity toward false positive and false negative (where the positive examples usually denote the rare class). Hence, the most common measures are precision (fraction of examples classified as positive that are truly positive) and recall (fraction of positive examples that are correctly labeled). A more general measure to evaluate the performance of a classifier in presence of rare events is the ROC curve, which plots the True Positive Rate (1- False Negative rate) vs the False Positive rate, when the classification threshold varies. The classifier performs as better as steeper is the ROC curve that is, as larger is the area underlying the curve.

b. Learning with rare events

Except the uninteresting situations where the classes are perfectly separated, when data with different labels overlap, most of standard classifiers tend to be overwhelmed by the prevalent classes and ignore the rare one. In fact, most of classification methods generally are conceived to estimate from the sample the simplest model that best fits the data. The simplest model, however, pays less attention to rare cases in an imbalanced data set. Therefore, classification rules that predict the rare class tend to be fewer and weaker than those that predict the prevalent class. Consequently, test examples belonging to the rare class are misclassified more often than those belonging to the prevalent class.

Several strategies for dealing with the class imbalance problem have been proposed in the literature. Almost all of them are designed for the binary scenario, where one class is represented by a large number of data while another is represented by only a few, tipically

with higher identification importance. Reported solutions are developed at both the data and algorithmic levels.

At the data level, the objective is to re-balance the class distribution by re-sampling the data space. Solutions at the data-level include random oversampling the rare class with replacement, random undersampling the prevalent class, directed oversampling the rare class (where 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 (Japkowicz, 2000).

At the algorithm level, solutions try to adapt existing classifier learning algorithms to strengthen learning with regards to the rare class. Cost-sensitive learning methods incorporate both the data and algorithmic level approaches by assuming higher misclassification costs of the rare examples and minimizing the high cost errors.

In this paper a recent strategy to deal with the class imbalance problem is applied to understand how the sample selection bias affects the classification. Namely, the imbalance problem is reduced by the artificial generation of new data. This new technique will be illustrated in section 4.

3. Effects of sample selection

Many financial companies which provide economical analysis and risk ratings for enterprises base their results on models which are estimated by using information about not randomly selected business, such as for example large corporations. Our natural conjecture is that the use of non random samples may affect the accuracy of the results.

a. Simulated data application

A simulation study has been conducted to show how non random criteria of sample selection may lead to misleading results.

A set of covariates have been generated as follows: x_1 has an Uniform distribution; x_2 has a Gaussian distribution with moments depending on the values of x_1, x_3 is an Exponential random variable which has been used to select the samples. The dependent variable y takes values in $\{0, 1\}$ and is related to the covariates through a non linear probit model with parameters chosen to control the dependence of y on the covariates.

We have splitted the generated population into three subpopulations according to the value of x_3: S (small values of x_3), M (medium values of x_3), L (large values of x_3) and we have supposed to be interested in classifying data with small values of x_3 (this choice corresponds to our focus on predicting the default event of SME, using data coming from SMEs or larger business).

The procedure to generate y has been adjusted in order to take account of different proportions of events (set to 50%, 10%, 5 ‰). We have randomly generated a training set from S, M, L and from the whole population (T) and we have estimated a logit model on the selected data. The accuracy of the model has been evaluated on a test set drawn from S. This procedure has been iterated several times in order to guarantee the stability of the results.

