ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li

***Abstract*—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.**

I. Introduction

LEARNING from imbalanced data sets is a relatively new challenge for many of today's data mining applications. From applications in Web mining to text categorization to biomedical data analysis [1], this challenge manifests itself in two common forms: minority interests and rare instances. Minority interests arise in domains where rare objects (minority class samples) are of great interest, and it is the objective of the machine learning algorithm to identify these minority class examples as accurately as possible. For instance, in financial engineering, it is important to detect fraudulent credit card activities in a pool of large transactions [2] [3]. Rare instances, on the other hand, concerns itself with situations where data representing a particular event is limited compared to other distributions [4] [5], such as the detection of oil spills from satellite images [6]. One should note that many imbalanced learning problems are caused by a combination of these two factors. For instance, in biomedical data analysis, the data samples for different kinds of cancers are normally very limited (rare instances) compared to normal non-cancerous cases; therefore, the ratio of the minority class to the majority class can be significant (at a ratio of 1 to 1000 or even more [4][7][8]). On the other hand, it is essential to predict the presence of cancers, or further classify different types of cancers as accurate as possible for earlier and proper treatment (minority interests).

Haibo He, Yang Bai, and Edwardo A. Garcia are with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, New Jersey 07030, USA (email: {hhe, ybai1, egarcia}@stevens.edu).

Shutao Li is with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China.(Email: shutao_li@hnu.cn)

This work was supported in part by the Center for Intelligent Networked Systems (iNetS) at Stevens Institute of Technology and the Excellent Youth Foundation of Hunan Province (Grant No. 06JJ1010).

Generally speaking, imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International Conference on Machine Learning workshop on Learning from Imbalanced Data Sets (ICML'03) [10], and the Association for Computing Machinery (ACM) Special Interest Group on Knowledge Discovery and Data Mining explorations (ACM SIGKDD Explorations'04) [11].

The state-of-the-art research methodologies to handle imbalanced learning problems can be categorized into the following five major directions:

(1) Sampling strategies. This method aims to develop various oversampling and/or undersampling techniques to compensate for imbalanced distributions in the original data sets. For instance, in [12] the cost curves technique was used to study the interaction of both oversampling and undersampling with decision tree based learning algorithms. Sampling techniques with the integration of probabilistic estimates, pruning, and data preprocessing were studied for decision tree learning in [13]. Additionally, in [14], "JOUS-Boost" was proposed to handle imbalanced data learning by integrating adaptive boosting with jittering sampling techniques.

(2) Synthetic data generation. This approach aims to overcome imbalance in the original data sets by artificially generating data samples. The SMOTE algorithm [15], generates an arbitrary number of synthetic minority examples to shift the classifier learning bias toward the minority class. SMOTEBoost, an extension work based on this idea, was proposed in [16], in which the synthetic procedure was integrated with adaptive boosting techniques to change the method of updating weights to better compensate for skewed distributions. In order to ensure optimal classification accuracy for minority and majority class, DataBoost-IM algorithm was proposed in [17] where synthetic data examples are generated for both minority and majority classes through the use of "seed" samples.

(3) Cost-sensitive learning. Instead of creating balanced data distributions by sampling strategies or synthetic data generation methods, cost-sensitive learning takes a different

978-1-4244-1821-3/08/$25.00©2008 IEEE

approach to address this issue: It uses a cost-matrix for different types of errors or instance to facilitate learning from imbalanced data sets. That is to say, cost-sensitive learning does not modify the imbalanced data distribution directly; instead, it targets this problem by using different cost-matrices that describe the cost for misclassifying any particular data sample. A theoretical analysis on optimal cost-sensitive learning for binary classification problems was studied in [18]. In [19] instead of using misclassification costs, an instance-weighting method was used to induce cost-sensitive trees and demonstrated better performance. In [20], Metacost, a general cost-sensitive learning framework was proposed. By wrapping a cost-minimizing procedure, Metacost can make any arbitrary classifier cost-sensitive according to different requirements. In [21], cost-sensitive neural network models were investigated for imbalanced classification problems. A threshold-moving technique was used in this method to adjust the output threshold toward inexpensive classes, such that high-cost (expensive) samples are unlikely to be misclassified.

