RESAMPLING BASED ON STATISTICAL PROPERTIES OF DATA SETS

Julia Bondarenko

Department of Economic and Social Sciences, Helmut-Schmidt University Hamburg
(University of the Federal Armed Forces Hamburg), Holstenhofweg 85, 22043 Hamburg, Germany

Keywords: Resampling, Classification algorithm C4.5, Uniform/(truncated) Normal distribution, Kurtosis, Chi-squared test, Kolmogorov-Smirnov test, Traffic injuries number.

Abstract: In imbalanced data sets, classes separated into majority (negative) and minority (positive) classes, are not approximately equally represented. That leads to impeding of accurate classification results. Well balanced data sets assume uniform distribution. The approach we present in the paper, is based on directed oversampling of minority class objects with simultaneous undersampling of majority class objects, to balance non-uniform data sets, and relies upon the certain statistical criteria. The resampling procedure is carried out for the daily traffic injuries data sets. The results obtained show the improving of rare cases (positive class objects) identification with accordance to several performance measures.

1 INTRODUCTION

Numerous machine learning classification methods currently give good performance in numerous practical problems, such as diagnosing medical problems, speech recognition, expert systems, robotic processing etc. A starting point for the present study was an investigation of road injuries number within the framework of the joint project with police departments of one of German federal states. The presented work focuses on temporal factors impact on daily traffic injuries number. The initial data sets of traffic injuries are imbalanced: daily injuries numbers are not approximately equally represented, that is, separated into majority (negative) and minority (positive) classes. As a result, minority class is poorly performed by classification (S. Ertekin and Giles, 2007), (S. Kotsiantis and Pintelas, 2006). But our aim is to detect efficiently the important rare cases in number of injured persons. In fact, the ability to predict periods of high incidence of road accidents, is really essential. Rebalancing the class distributions for the further classification, which includes over- and undersampling techniques, can be applied in order to solve this problem at the data level. We propose here a simple and general resampling procedure, improving a classification performance of daily road injuries number.

Oversampling method balances data set by increasing the number of minority class objects (examples). The simplest oversampling method - random oversampling - increases the minority class size by randomly replicating existing minority class examples (oversampling with replacement). This technique is attractive exactly due to its simplicity, but unfortunately, since random oversampling only replicates existing data, it does not add any actual information to the data set. The another approach is to oversample the positive minority class by creating new examples. SMOTE (Synthetic Minority Over-Sampling Technique, see (N. Chawla and Kegelmeyer, 2002)) is the most popular oversampling method here. In SMOTE minority classes are oversampled by generating ”synthetic” examples of minority class and adding them to the data set. As a result, the class distribution in the data set changes and probability of correctly classifying minority class increases. Other oversampling approaches were also proposed ((V. Garcha and Mollineda, 2008), (H. Han and Mao, 2005)). Undersampling approaches try to decrease the number of major class examples. However this method may involve information loss (that is, discard potentially important for learning and prediction examples, see (X.-Y. Liu and Zhou, 2006)).

Naturally, well balanced data sets assume uniform distribution. The approach we present in the paper, is based on directed oversampling of minority class objects with simultaneous undersampling of major-

Bondarenko J.
RESAMPLING BASED ON STATISTICAL PROPERTIES OF DATA SETS.
DOI: 10.5220/0002171701430148
In *Proceedings of the 6th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2009)*, page
ISBN: 978-989-8111-99-9
Copyright © 2009 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

ity class objects, to balance non-uniform data sets (to ensure the uniform distribution), and relies upon the statistical criteria. The remaining parts of the paper are organized as follows. In Section 2, we discuss the initial injuries data sets used in our work and perform their classification according to several metrics. Resampling procedure description and new classification results are presented in Section 3. Section 4 contains conclusions, brief discussion of ongoing and potential future research topics.

2 CLASSIFICATION OF INITIAL DATA SETS

2.1 Data Sets and Attributes

Data sets used in our study include the daily number of injured persons in traffic accidents in two cities of North Rhine-Westphalia, Germany - Duesseldorf and Duisburg, for the period 2004-2006. The daily data for each city represent an age group 25-59 years old people. Table 1 shows the mean, mode, median, minimum and maximum values for each region, as well as skewness and kurtosis, which are the measures of the data sets asymmetry and peakedness, respectively. Note, that we removed before all ”heavy” outliers from the data sets, as sensitive to outliers. ”Heavy” outliers are identified as all values more than $3IQR$ above the 3rd or below the 1st quartile, where $IQR = Q_3 - Q_1$ is interquartile range, Q_3 and Q_1 are 3rd and 1st quartiles correspondingly.