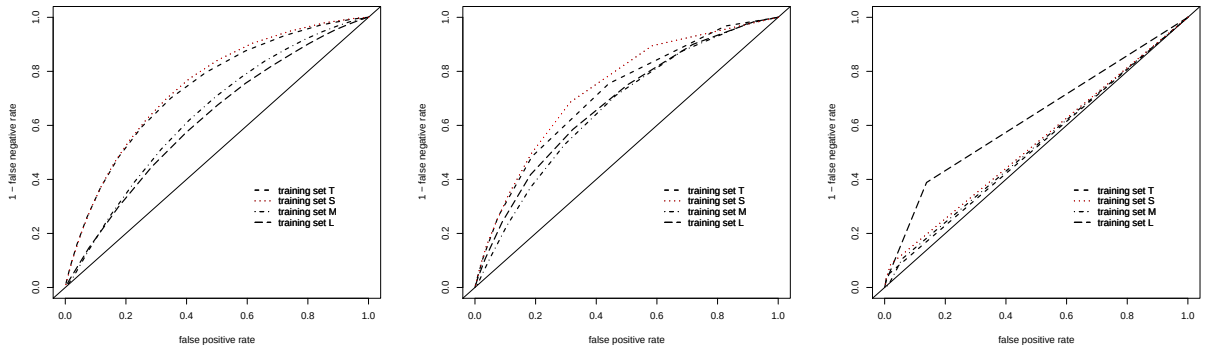


FIG. 1. Comparison between models built on simulated data. On each panel different curves correspond to the sample selection from different subpopulations. The test set has been drawn from S. The three panels refer to different proportions of events in the population (50%, 10%, $5\%_o$ respectively).

The left panel of Figure 1 shows the effect of selecting samples from subpopulations different from the target S. The loss of accuracy due to the sample selection effect is remarkable. However, this effect get smaller as the imbalance level increases and it is completely hidden when the class distribution is highly skew.

For this reason we have adopted one of the standard procedures recommended to cope with rare events, consisting in oversampling the rare class in such a way to obtain balanced samples. Results reported in Figure 2 show an increased accuracy of the models and the sample selection effect arises even in the most unbalanced situations.

This procedure has been implemented also by using decision trees and linear discriminant analysis as a classification technique with similar results.

b. Real data application

The population of enterprises may be partitioned into the subsets of micro enterprises (MIE), small and medium enterprises (SME) and big enterprises (BE). This partition is made according to some criteria which depends on the headcount, the turnover and the balance sheet total of the company.

Our goal is to find an accurate rule to classify SMEs according to their risk of default. In order to use the balance sheet information, we have to restrict the attention on corporations, which are the only companies required by law to present the balance sheet to the Business Register.

Data at hand consist of vital statistics, balance sheet records and financial ratio of all the commercial companies enrolled to the Business Register and located in a province of the North Eastern part of Italy. The occurrence of a bankruptcy condition is considered as the default event. The whole data set includes 11199 enterprises with 76 defaulter firms only. Stratification of the data into the categories of and MIEs, SMEs and BEs results in increasing the gravity of the class imbalance, because fewer positive examples are available in each class of firms.

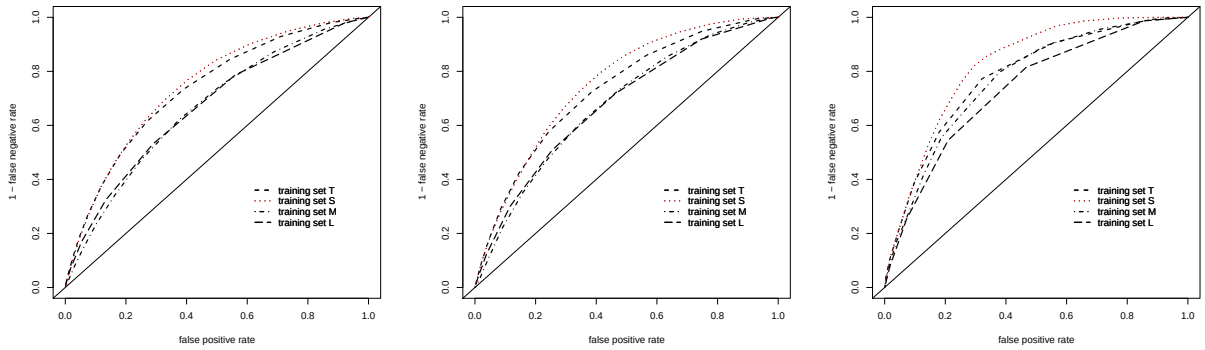


FIG. 2. See Figure 1. The training sets have been selected according to a disproportioned stratified sample in order to balance the classes.