(4) Active learning. Active learning techniques are conventionally used to solve problems related to unlabeled training data. Recently, various approaches on active learning from imbalanced data sets have been proposed in literature [1] [22] [23] [24]. In particular, an active learning method based on support vector machines (SVM) was proposed in [23] [24]. Instead of searching the entire training data space, this method can effectively select informative instances from a random set of training populations, therefore significantly reducing the computational cost when dealing with large imbalanced data sets. In [25], active learning was used to study the class imbalance problems of word sense disambiguation (WSD) applications. Various strategies including max-confidence and min-error were investigated as the stopping criteria for the proposed active learning methods.

(5) Kernel-based methods. Kernel-based methods have also been used to study the imbalanced learning problem. By integrating the regularized orthogonal weighted least squares (ROWLS) estimator, a kernel classifier construction algorithm based on orthogonal forward selection (OFS) was proposed in [26] to optimize the model generalization for learning from two-class imbalanced data sets. In [27], a kernel-boundary-alignment (KBA) algorithm based on the idea of modifying the kernel matrix according to the imbalanced data distribution was proposed to solve this problem. Theoretical analyses in addition to empirical studies were used to demonstrate the effectiveness of this method.

In this paper, we propose an adaptive synthetic (ADASYN) sampling approach to address this problem. ADASYN is based on the idea of adaptively generating minority data samples according to their distributions: more synthetic data is generated for minority class samples that are harder to learn compared to those minority samples that are easier to learn. The ADASYN method can not only reduce the learning bias introduced by the original imbalance data distribution, but can also adaptively shift the decision boundary to focus on those difficult to learn samples.

The remainder of this paper is organized as follow. Section II presents the ADASYN algorithm in detail, and discusses the major advantages of this method compared to conventional synthetic approaches for imbalanced learning problems. In section III, we test the performance of ADASYN on various machine learning test benches. Various evaluation metrics are used to assess the performance of this method against existing methods. Finally, a conclusion is presented in Section IV.

II. ADASYN ALGORITHM

Motivated by the success of recent synthetic approaches including SMOTE [15], SMOTEBoost [16], and DataBoost-IM [17], we propose an adaptive method to facilitate learning from imbalanced data sets. The objective here is two-fold: reducing the bias and adaptively learning. The proposed algorithm for the two-class classification problem is described in [Algorithm ADASYN]:

[Algorithm - ADASYN]

Input

(1) Training data set D_{tr} with m samples $\{\mathbf{x_i}, y_i\}$, $i = 1, ..., m$, where $\boldsymbol{x}_i$ is an instance in the n dimensional feature space $\boldsymbol{X}$ and $y_i \in Y = \{1, -1\}$ is the class identity label associated with $\boldsymbol{x}_i$. Define m_s and m_l as the number of minority class examples and the number of majority class examples, respectively. Therefore, $m_s \leq m_l$ and $m_s + m_l = m$.

Procedure

(1) Calculate the degree of class imbalance:

$$d = m_s/m_l \tag{1}$$

where $d \in (0,\ 1]$.

(2) If $d < d_{th}$ then (d_{th} is a preset threshold for the maximum tolerated degree of class imbalance ratio):

(a) Calculate the number of synthetic data examples that need to be generated for the minority class:

$$G = (m_l - m_s) \times \beta \tag{2}$$

Where $\beta \in [0,\ 1]$ is a parameter used to specify the desired balance level after generation of the synthetic data. $\beta = 1$ means a fully balanced data set is created after the generalization process.

(b) For each example $\boldsymbol{x_i} \in minorityclass$, find K nearest neighbors based on the Euclidean distance in n dimensional space, and calculate the ratio r_i defined as:

$$r_i = \Delta_i/K, \quad i = 1, ..., m_s \tag{3}$$

where Δ_i is the number of examples in the K nearest neighbors of $\boldsymbol{x_i}$ that belong to the majority class, therefore $r_i \in [0,\ 1]$;

(c) Normalize r_i according to $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$, so that $\hat{r}_i$ is

a density distribution ($\sum_i \hat{r_i} = 1$)

(d) Calculate the number of synthetic data examples that need to be generated for each minority example $\boldsymbol{x}_i$:

$$g_i = \hat{r_i} \times G \tag{4}$$

where G is the total number of synthetic data examples that need to be generated for the minority class as defined in Equation (2).