Table 1: Descriptive Statistics of Data Sets.

value	Duesseldorf	Duisbirg
mean	5.03	2.43
min	0	0
max	19	12
mode	4	1
skewness	0.65	0.92
kurtosis	3.67	4.02

Histograms showing distribution of injured persons for each urban region for period 2004-2006, are presented in Figure 1.

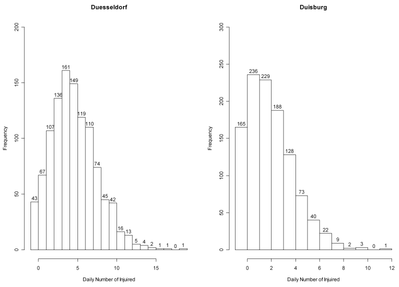

Figure 1: Histograms of Daily Number of Traffic Injured Persons for Age Range 25-59, 01.01.2004-31.12.2006.

Let $Y_{i,t}$ be the number of injured at the day t, $t = 1,...,1096$ (3 years, 2004-2006) in region i, $i \in I = \{1,2\}$ (1=Duesseldof, 2=Duisburg). We treat as attributes the certain day features related to corresponding outcomes $Y_{i,t}$, namely:

- year: 1=2004, 2=2005, 3=2006;
- month: 1=January, 2=February,..., 12=December;
- school vocation day: 1=yes or 0=no;
- holiday (official): 1=yes or 0=no;
- weekday: 1=Monday, 2=Tuesday,..., 7=Sunday;
- bridge day: 1=yes or 0=no.

The distribution for Region 1 looks roughly symmetric and just lightly (positively) skewed, data set for Region 2 shows more skewed behavior. Each data set has relatively narrow range of values - $0 \div 19$ for Region 1 and $0 \div 12$ for Region 2.

A number of algorithms has been developed for decision tree construction, but we will dwell upon one of them - C4.5-algorithm (Quinlan, 1993), based on the computing the metrics known as the information gain (IG) and gain ratio (GR). We introduce a notion of IG as follows:

$$IG(S_n,\, A) = \mathcal{E}(S_n) - \sum_{i \in values(A)} \frac{l_i}{n} \mathcal{E}(S_{A_i}) \;, \qquad (1)$$

where $\mathcal{E}(S_n) = -\sum_k \frac{n_k}{n} \log\left(\frac{n_k}{n}\right)$ is entropy of the entire data set $S_n = (Y_1,...,Y_n)$ of size n, n_k is the number of instances in S_n with value k, $\mathcal{E}(S_{A_i})$ is entropy of the sample S_{A_i} of size l_i involving elements from $\mathbf{Y}$, which correspond to the outcome (value) A_i of the attribute (feature) A, and $\frac{l_i}{n}$ represents the fraction of the data in S_n that goes into S_{A_i}.

In C4.5 algorithm below we adopt the information gain ratio to select the best day attribute to branch on at each stage. Attribute with the highest gain ratio gives us the crucial information concerning the temporal distribution of traffic injuries number inside the each region.

The formula aggregates over the different values A_i attribute A can have. But IG would be biased towards selecting attributes with more values. To mitigate this effect, we use a normalized version of IG - Gain Ratio (GR), defined as follows:

$$GR(S_n,\, A) = \frac{IG(S_n,\, A)}{Split\; Info(S_n,\, A)} \;, \qquad (2)$$

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

where split information $Split\ Info(S_n, A) = \sum_{i \in values(A)} \frac{l_i}{n} \ln\left(\frac{l_i}{n}\right)$ is the entropy of partitioning, or in other words, entropy associated with the distribution of the attribute A, where $\frac{l_i}{n}$ is the probability (proportion) of observing the ith value of A. Thus, a large number of small partitions is penalized there.