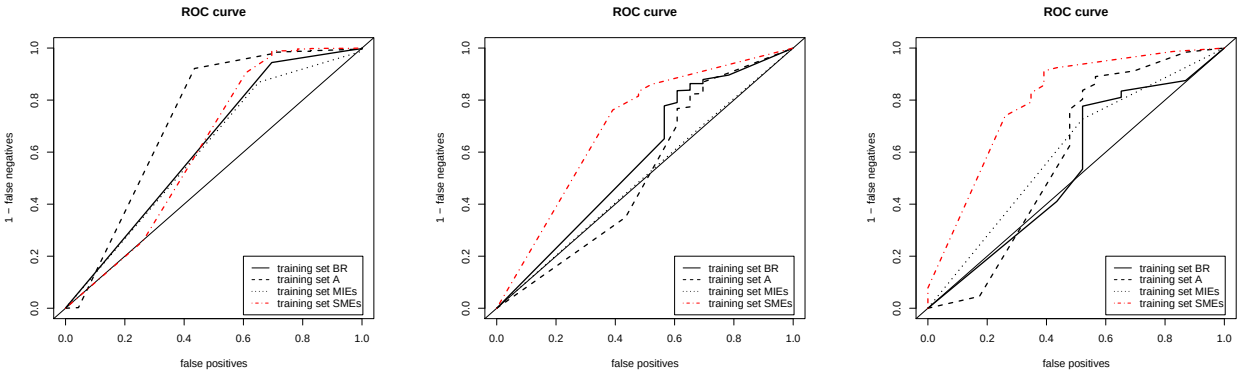
We have considered the following sets of companies: the micro enterprises, the small and medium enterprises and the whole set of enterprises (BR). Moreover, we have considered a sample of firms (A) which exceed a certain threshold of the turnover. This threshold has been set equal to 500000 euros and this choice can be motivated because data commonly available for classification purposes are selected with similar criteria (for example the well known AIDA data provided from the Bureau van Dijk Electronic Publishing). Considering a set of BE has not been possible because of the presence of too less defaulter events in that set of business.
From each set of firms we have selected a training set and we have built some standard classifiers (logit models, decision trees, discriminant analysis methods). After, we have evaluated the performances of the estimated models in classifying new sample points belonging to a test set selected from the SMEs subpopulation.

In a first analysis we have ignored the class imbalance by randomly partitioning the sets of business in a training set and a test set.

The left panel of Figure 3 compares the ROC curves built by estimating a logit model on the 4 training samples and evaluating it on the SME test sets. Analogue results derive from the application of different classification models as linear discriminant analysis or decision trees.

The models estimated by using the MIE, BR and SMEs training samples perform just slightly better than the random choice. Better results derive from the use of the models built on the sample selected from A. Our analysis suggests that these results derive from the unequal conditions of the comparison: the 4 models have been estimated by using differently sized training sets (75% of the size of the population from which we have drawn the sample) with a different rate of imbalance between the classes. In fact, besides the proportion of rare events is very close to zero in each set of business, the A set is more than twice sized than the SMEs set and it has an higher rate of positive examples than the other sets. These differences affect the classification and hinder the sample selection effects.

Therefore, we have conducted a comparison between the models on equal terms, by using the same sample size as well as the same proportion of events. The selection of balanced training samples from the four sets has been not possible because the sample sizes would



SMEs vs BR	1.000	SMEs vs BR	0.159	SMEs vs BR	0.006
SMEs vs A	0.630	SMEs vs A	0.080	SMEs vs A	0.034
SMEs vs MIEs	1.000	SMEs vs MIEs	0.072	SMEs vs MIEs	0.021

FIG. 3. ROC curves relative to the accuracy of the logit models on the SMEs. In the left panel the training sets have been randomly selected. The central panel refer to models built on equally sized and balanced training sets. The right panel displays the ROC curves obtain after applying ROSE. Below, the p values obtained by testing the difference between the areas under the curves. A Bonferron-Holms correction has been applied to take into account the multiple comparisons.

have been too small. Instead, we have selected samples with 25% of positive example and 75% of negative ones.

The central panel of Figure 3 suggests that our conjecture about the contradictory results was right. When using equally sized and more balanced training sets, the accuracy of the models downsizes, except for the model built on the SMEs. Hence, the effect of sample selection arises resulting in less accurate model estimated on samples different from the target population. However, in order to compare the models other things being equal, the sample size has been considerably reduced, thus leading to wonder if results may be considered reliable. In fact, besides the area under the ROC curve corresponding to the SMEs model is larger than the others, the difference is not significant (a Mann-Withney statistic based test has been used to test the difference between the areas under the ROC curves according to Hanley and McNeil (1982)).