(e) For each minority class data example $\boldsymbol{x}_i$, generate g_i synthetic data examples according to the following steps:

Do the **Loop** from 1 to g_i:

(i) Randomly choose one minority data example, $\boldsymbol{x}_{zi}$, from the K nearest neighbors for data $\boldsymbol{x}_i$.

(ii) Generate the synthetic data example:

$$\boldsymbol{s_i} = \boldsymbol{x_i} + (\boldsymbol{x_{zi}} - \boldsymbol{x_i}) \times \lambda \tag{5}$$

where $(\boldsymbol{x_{zi}} - \boldsymbol{x_i})$ is the difference vector in n dimensional spaces, and λ is a random number: $\lambda \in [0,\ 1]$.

End **Loop**

The key idea of ADASYN algorithm is to use a density distribution $\hat{r_i}$ as a criterion to automatically decide the number of synthetic samples that need to be generated for each minority data example. Physically, $\hat{r_i}$ is a measurement of the distribution of weights for different minority class examples according to their level of difficulty in learning. The resulting dataset post ADASYN will not only provide a balanced representation of the data distribution (according to the desired balance level defined by the β coefficient), but it will also force the learning algorithm to focus on those difficult to learn examples. This is a major difference compared to the SMOTE [15] algorithm, in which equal numbers of synthetic samples are generated for each minority data example. Our objective here is similar to those in SMOTEBoost [16] and DataBoost-IM [17] algorithms: providing different weights for different minority examples to compensate for the skewed distributions. However, the approach used in ADASYN is more efficient since both SMOTEBoost and DataBoost-IM rely on the evaluation of hypothesis performance to update the distribution function, whereas our algorithm adaptively updates the distribution based on the data distribution characteristics. Hence, there is no hypothesis evaluation required for generating synthetic data samples in our algorithm.

Fig. 1 shows the classification error performance for different β coefficients for an artificial two-class imbalanced data set. The training data set includes 50 minority class examples and 200 majority class examples, and the testing data set includes 200 examples. All data examples are generated by multidimensional Gaussian distributions with different mean and covariance matrix parameters. These results are based on the average of 100 runs with a decision tree as the base classifier. In Fig. 1, $\beta = 0$ corresponds to the classification error based on the original imbalanced data set, while $\beta = 1$ represents a fully balanced data set generated by the ADASYN algorithm. Fig. 1 shows that the ADASYN algorithm can improve the classification performance by reducing the bias introduced in the original imbalanced data sets. Further more, it also demonstrates the tendency in error reduction as balance level is increased by ADASYN.

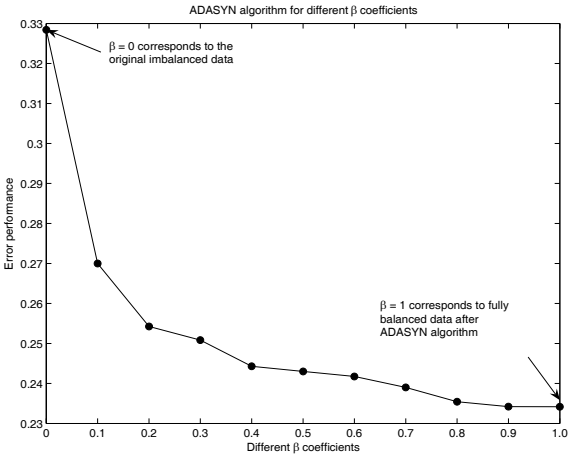

Fig. 1. ADASYN algorithm for imbalanced learning

III. Simulation Analysis and Discussions

A. Data set analysis

We test our algorithm on various real-world machine learning data sets as summarized in Table 1. All these data sets are available from the UCI Machine Learning Repository [28]. In addition, since our interest here is to test the learning capabilities from two-class imbalanced problems, we made modifications on several of the original data sets according to various literary results from similar experiments [17] [29]. A brief description of such modifications is discussed as follows.

TABLE I
Data set characteristics used in this paper.

