The Novelty Detection Approach for Different Degrees of Class Imbalance

Hyoung-joo Lee and Sungzoon Cho*

Seoul National University, San 56-1, Shillim-dong, Kwanak-gu, 151-744, Seoul, Korea
imhjlee@gmail.com, zoon@snu.ac.kr

Abstract. We show that the novelty detection approach is a viable solution to the class imbalance and examine which approach is suitable for different degrees of imbalance. In experiments using SVM-based classifiers, when the imbalance is extreme, novelty detectors are more accurate than balanced and unbalanced binary classifiers. However, with a relatively moderate imbalance, balanced binary classifiers should be employed. In addition, novelty detectors are more effective when the classes have a non-symmetrical class relationship.

1 Introduction

The class imbalance refers to a situation where one class is heavily underrepresented compared to the other class in a classification problem [1]. Dealing with the class imbalance is of importance since it is not only very prevalent in various domains of problems but also a major cause for performance deterioration [2]. When one constructs a binary classifier with an imbalanced training dataset, the classifier produces lopsided outputs to the majority class. In other words, it classifies far more patterns to belong to the majority class than it should. Real world examples include fault detection in a machine, fraud detection, response modeling, and so on.

A vast number of approaches have been proposed to deal with the class imbalance [1,2,3,4,5,6]. The most popular methods try to balance the dataset with under-/over-sampling, and cost modification. A balanced binary classifier is constructed using one of the balancing methods, while a classifier is called unbalanced when no balancing method is implemented. On the other hand, the drastic solution of totally ignoring one class during training can work well for some imbalanced problems [7,8,9,10]. This approach is called novelty detection or one-class classification [11,12] where the majority class is designated as normal while the minority class as novel. A classifier learns the characteristics of the normal patterns in training data and detects novel patterns that are different from the normal ones. Geometrically speaking, a novelty detector generates a closed boundary around the normal patterns [13]. Although a novelty detector usually learns only one class, it can also learn two classes. It has been empirically shown that a novelty detector trained with a few novel patterns as well can generate a more accurate and tighter boundary [9,12].

* Corresponding author.

I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 21–30, 2006.

In this paper, we show that the novelty detection approach is a viable solution to the class imbalance. In particular, two types of novelty detectors, 1-SVM trained only with one class [13] and 1-SVM trained with two classes (1-SVM$_2$) [14], are compared with balanced and unbalanced SVMs. In order to investigate which approach is suitable for different degrees of class imbalance, experiments are conducted on artificial and real-world problems with varying degrees of imbalance. In the end, we examine the following conjectures:

(a) Novelty detectors are suitable for an extreme imbalance while balanced binary classifiers are suitable for a relatively moderate imbalance.
(b) A problem is called symmetrical when each class originally consists of homogeneous patterns and a classifier discriminates two classes, e.g. apples and oranges, or males and females. A problem is called non-symmetrical, when only one class is of interest and everything else belongs to another class. A classifier distinguishes one class from all other classes, e.g. apples from all other fruits. Novelty detectors are more suitable for datasets with non-symmetrical class relationships than with symmetrical relationships.
(c) As the class imbalance diminishes, a novelty detector trained with two classes improves more, compared to one trained with one class.

The following section briefly reviews the support vector-based classifiers used in this paper and Section 3 presents the experimental results. Conclusion and some remarks are given in Section 4.

2 Support Vector-Based Classifiers

Suppose a dataset $\mathbf{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{x}_i$ is a d-dimensional input pattern and y_i is its class label. Let us define the majority and the minority classes as $\mathbf{X}^+ = \{\mathbf{x}_i | y_i = +1\}$ and $\mathbf{X}^- = \{\mathbf{x}_i | y_i = -1\}$, respectively. In an imbalanced dataset, $N^+ \gg N^-$ where N^+ and N^- are the numbers of patterns in $\mathbf{X}^+$ and $\mathbf{X}^-$, respectively. We employ unbalanced SVM, balanced SVMs, 1-SVM, and 1-SVM$_2$ as listed in Table 1.

Table 1. Classifiers used: SVM indicates the standard two-class SVM. SVM-U, SVM-O, and SVM-C are balanced SVMs using under-sampling, over-sampling, and cost modification, respectively. 1-SVM and 1-SVM2 indicate one-class SVMs trained with one class and with two classes, respectively.