2.2 Evaluation Measures

In the first place, we employ the Accuracy traditional metric, that is the percentage of the correctly classified data. An information about actual and predicted examples is contained in confusion matrix (Kohavi and Provost, 1998). The entries in the confusion matrix have the following meaning in our problem:
TN is the number of correct predictions that an example is from negative class;
TP is the number of correct predictions that an example is from positive class;
FN is the number of incorrect predictions that an example is from negative class;
FP is the number of incorrect predictions that an example is from positive class.
The Accuracy is computed then as a proportion: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.

But for classification of imbalanced data sets, accuracy is no longer a proper measure since minority class has very small impact on the accuracy. Therefore, the following alternative evaluation measures (metrics) were proposed and used, for instance, in (G. Cohen and Geissbuhler, 2005), (Hido and Kashima, 2008), (Kubat and Matwin, 1997):

- True Positive Rate (TPR), or Recall. The proportion between correctly classified positive examples and that are calculated: Recall= $TPR = \frac{TP}{TP+FN}$.

- Precision. The proportion between correctly classified positive examples and that are actually correct: Precision= $\frac{TP}{TP+FP}$.

- G-mean (Geometric mean). Tries to maximize the accuracy on each of the two classes while keeping these accuracies balanced: $G - mean = \sqrt{Positive\ Accuracy * Negative\ Accuracy}$, where $Positive\ Accuracy = \frac{TP}{TP+FN}$ =Recall, $Negative\ Accuracy = \frac{TN}{TN+FP}$.

- F-measure. "The trade-off" between precision and recall, drops rapidly if either precision or recall is poor: $FM = \frac{2\text{Recall}*\text{Precision}}{Recall+Precision}$.

2.3 Classification Results

In Section 2.1 we have noted that each data set S is kept within a certain limited relatively narrow range of values. That allows us to consider every value of the range ν_i, where $v = (v_i)^T = (\min(S), ..., \max(S))^T$, as a separate class:

Class 1: "0 Injured Persons per Day", Class 2: "1 Injured Person per Day",..., Class r "r Injured Persons per Day", where $r = 19 + 1 = 20$ for Region 1 and $r = 12 + 1 = 13$ for Region 2. Thus, we transform here count data into categorical ones. Of course, we could group the data or consider continuous data as well.

We make no assumption about the distribution of the daily number of persons injured in traffic accidents, except its unimodality during the resampling procedure. The unimodality of initial data sets is also established in the frequency distribution histograms, see Fig. 1. Remind, that a distribution is called unimodal if there is only one major "peak" (mode) in the distribution. Let M be mode of the data set, and φ_M - its frequency. We select all values of daily number of injured persons ν, those frequencies are lying within the interval from 0.8 φ_M till φ_M, to be "negative", or majority, class S^{neg}. All the other values combined together represent "positive", or minority, class S^{pos}.

The classification procedure with algorithm C4.5 gives the following results for both regions, which we present below. The classification is performed by means function J48() implemented in RWeka package, R statistical software. The minimal number of instances per leaf we set to 20. The evaluation measures defined in previous subsection, are computed either.

The classification results for both regions are reported below. Only about 18% of examples are classified correctly for Region 1, with low magnitudes of alternative measures.

```
CLASSIFICATION RESULTS FOR REGION 1:

Correctly Classified Instances 18.3394 %,

Recall = 0.1325, Precision = 0.0676,
G-mean = 0.1650, F-measure = 0.0895.
```

Majority (negative) class is presented here by numbers 3, 4 and 5. All the others are classified into the minority class. For the Region 2 we have at the beginning about 26% of correctly classified examples

```
CLASSIFICATION RESULTS FOR REGION 2:

Correctly Classified Instances 26.2774 %

Recall = 0.2449, Precision = 0.1347,
G-mean = 0.2577, F-measure = 0.1738.
```

Consistently, negative class consists of numbers 1, 2 and 3, the rest of the numbers is in positive class. In the next Section we describe a resampling strategy, which improves classification performance.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

3 NEW RESAMPLING PROCEDURE

3.1 Resampling Motivation and Procedure Illustration

As one can see, in both our data sets, the examples from minority class are much less ”beloved” by classification algorithm than from majority one (common problem of imbalanced data). Below we present an algorithm, which resizes/rebalances our data sets. With accordance to certain criteria, we generate artificial data from minority class and simultaneously withdraw data from majority class, until the classes are approximately equally represented (data are uniformly distributed). It may be considered as a preprocessing procedure for further classification and prediction.