Data set Name	# total examples	# minority examples	# majority examples	# attributes
Vehicle	846	199	647	18
Diabetes (PID)	768	268	500	8
Vowel	990	90	900	10
Ionosphere	351	126	225	34
Abalone	731	42	689	7

Vehicle dataset. This data set is used to classify a given silhouette as one of four types of vehicles [30]. This dataset has a total of 846 data examples and 4 classes (opel, saab, bus and van). Each example is represented by 18 attributes. We choose "Van" as the minority class and collapse the remaining classes into one majority class. This gives us an imbalanced two-class dataset, with 199 minority class examples and 647 majority class examples.

Pima Indian Diabetes dataset. This is a two-class data set and is used to predict positive diabetes cases. It includes a total of 768 cases with 8 attributes. We use the positive cases as the minority class, which give us 268 minority class cases and 500 majority class cases.

Vowel recognition dataset. This is a speech recognition dataset used to classify different vowels. The original dataset includes 990 examples and 11 classes. Each example is represented by 10 attributes. Since each vowel in the original data set has 10 examples, we choose the first vowel as the minority class and collapse the rest to be the majority class, which gives 90 and 900 minority and majority examples, respectively.

Ionosphere dataset. This data set includes 351 examples with 2 classes (good radar returns versus bad radar returns). Each example is represented by 34 numeric attributes. We choose the "bad radar" instances as minority class and "good radar" instance as the majority class, which gives us 126 minority class examples and 225 majority class examples.

Abalone dataset. This data set is used to predict the age of abalone from physical measurements. The original data set includes 4177 examples and 29 classes, and each example is represented by 8 attributes. We choose class "18" as the minority class and class "9" as the majority class as suggested in [17]. In addition, we also removed the discrete feature (feature "sex") in our current simulation. This gives us 42 minority class examples and 689 majority class examples; each represented by 7 numerical attributes.

B. Evaluation metrics for imbalanced data sets

Instead of using the overall classification accuracy as a single evaluation criterion, we use a set of assessment metrics related to receiver operating characteristics (ROC) graphs [31] to evaluate the performance of ADASYN algorithm. We use ROC based evaluation metrics because under the imbalanced learning condition, traditional overall classification accuracy may not be able to provide a comprehensive assessment of the observed learning algorithm [17] [31] [32] [33] [6] [34] [16]. Let $\{p, n\}$ be the positive and negative testing examples and $\{Y, N\}$ be the classification results given by a learning algorithm for positive and negative predictions. A representation of classification performance can be formulated by a confusion matrix (contingency table) as illustrated in Fig. 2. We followed the suggestions of [15] [34] and use the minority class as the positive class and majority class as the negative class.

Based on Fig. 2, the evaluation metrics used to assess learning from imbalanced data sets are defined as:

Overall Accuracy (OA):

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

Precision:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

True class \ Hypothesis output	Y	N
p	TP (True Positives)	FN (False Negatives)
n	FP (False Positives)	TN (True Negatives)

Fig. 2. Confusion matrix for performance evaluation

F_Measure:

$$F\_Measure = \frac{(1 + \beta^2) \cdot recall \cdot precision}{\beta^2 \cdot recall + precision} \tag{9}$$

Where β is a coefficient to adjust the relative importance of precision versus recall (usually $\beta = 1$).

G_mean:

$$\begin{aligned} G\_mean &= \sqrt{PositiveAccuracy \times NegativeAccuracy} \\ &= \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \end{aligned} \tag{10}$$

C. Simulation analyses

We use the decision tree as the base learning model in our current study. According to the assessment metrics presented in Section III-B, Table 2 illustrates the performance of the ADASYN algorithm compared to the SMOTE algorithm. As a reference, we also give the performance of the decision tree learning based on the original imbalanced data sets. These results are based on the average of 100 runs. At each run, we randomly select half of the minority class and majority class examples as the training data, and use the remaining half for testing purpose. For both SMOTE and ADASYN, we set the number of nearest neighbors $K = 5$. Other parameters include $N = 200$ for SMOTE according to [15], $\beta = 1$ and $d_{th} = 0.75$ for ADASYN.

For each method, the best performance is highlighted in each category. In addition, the total winning times for each method across different evaluation metrics are also shown in Table 2. Based on these simulation results, the ADASYN algorithm can achieve competitive results on these five test benches. As far as the overall winning times are concerned, ADASYN outperforms the other methods. Further more, ADASYN algorithm also provides the best performance in terms of G-mean for all data sets. This means our algorithm provides improved accuracy for both minority and majority classes and does not sacrifice one class in preference for another. This is one of the advantages of our method to handle the imbalanced learning problems.