Unbalanced binary classifier	Balanced binary classifiers	Novelty detector with one class	Novelty detector with two classes
SVM	SVM-U SVM-O SVM-C	1-SVM	1-SVM$_2$

2.1 Support Vector Machine (SVM)

SVM finds a hyperplane that separates two classes with a maximal margin in a feature space [15]. An optimization problem can be considered:

$$\begin{aligned} \min \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i, \\ \text{s.t} \ & y_i(\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \cdots, N, \end{aligned} \tag{1}$$

where $C \in (0, \infty]$ is the cost coefficient which controls the trade-off between the margin and the training error. The solution can be obtained by the quadratic programming techniques. In the optimal solution, only a small number of α_i's are positive where α_i's are the Lagrangian multipliers related to the training patterns. Those patterns for which α_i's are positive are called support vectors and the subset of support vectors is denoted as SV. The SVM decision function for a test pattern $\mathbf{x}$ is computed as

$$f(\mathbf{x}) = \text{sign}\Big[\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) + b\Big] = \text{sign}\Bigg[\sum_{\mathbf{x}_i \in \text{SV}} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\Bigg]. \tag{2}$$

2.2 Balancing with SVM

In an imbalanced problem, a typical binary classifier predict most or even all patterns to belong to the majority class [1]. Although the classification accuracy may be very high, this is not what we are interested in. We would like to construct a classifier which identifies both classes. Therefore, a balanced classifier is preferred although its accuracy may be lower than an unbalanced one. Various balancing methods have been proposed [1,2,3,4,5,6]. We apply a few of the simplest methods: under-sampling, over-sampling, and cost modification.

With under-sampling [1], N^- patterns are randomly sampled from X^+ to equate the numbers of patterns in two classes. With over-sampling [5], patterns from X^- are randomly sampled N^+ times with replacement. The two sampling methods are the most simple and the most popular. However, under-sampling may discard important information from the majority class. Over-sampling do not make additional information while increasing the number of patterns significantly. In this paper, SVM-U and SVM-O denote SVM classifiers using the under- and over-sampling methods, respectively.

For SVM, the cost modification method [3,6] is readily applicable by assigning a smaller cost to the majority class and a larger cost to the minority class to assure that the minority class is not ignored. One way to accomplish it is to modify the objective function in (1) as follows,

$$\min \ \frac{1}{2}\|\mathbf{w}\|^2 + C^+ \sum_{\mathbf{x}_i \in \mathbf{X}^+} \xi_i + C^- \sum_{\mathbf{x}_i \in \mathbf{X}^-} \xi_i, \tag{3}$$

where $C^+ = \frac{N^-}{N}C$ and $C^- = \frac{N^+}{N}C$. The classifier obtained by solving (3) is denoted as SVM-C. SVM-C may lead to seriously biased results since the costs assigned entirely based on the numbers of patterns can be incorrect.

2.3 Support Vector Machine for Novelty Detection

1-SVM [13] finds a function that returns +1 for a small region containing training data and −1 for all other regions. A hyperplane $\mathbf{w}$ is defined to separate a fraction of patterns from the origin with a maximal margin in a feature space. The conventional 1-SVM performs a kind of unsupervised learning, learning only the majority class and not considering the class labels. Thus an optimization problem can be considered as follows,

$$\begin{aligned} \min \ & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu N^+} \sum_{\mathbf{x}_i \in \mathbf{X}^+} \xi_i, \\ \text{s.t } \ & \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_i) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall \mathbf{x}_i \in \mathbf{X}^+. \end{aligned} \tag{4}$$

where $\nu \in (0, 1]$ is a cost coefficient.

One can construct 1-SVM$_2$ [14] by incorporating patterns from the minority class into (4) as follows,

$$\begin{aligned} \min \ & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu N} \sum_{i} \xi_i, \\ \text{s.t } \ & y_i(\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_i) - \rho) \geq \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \cdots, N. \end{aligned} \tag{5}$$

Note that this is not for binary classification. The objective function is not to separate two classes but to separate the majority patterns from the origin while keeping the errors as small as possible. The solutions of (4-5) can be obtained analogously to SVM.

3 Experimental Results

The classifiers were applied to ten artificial and 24 real-world problems. For each training dataset, the degree of class imbalance varied with the fractions of the minority class being 1, 3, 5, 7, 10, 20, 30, and 40%. Each classifier was constructed based on a training dataset and evaluated on a test set which has a relatively balanced class distribution. Ten different training and test sets were randomly sampled for each problem to reduce a sampling bias.