The algorithm assumes unimodal character for frequency distribution of classes, without reference to skewness (asymmetry) of the distribution. Majority classes are concentrated around peak M (Figure 2a), while minority classes are more tails-sited. Our suggestion is to generate new synthetic data from singly left- and right-truncated normal (half-normal) distributions, with truncation points $u_{left} = \min(S)$ and $u_{right} = \max(S)$, respectively, where u_{left} and u_{right} are also the means of truncated normal distributions. At the same time, we flatten the peak of distribution, picking out the data placed around it. Such combination of the oversampling the minority class with undersampling the majority class helps to achieve better classifier performance.

If more detailed, a random variable is said to be from a left-truncated normal distribution if its density is

$$f(x) = \frac{q}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

$$\text{for } x \geq u_{left},\ f(x) = 0 \text{ for } x < u_{left}\ , \qquad (3)$$

where m is a mean and σ is a standard deviation of the distribution, q is a normalizing quantity with value obtained from the equation $\int_{u_{left}}^{\infty} f(x)\,dx = 1$. A right-truncated normal distribution is defined analogously. We discard elements outside the limit points u_{left} and u_{right}, and choose $m_{left} = u_{left}$ and $m_{right} = u_{right}$. This guarantees, that the least popular classes lying at the ends of data range interval, will be treated most intensive. Choosing the desired level of significance α, we put the critical points (confidence intervals endpoints with $100(1-\alpha)\%$ -confidence level) $C_{left}^{1-\alpha}$ and C_{right}^{α} for both truncated distributions equal to M. Thus, the majority class will be also maintained, as artificially generated data can be also out of endpoints $C_{left}^{1-\alpha}$ and C_{right}^{α}. The standard deviations of the truncated normal distributions can be obtained as $\sigma_{left} = \frac{M-m_{left}}{C^{1-\alpha}}$ and $\sigma_{right} = \frac{m_{right}-M}{C^{\alpha}}$, where C^{α}, $C^{1-\alpha}$ are critical values of truncated at 0 standard normal distribution. For example, for $\alpha = 0.05$: $C^{1-\alpha} = -C^{\alpha} = 1.959964$. This is shown schematically in Figure 2b.

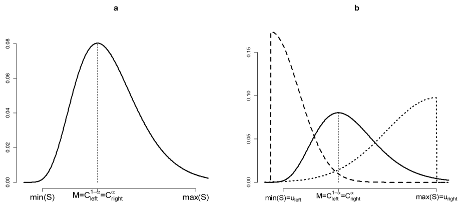

Figure 2: Scheme of Resampling Procedure: a - Underlying Unimodal Distribution; b - Underlying Distribution (solid line) and Two Truncated Normal Distributions (dashed and dotted lines).

As we have noted above, simultaneously with adding new ”artificial” observation to positive class, we drop randomly observation from negative class. Therefore, we keep the sample size fixed. That helps us to avoid increasing computational time and losing information.

In the procedure, we will use the following criteria:

- Chi-squared test (Pearson goodness-of-fit test). We have data set S_n that is grouped into discrete classes. Let φ denote a vector of observed frequencies of classes and let φ^0 denote the corresponding vector of expected (specified) frequencies. We then calculate our test statistic:

$$T = \sum_{i=1}^{r} \frac{\left(\varphi_i - \varphi_i^0\right)^2}{\varphi_i}\ , \qquad (4)$$

where r is a number of classes. Under the null hypothesis, this statistic is chi-squared distributed, with degrees of freedom equal to $r-1-j$, where j is a number of parameters that should be estimated (for parametric case). That is, we can test here a null hypothesis that the frequencies of observed outcomes (daily numbers of road injuries) follow a specified (uniform) distribution, at some chosen level γ. Unfortunately, this test can be also unsuitable for samples of considerably large size ($>$ 3000). It cannot be also applied when the expected frequency of any cell is less than 5 or the total n is less than 50. But in our case this test is appropriate, so further we apply it as a primary criterion.