There is another interesting observation that merit further discussion. From Table 2 one can see there are situations that learning from the original data set can actually achieve better performance for certain assessment criterion, such as the precision assessment. This raises an important question: generally speaking, to what level the imbalanced learning

TABLE II
EVALUATION METRICS AND PERFORMANCE COMPARISON

Dataset	Methods	OA	Precision	Recall	F_measure	G_mean
Vehicle	Decision tree	0.9220	**0.8454**	0.8199	0.8308	0.8834
	SMOTE	0.9239	0.8236	0.8638	0.8418	0.9018
	ADASYN	**0.9257**	0.8067	**0.9015**	**0.8505**	**0.9168**
Pima Indian Diabetes	Decision tree	0.6831	**0.5460**	0.5500	0.5469	0.6430
	SMOTE	0.6557	0.5049	**0.6201**	0.5556	0.6454
	ADASYN	**0.6837**	0.5412	0.6097	**0.5726**	**0.6625**
Vowel recognition	Decision tree	**0.9760**	**0.8710**	0.8700	0.8681	0.9256
	SMOTE	0.9753	0.8365	0.9147	**0.8717**	0.9470
	ADASYN	0.9678	0.7603	**0.9560**	0.8453	**0.9622**
Ionosphere	Decision tree	0.8617	**0.8403**	0.7698	0.8003	0.8371
	SMOTE	0.8646	0.8211	0.8032	0.8101	0.8489
	ADASYN	**0.8686**	0.8298	**0.8095**	**0.8162**	**0.8530**
Abalone	Decision tree	**0.9307**	**0.3877**	0.2929	**0.3249**	0.5227
	SMOTE	0.9121	0.2876	0.3414	0.3060	0.5588
	ADASYN	0.8659	0.2073	**0.4538**	0.2805	**0.6291**
Winning times	Decision tree	2	**5**	0	1	0
	SMOTE	0	0	1	1	0
	ADASYN	**3**	0	**4**	**3**	**5**

methods such as adjusting the class balance can help the learning capabilities? This is a fundamental and critical question in this domain. In fact, the importance of this question has been previously addressed by F. Provost in the invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets [1]:

"Isn't the best research strategy to concentrate on how machine learning algorithms can deal most effectively with whatever data they are given?"

Based on our simulation results, we believe that this fundamental question should be investigated in more depth both theoretically and empirically in the research community to correctly understand the essence of imbalanced learning problems.

D. Discussions

As a new learning method, ADASYN can be further extended to handle imbalanced learning in different scenarios, therefore potentially benefit a wide range of real-world applications for learning from imbalanced data sets. We give a brief discussion on possible future research directions in this Section.

Firstly of all, in our current study, we compared the ADASYN algorithm to single decision tree and SMTOE algorithm [15] for performance assessment. This is mainly because all of these methods are single-model based learning algorithms. Statistically speaking, ensemble based learning algorithms can improve the accuracy and robustness of learning performance, thus as a future research direction, the ADASYN algorithm can be extended for integration with ensemble based learning algorithms. To do this, one will need to use a bootstrap sampling technique to sample the original training data sets, and then embed ADASYN to each sampled set to train a hypothesis. Finally, a weighted combination voting rule similar to AdaBoost.M1 [35] [36] can be used to combine all decisions from different hypotheses for the final predicted outputs. In such situation, it would be interesting to see the performance of such boosted ADASYN algorithm with those of SMOTEBoost [16], DataBoost-IM [17] and other ensemble

based imbalanced learning algorithms.

Secondly, ADASYN can be generalized to multiple-class imbalanced learning problems as well. Although two-class imbalanced classification problems dominate the research activities in today's research community, this is not a limitation to our method. To extend the ADASYN idea to multi-class problems, one first needs to calculate and sort the degree of class imbalance for each class with respect to the most significant class, $y_s \in Y = \{1, ..., C\}$, which is defined as the class identity label with the largest number of examples. Then for all classes that satisfy the condition $d < d_{th}$, the ADASYN algorithm is executed to balance them according to their own data distribution characteristics. In this situation, the update of r_i in equation (3) can be modified to reflect different needs in different applications. For instance, if one would like to balance the examples in class y_k, ($y_k \in \{1, ..., C\}$ and $y_k \neq y_s$), then the definition of Δ_i in equation (3) can be defined as the number of examples in the nearest neighbors belonging to class y_s, or belonging to all other classes except y_k (similar to transforming the calculation of the nearest neighbors to a Boolean type function: belonging to y_k or not belonging to y_k).