To train the SV-based classifiers, two hyper-parameters have to be specified in advance, the RBF kernel width, σ, and the cost coefficient, C or ν. For each problem, we chose the best parameters on a hold-out dataset which has an equal number of patterns from the two classes.

3.1 Artificial Datasets

We generated five types of majority classes which reflect features such as scaling, clustering, convexity, and multi-modality. For each distribution, two types

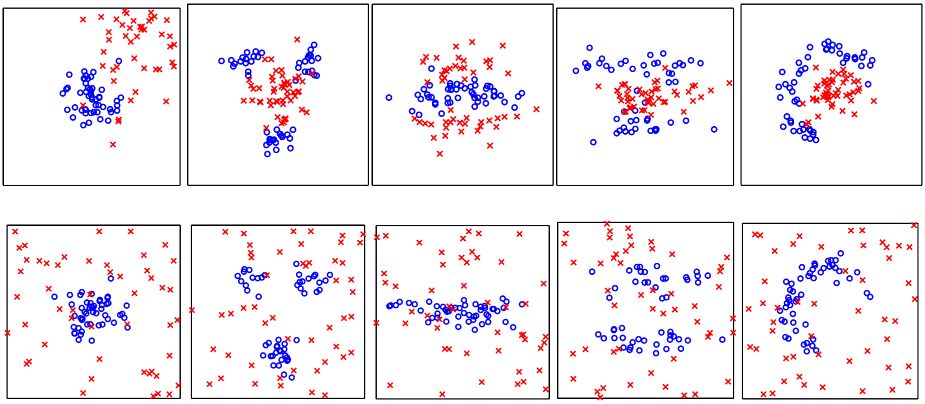
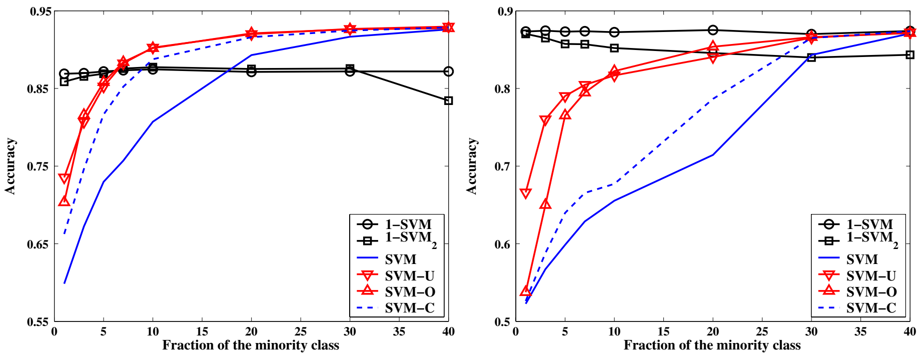


Fig. 1. The artificial datasets: The first and second rows correspond to symmetrical and non-symmetrical cases, respectively. The columns correspond to “Gauss”, “Gauss3”, “Ellipse”, “Ellipse2”, and “Horseshoe” from left to right. The circles and the crosses represent patterns from the majority and the minority classes, respectively.

of minority classes were generated. One has a multivariate Gaussian distribution while the other has the uniform distribution over the whole input space. The former corresponds to the symmetrical case while the latter to the non-symmetrical case. Thus, ten (= 5 × 2) artificial datasets were generated as shown in Fig. 1. For each dataset, 200 and 1,000 patterns were sampled from the majority class for training and test, respectively, and 1,000 patterns were sampled from the minority class for test.

Fig. 2(a) shows the average accuracies over the five symmetrical artificial datasets. When the fraction of the minority class is 5% or lower, 1-SVM and 1-SVM$_2$ are superior to the binary classifiers, balanced or not. Then, balanced classifiers, especially SVM-U and SVM-O, improved and came ahead of them as the fraction of the minority class increases. 1-SVM is generally slightly better than 1-SVM$_2$. The average accuracies over the non-symmetrical datasets are shown in Fig. 2(b). Novelty detectors are even better than in Fig. 2(a). In particular, 1-SVM is the best classifier or tied for the best for all the fractions. Unexpectedly, 1-SVM$_2$ gets gradually worse as the fraction of the minority class increases. Novelty detection is more effective for non-symmetrical datasets than for symmetrical ones. Considering that 1-SVM is better than 1-SVM$_2$, utilizing two classes does not necessarily lead to better results. As expected, unbalanced SVM did not work well and performed worst in both cases, although it caught up with the others as the fraction increased.