- One-sample Kolmogorov-Smirnov nonparametric test (KS-test). Let $S_n = \{Y_1, Y_2, ..., Y_n\}$ be as before our data set, with distribution function $F(y)$. We wish to test the null hypothesis H_0: $F(y) = F_0(y)$ for all y against the alternative H_1: $F(y) \neq F_0(y)$, where $F_0(y)$ is a completely specified distribution function (in our case - function of uniform distribution). Test on H_0 vs H_1 is determined by Kolmogorov-Smirnov statistic

$$D_n = \sup_{-\infty<y<\infty} |F_n(y) - F_0(y)| \, , \qquad (5)$$

where $F_n(y)$ is the empirical distribution function defined by $F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i < y\}$. That is, the Kolmogorov-Smirnov test tries to determine if distribution of our data set differs significantly from the specified hypothetical distribution (here - the uniform hypothetical distribution). The null hypothesis is rejected at level γ if the computed value D_n is larger than the critical value C_γ. The critical values of Kolmogorov-Smirnov test statistic depend on the sample size. For large samples (≥ 40) one uses asymptotic critical values, which are strictly decreasing functions of the sample size: for example, the critical value at the $\gamma = 0.05$ level is approximately $\frac{1.36}{\sqrt{n}}$, where n is sample size. Obtaining the value $D_n < C_\gamma$ for considerably large n (> 3000) may dramatically increase computational time. Another problem is that Kolmogorov-Smirnov test doesn't work well with discrete (count) data. This problem could be solved, for example, by the following way: we can transform our data to continuous ones by means of Monte Carlo simulation, and then apply Kolmogorov-Smirnov test to the empirical distribution function of continuous simulated data and the specified distribution $F_0(y)$.

- As we constrain a uniform distribution, one can use a value of kurtosis for distribution control. Remind, that kurtosis is the degree of peakedness of a distribution. Removing observations from the center to the tails and "shoulders" of the distribution will decrease kurtosis, making the initial leptokurtic distributions more platykurtic. A uniform distribution has a kurtosis of 1.8. Thus, kurtosis could be used. as a simple parameter for comparison data set distribution with uniform ones: if

$$|\mu(S_n) - \mu_0| < \varepsilon \, , \qquad (6)$$

where $\mu_0 = 1.8$, $\mu(S_n)$ is kurtosis of S_n, and $0 < \varepsilon < 1$, we stop resampling procedure.

3.2 Classification Results

We can now see that the C4.5 algorithm improves its classification performance when we apply it to rebalanced data sets. In our example, we have chosen $\gamma = \alpha = 5\%$. The percentage of correctly classified instances after resampling procedure have increased from 18% to 35% for Region 1, from 26% to 37% - for Region 2 (see below). Our approach also yields promising results in terms of the alternative performance measures.

```
CLASSIFICATION RESULTS FOR REGION 1:
Correctly Classified Instances 35.0365 %,
Recall = 0.3089, Precision = 0.5783,
G-mean = 0.3734, F-measure = 0.4027.

CLASSIFICATION RESULTS FOR REGION 2:
Correctly Classified Instances 37.5 %,
Recall = 0.3477, Precision = 0.6476,
G-mean = 0.3972, F-measure = 0.4525.
```

The histograms indicate approximately uniform distributions after procedure completion, with kurtosis magnitudes equal to 1.8643 and 1.7894, respectively.

Reporting our experimental results in dynamics, one can see in Figures 3-5, that the measures those we are interested in, are increasing (although non-monotonically) with the iteration number. In each figure, we plot every classification measure for both regions: percentage of correctly classified examples (Figure 3), Recall (Figure 4), Precision (Figure 5) (G-mean and F-measure plots are not presented here by lack of space).

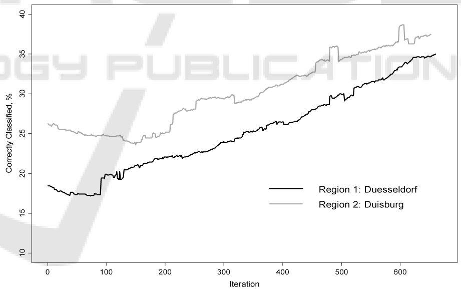

Figure 3: Algorithm Performance for Regions: Accuracy, %.

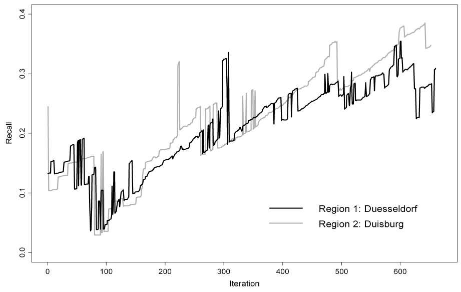

Figure 4: Algorithm Performance for Regions: Recall.