4. Generating new events prior to sample selection evaluation

In the previous paragraph we have highlighted the risk of building models based on samples selected from a population different from the target one, but this problem is strongly related to the presence of a rare class. In fact, the skewer the class distribution is, the more hidden the sample selection effect is. Hence, applying a remedy against the problems caused by the class imbalance, prior to any deeper analysis of the sample selection effects, turns out to be necessary.

However, when the distribution of the classes is extremely skewed, even the most widespread

remedies for class imbalance have known drawbacks (McCarthy et al., 2005). One problem with the application of cost sensitive algorithms is that specific cost information is usually not available. Methods of undersampling may discard potentially useful data thus reducing the sample size (as seen in the previous section), while over-sampling may increase the likelihood of occurring overfitting, since it is bound to produce ties in the sample, especially as the sampling rate increases.

A recent strategy to cope with imbalanced learning has been proposed by Menardi and Torelli (2010). Aimed at balancing the distribution of the classes, the approach rests on the same idea of over/undersampling methods, but the imbalance is managed by generating new synthetic data to be used for training the classifier. This strategy may be referred to as ROSE (Random Over Sampling Examples).

Each sample point may be described by a vector $\mathbf{x} = (x_1, \ldots, x_p)$ of observed covariates and a label class y belonging to the set $\{Y_0, Y_1\}$, where Y_1 denotes the rare class and Y_0 the frequent one. In classification problems, learning methods are basically characterized by an implicit or explicit different approach to the estimation of the unknown conditional probabilities of belonging to the classes. Accordingly, the classification rule allocates a sample unit to the class Y_j, $j = 0, 1$ if the estimated conditional probability $Pr(y = Y_j|\mathbf{x})$ of belonging to that class exceeds a fixed threshold, that is

$$Pr(y = Y_j|\mathbf{x}) = \frac{f(\mathbf{x}|y = Y_j) \cdot Pr(Y_j)}{f(\mathbf{x})} > k, \quad 0 < k < 1 \tag{1}$$

with $Pr(Y_j)$ the probability of the class Y_j and f the probability density function of $\mathbf{x}$. Here, $\mathbf{x}$ is supposed to be continuous. Usually the threshold is $\frac{1}{2}$ but in a class imbalance framework it may be set differently. See for details, Prentice and Pyke (1979).

In order to obtain a balanced sample, ROSE aims at generating new examples from both the unknown $f(\mathbf{x}|Y_0)$ and $f(\mathbf{x}|Y_1)$. The generation of data from both the classes differentiates such approach from other existing remedies for class imbalance, allowing for creating a previously unobserved sample to be used for training the model. The originally observed data may be thereby employed to evaluate the model's accuracy.

Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be the training set used to perform the classification, with $\mathbf{x}_i$ the vector of covariates and $y_i \in \{Y_0, Y_1\}$, the label class. Without loss of generality, we may consider that $n_j < n$ is the size of the class Y_j and the first $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_r, y_{n_j})\}$ data belong to the rare class. The ROSE procedure for generating one example from class Y_j consists of two steps:

i. select $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_j}\}$, with probability $P(\mathbf{x}_i) = \frac{1}{n_j}$

ii. sample $\mathbf{x}$ from $K_{H_j}(\mathbf{x}_i)$, with K_{H_j} a probability distribution centered at $\mathbf{x}_i$ and H_j a matrix of scale parameters. $K_{H_j}(\mathbf{x}_i)$ is usually chosen in the set of the symmetric distributions (e.g. K is a Gaussian distribution) and it is an estimate of the local density of $\mathbf{x}_i$.

Essentially, we select an observed data belonging to one class and generate a new data in its neighborhood, where the width of the neighborhood is determined by H_j. It is worthwhile

to note that:

$$\begin{aligned}\hat{f}(\mathbf{x}|y=Y_j) &= \sum_{i=1}^{n_j} \hat{Pr}(\mathbf{x}_i)\hat{Pr}(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j}\hat{Pr}(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(\mathbf{x}-\mathbf{x}_i)\end{aligned}$$

which entails that the generation of events according to the ROSE procedure corresponds to the generation of data from the kernel density estimate of $f(\mathbf{x}|Y_j)$. This desirable property allows us to consider H_j as a smoothing matrix and to choose it proportional to the solution of one of the several methods of bandwidth selection proposed in the literature (for a review, see for example Wand and Jones (1995.)).