Fig. 3 shows examples of decision boundaries with 10% of patterns from the minority class. For the symmetrical dataset, every classifier generated a reasonable boundary. The boundaries by the binary classifiers resembled the “optimal” one. While the boundaries by the novelty detectors were different from the optimal one, they could effectively discriminate the two classes. On the other hand, for the non-symmetrical dataset, the binary classifiers failed to generate good



(a) Symmetrical (b) Non-symmetrical

Fig. 2. The average accuracies for the artificial datasets

decision boundaries. SVM generated boundaries that will classify too large a region as the majority class. Remember that crosses can appear anywhere in the 2D space. SVM-U did its best given the dataset, but generated a boundary that was much different from the optimal one because too many patterns from the majority class were discarded. Another drawback of SVM-U is its instability. A boundary in one trial was very different from a boundary in another. Note that we present the best looking boundary in our experiments. SVM-O and SVM-C performed poorly since the patterns from the minority class were too scarce to balance the imbalance. The novelty detectors generated boundaries similar to the optimal one, though the boundaries by 1-SVM and 1-SVM$_2$ were not exactly identical.

3.2 Real-World Datasets

A total of 21 real-world datasets were selected from UCI machine learning repository[1], Data Mining Institute (DMI)[2], Rätsch's benchmark repository[3], and Tax[4] as listed in Table 2. Digit and letter recognition problems are non-symmetrical since they were formulated to distinguish one class from all others. For the digit dataset, '1' and '3' were designated in turn as the majority classes and discriminated from all other digits, respectively. For the letter dataset, 'a', 'o', and 's' were designated in turn as the majority class. Also, the pump dataset is non-symmetrical since a small non-faulty region is to be recognized in the whole input space. Therefore, six non-symmetrical problems were formulated.

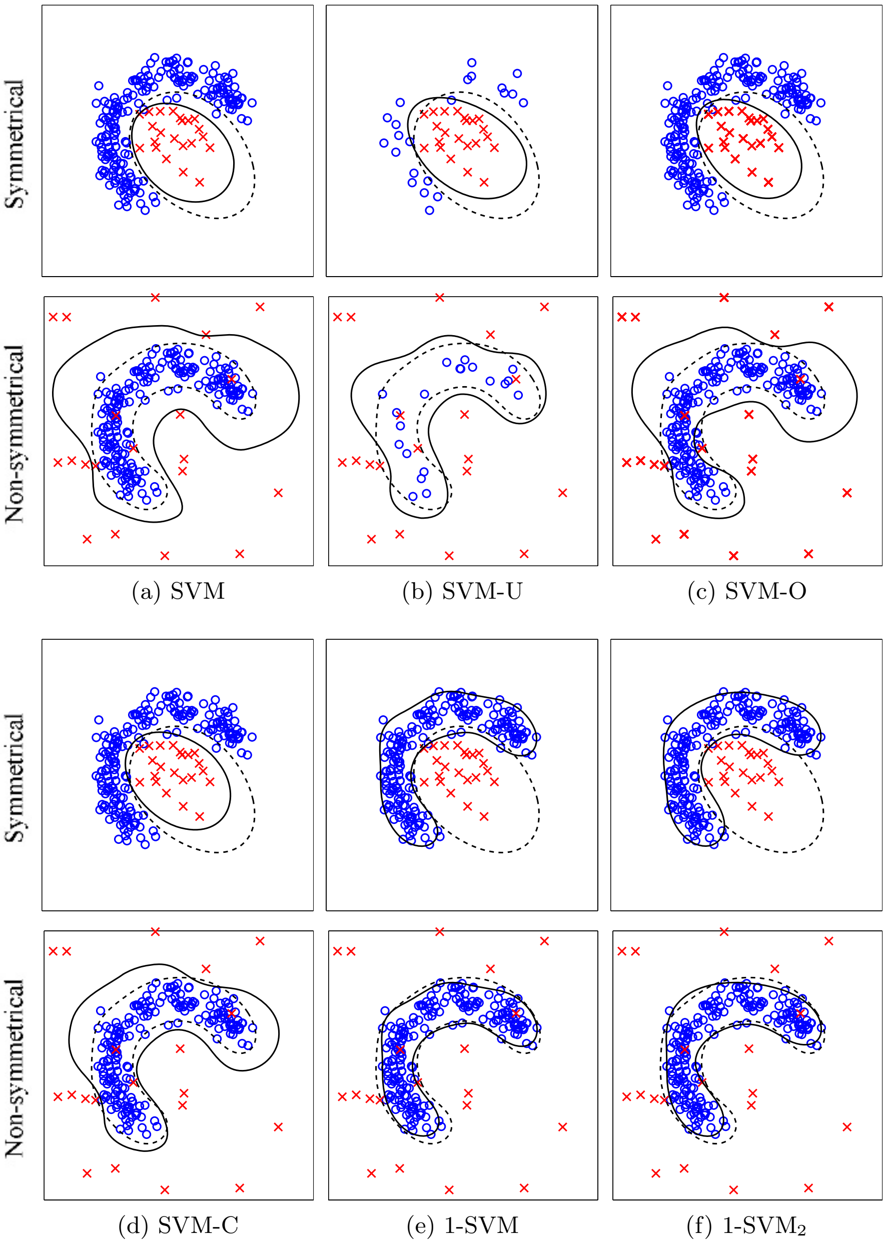
Fig. 4(a) shows the average accuracies over the 18 symmetrical real-world problems. The novelty detectors are better than the binary classifiers when the fraction is lower than 5%. Their accuracies remain still for all fractions while