Note, that Precision lines are lying almost everywhere above all other lines. This tells us that proportion of

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

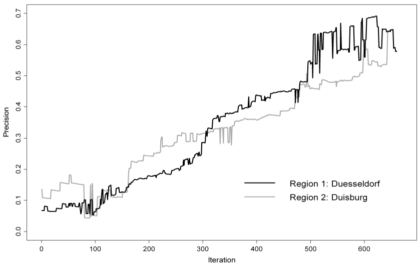

Figure 5: Algorithm Performance for Regions: Precision.

examples that were classified as elements from positive classes and those that are actually positive, grows most rapidly.

4 CONCLUSIONS AND FUTURE RESEARCH WORK

In this paper, a resampling technique based on statistical properties of data set, was proposed. We have tested our technique in terms of its accuracy and four performance measures: Recall, Precision, G-mean and F-measure. As investigation reveals, C4.5 algorithm applied to resampled data sets produced better results. But, in spite of the presented promising direction of rather general resampling techniques, the algorithm has to be yet improved in terms of classification performance. The effect of its application to various forms of data sets structure (highly skewed data sets, multimodal data sets, etc.) should be investigated as well. The comparison with other resampling methods also has to be carried out.

The resampling algorithm can be also carried out on the basis of the likelihood ratio test. The Neyman-Pearson Lemma implies that likelihood ratio test gives the best result in fixed size samples.

Further, for the start-up problem we were interested in, an accurate classification can result in injuries control boundaries analogous to presented in (Bondarenko, 2006a), (Bondarenko, 2006b), (F. Pokropp and Sever, 2006). The trees obtained by classification, can be very large (a lot of nodes and leaves), and in this since they are less comprehensible for control boundaries illustration. But we can simplify the obtained classification results by transforming every decision tree into a set of ”if-then” rules (”Traffic Injuries Rules”), which seem to be easier for understanding and interpreting. Using real traffic injuries data, it is possible to develop realistic model for daily injuries number prediction, depending on temporal factors (year, month, day type). Of course, this research direction is open for other practical implications as well.

REFERENCES

Bondarenko, J. (2006a). Analysis of traffic injuries among children based on generalized linear model with a latent process in the mean. *Discussion Paper in Statistics and Quantitative Economics, Helmut-Schmidt University Hamburg*, (116).

Bondarenko, J. (2006b). Children traffic accidents models: Analysis and comparison. *Discussion Paper in Statistics and Quantitative Economics, Helmut-Schmidt University Hamburg*, (117).

F. Pokropp, W. Seidel, A. B. M. H. and Sever, K. (2006). Control charts for the number of children injured in traffic accidents. In H.-J. Lenz, P.-T. W., editor, *Frontiers in Statistical Quality Control*, pages 151–171. Physica, Heidelberg, 5 edition.

G. Cohen, M. Hilario, H. S. S. H. and Geissbuhler, A. (2005). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37:7–18.

H. Han, W. W. and Mao, B. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing*, pages 878–887.

Hido, S. and Kashima, H. (2008). Roughly balanced bagging for imbalanced data. In *Proceedings of the SIAM International Conference on Data Mining*, pages 143–152.

Kohavi, R. and Provost, F. (1998). Glossary of terms. editorial for the special issue on applications of machine learning and the knowledge discovery process. *Machine Learning*, 30:271–274.

Kubat, M. and Matwin, S. (1997). Adressing the curse of imbalanced training sets: Onesided selection. In *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186.

N. Chawla, K. W. Bowyer, L. O. H. and Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California.

S. Ertekin, J. H. and Giles, C. L. (2007). Active learning for class imbalance problem. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 823–824.

S. Kotsiantis, D. K. and Pintelas, P. (2006). Handling imbalanced datasets: a review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36.

V. Garcha, J. S. and Mollineda, R. (2008). On the use of surrounding neighbors for synthetic over-sampling of the minority class. In *Proceedings of the 8th WSEAS International Conference on Simulation, Modelling and Optimization*, pages 389–394.

X.-Y. Liu, J. W. and Zhou, Z.-H. (2006). Exploratory undersampling for class-imbalance learning. In *Proceedings of the International Conference on Data Mining*, pages 965–969.