We have applied the ROSE strategy to our real data set in order to generate the attributes of new artificial firms belonging to the described subpopulation MIE, SME, BR, A. Given the new synthetic balanced sample, we have repeated the experiment of comparing the performance of the models in detecting the SME default event. Results are displayed in the right panel of Figure 3. While models trained on the samples selected from MIE, BR and A still report a poor accuracy, classification performed by using artificial samples drawn from the target population show significantly improved results, thus confirming our conjecture about the existence of a sample selection effect, hidden by the unmbalanced distribution of the classes.

5. Concluding remarks

In this work we have empirically analyzed how different criteria of sample selection may affect the accuracy of a classification model in presence of rare classes. An application of several classification techniques to simulated data has shown the risk of building models on non random samples, but it has also pointed out that this problem is strongly related to the eventual imbalance of the classes. In fact, the skewer is the class distribution, the less accurate are the models and the more hidden is the sample selection effect. An application to real data aimed at separating defaulter firms from non defaulter ones has confirmed the described behavior.

A recent solution to the class imbalance problem consisting in generating new synthetic examples from the rare class has been considered and its application to the real data set has determined an increased accuracy of the classification as well as the evidence of the sample selection effect.

A deeper insight to this framework will be the focus of future research. In particular, our purpose is to propose possible remedies after exploring how the sample selection effects vary with different degrees of dependence between the response variable and the variables according to which the sample selection is done and how the problem is related to the concern of model selection.

Acknowledgments.

This work has been conducted in the framework of the FIRB project "Modelli di data mining e knowledge management per le piccole e medie imprese", supported by MIUR, 2003

REFERENCES

E. Altman and G. Sabato. Effects of the new basel capital accord on bank capital requirements for smes. *Journal of Financial Services Research*, 28:15–42, 2005.

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 the Fourth International Conference on Knowledge Discovery and Data Mining*, 2001.

M. Dietsch and J. Petey. Should sme exposures be treated as retail or as corporate exposures? a comparative analysis of default probabilities and asset correlation in french and german smes. *Journal of Banking and Finance*, 28:773–788, 2004.

J. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:28–36, 1982.

N. Japkowicz. Learning from imbalanced data sets: a comparison of various strategies. Technical report, Papers from the AAAI Workshop on Learning from Imbalanced Data Sets. Tech. rep. WS-00-05, Menlo Park, CA: AAAI Press, 2000.

N. Japkowicz and S. Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis.*, 6:429–450, 2002.

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

K. McCarthy, B. Zabar, and G. Weiss. Does cost-sensitive learning beat sampling for classifying rare classes? In *UBDM Š05. New York, NY, USA: ACM Press*, pages 69–77, 2005.

G. Menardi and N. Torelli. Training and assessing classification rules with unbalanced data. Technical report, Submitted, 2010.

R.L. Prentice and R.. Pyke. Logistic disease incidence models and case control studies. *Biometrika*, 66:403–411, 1979.

E. Stanghellini. On statistical issues raised by the new capital accord. *Statistica Applicata*, 18:389–405, 2006.

L. C. Thomas, D. B. Edelman, and J. N. Crook. *Credit Scoring and its Applications*. Society for Industrial and Applied Mathematics, 2002.

A. Van den Bosch, T. Weijters, H. J. van den Herik, and W. Daelemans. When small disjuncts abound, try lazy learning: A case study. In *Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning.*, pages 109–118, 1997.

M. P. Wand and M. C. Jones. *Kernel smoothing*. Chapman and Hall, 1995.

G. M. Weiss. Mining with rarity a unifying framework. In *SIGKDD Explorations, Chicago*, number 6, pages 7–19, 2004.

K. Woods, C. Doss, K. Bowyer, J. Solka, C. Priebe, and P. Kegelmeyer. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 7: 1417–1436, 1993.