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[2] `http://www.cs.wisc.edu/dmi/`.

[3] `http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm`.

[4] Pump vibration datasets for fault detection used in [12]. Personal communication.



(a) SVM (b) SVM-U (c) SVM-O

(d) SVM-C (e) 1-SVM (f) 1-SVM$_2$

Fig. 3. Decision boundaries for the horseshoe dataset: Six classifiers were trained with 100 circles and ten crosses. The solid boundaries were generated by the classifiers while the broken ones are the "optimal" ones.

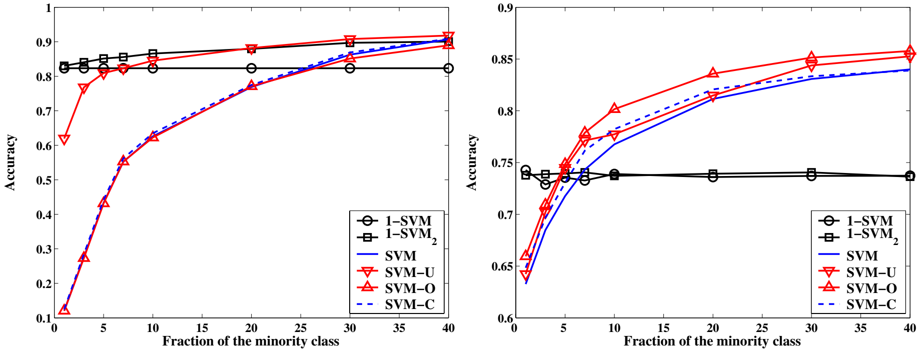
Table 2. Real-world datasets: 18 of 24 have symmetrical class distributions while three have non-symmetrical distributions

Symmetrical classes

Dataset	Source	Dataset	Source	Dataset	Source
banana	Rätsch	breast-cancer	Rätsch	bright	DMI
bupa	Rätsch	check	DMI	diabetes	Rätsch
dim	DMI	german	Rätsch	heart	Rätsch
housing	DMI	image	Rätsch	ionosphere	UCI
mush	DMI	thyroid	Rätsch	titanic	Rätsch
twonorm	Rätsch	vehicle	UCI	waveform	Rätsch

Non-symmetrical classes

Dataset	Source	Dataset	Source	Dataset	Source
digit	UCI	letter	UCI	pump	Tax



(a) Symmetrical classes (b) Non-symmetrical classes

Fig. 4. The average accuracies for the real-world datasets

the accuracies of the binary classifiers increase steeply. When the fraction exceeds 5%, SVM-O is the best classifier. 1-SVM and 1-SVM$_2$ are equivalent to each other. Fig. 4(b) shows the average accuracies over the six non-symmetrical real-world problems. 1-SVM$_2$ is the best or tied for the best when the fraction is 20% or lower. 1-SVM$_2$ improves steadily as the fraction increases while the accuracy of 1-SVM changes little. 1-SVM is better than the binary classifiers until the fraction increases to 7%. Among the binary classifiers, SVM-U is the most accurate. The other classifiers show little difference in accuracy.

4 Conclusions and Discussion

In our experiments, the conjectures in Section 1 were investigated:

(a) With an extreme imbalance, e.g. with 5% or lower fraction of the minority class, novelty detectors are generally more accurate than binary classifiers.

On the other hand, with a moderate imbalance, e.g. with 20% or higher fraction of the minority class, balanced binary classifiers are more accurate than unbalanced binary classifier and novelty detectors. With a fraction of 5 to 20% of the minority class, the results are not conclusive.

(b) Novelty detectors perform better for the non-symmetrical problems than for the symmetrical ones, in comparison to binary classifiers. That is not surprising since solving a non-symmetrical problem is naturally fit for the novelty detection approach.

(c) The results are conflicting regarding the third conjecture. For the artificial datasets, 1-SVM$_2$ is no better than 1-SVM and its accuracy even decreases as the fraction of the minority class increases. For the real-world dataset, on the other hand, 1-SVM$_2$ is slightly better than 1-SVM. Its accuracy increases gradually for the non-symmetrical datasets. We speculate that learning only one class can be sufficient for a relatively noise-free dataset such as the artificial ones while learning two classes helps a novelty detector refine its boundary for a noisy dataset.

In summary, novelty detection approach should be considered as a candidate for imbalanced problems, especially when the imbalance is extreme. Balanced binary classifiers have comparable performances. So a balancing method should be chosen empirically depending on the problem at hand.

A few limitations have to be addressed. First, we only have considered degrees of class imbalance. There are many other factors to influence the class imbalance such as data fragmentation, complexity of data, data size to name a few [2,4]. The novelty detection approach needs to be analyzed with respect to them. Second, parameter selection was based on a balanced hold-out dataset. How to perform parameter selection with an imbalanced dataset demands further research. Third, we restricted our base classifiers to SVM in the experiments. Other families of algorithms such as neural networks and codebook-based methods need to be investigated as well.

Acknowledgement

The authors would like to thank David M.J. Tax for having provided us with the pump data and for valuable comments. This work was partially supported by grant No. R01-2005-000-103900-0 from Basic Research Program of the Korea Science and Engineering Foundation, Brain Korea 21, and Engineering Research Institute of SNU.

References

1. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One-sided Selection. In: Proceedings of 14th International Conference on Machine Learning (1997) 179-186
2. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis 6(5) (2002) 429-450

3. Elkan, C.: The Foundations of Cost-sensitive Learning. In: Proceedings of the Seventh International Joint Conference on Artificial Intelligence (2001) 973-978
4. Weiss, G.M.: Mining with Rarity: A Unifying Framework. SIGKDD Explorations 6(1) (2004) 7-19
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE : Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321-357
6. Shin, H.J., Cho, S.: Response Modeling with Support Vector Machines. Expert Systems with Applications 30(4) (2006) 746-760
7. He, C., Girolami, M., Ross, G.: Employing Optimized Combinations of One-class Classifiers for Automated Currency Validation. Pattern Recognition 37 (2004) 1085-1096
8. Japkowicz, N.: Concept-Learning in the Absence of Counter-Examples: An Autoassociation-based Approach to Classification. PhD thesis. Rutgers University, New Jersey (1999)
9. Lee, H., Cho, S.:. SOM-based Novelty Detection Using Novel Data. In: Proceedings of Sixth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), Lecture Notes in Computer Science 3578 (2005) 359-366
10. Raskutti, B., Kowalczyk, A.: Extreme Re-balancing for SVMs: A Case Study. SIGKDD Explorations 6(1) (2004) 60-69
11. Bishop, C.: Novelty Detection and Neural Network Validation. In: Proceedings of IEE Conference on Vision, Image and Signal Processing 141(4) (1994) 217-222
12. Tax, D.M.J., Duin, R.P.W.: Support Vector Data Description. Machine Learning 54 (2004) 45-66
13. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-dimensional Distribution. Neural Computation 13 (2001) 1443-1471
14. Schölkopf, B., Platt, J.C., Smola, A.J.: Kernel Method for Percentile Feature Extraction. Technical Report, MSR-TR-2000-22. Microsoft Research, WA (2000)
15. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)