Further more, the ADASYN algorithm can also be modified to facilitate incremental learning applications. Most current imbalanced learning algorithms assume that representative data samples are available during the training process. However, in many real-world applications such as mobile sensor networks, Web mining, surveillance, homeland security, and communication networks, training data may continuously become available in small chunks over a period of time. In this situation, a learning algorithm should have the capability to accumulate previous experience and use this knowledge to learn additional new information to aid prediction and future decision-making processes. The ADASYN algorithm can potentially be adapted to such an incremental learning scenario. To do this, one will need to dynamically update the r_i distribution whenever a new chunk of data samples is received. This can be accomplished by an online learning and evaluation process.

IV. Conclusion

In this paper, we propose a novel adaptive learning algorithm ADASYN for imbalanced data classification problems. Based on the original data distribution, ADASYN can adaptively generate synthetic data samples for the minority class to reduce the bias introduced by the imbalanced data distribution. Further more, ADASYN can also autonomously shift the classifier decision boundary to be more focused on those difficult to learn examples, therefore improving learning performance. These two objectives are accomplished by a dynamic adjustment of weights and an adaptive learning procedure according to data distributions. Simulation results on five data sets based on various evaluation metrics show the effectiveness of this method.

Imbalanced learning is a challenging and active research topic in the artificial intelligence, machine learning, data mining and many related areas. We are currently investigating various issues, such as multiple classes imbalanced learning and incremental imbalanced learning. Motivated by the results in this paper, we believe that ADASYN may provide a powerful method in this domain.

References

[1] F. Provost, "Machine Learning from Imbalanced Data Sets 101," Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets, Menlo Park, CA, 2000.

[2] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolf, "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intelligent Systems*, pp. 67-74, November/December 1999.

[3] 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 *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'01)*, pp. 164-168, 2001.

[4] G. M. Weiss, "Mining with Rarity: A Unifying Framework," *SIGKDD Explorations*, 6(1):7-19, 2004.

[5] G. M. Weiss "Mining Rare Cases," In O. Maimon and L. Rokach (.eds), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers, pp. 765-776, 2005.

[6] M. Kubat, R. C. Holte, and S. Matwin, "Machine Learning for the Detection of Oil Spills in Satellite Radar Images," *Machine Learning*, 30(2):195-215, 1998.

[7] H. He and X. Shen, "A Ranked Subspace Learning Method for Gene Expression Data Classification," in *Proc. Int. Conf. Artificial Intelligence (ICAI'07)*, pp. 358 - 364, June 2007

[8] R. Pearson, G. Goney, and J. Shwaber, "Imbalanced Clustering for Microarray Time-Series," in *Proc. ICML'03 workshop on Learning from Imbalanced Data Sets*, 2003

[9] N. Japkowicz, (Ed.), "Learning from Imbalanced Data Sets," the AAAI Workshop, Technical Report WS-00-05, American Association for Artificial Intelligence, Menlo Park, CA, 2000.

[10] N. V. Chawla, N. Japkowicz, and A. Ko lcz, (Ed.), "Imbalanced Clustering for Microarray Time-Series," in *Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003

[11] N. V. Chawla, N. Japkowicz and A. Kolcz, SIGKDD Explorations: Special issue on Learning from Imbalanced Datasets, vol.6, issue 1, 2004.

[12] C. Drummond and R. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling Beats Oversampling," in *Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets*, 2003

[13] N. Chawla, "C4.5 and Imbalanced Datasets: Investigating the Effect of Sampling Method, Probalistic Estimate, and Decision Tree Structure," in *ICML-KDD'03 Workshop: Learning from Imbalanced Data Sets*, 2003

[14] D. Mease, A. J. Wyner, and A. Buja, "Boosted Classification Trees and Class Probability/Quantile Estimation," *Journal of Machine Learning Research*, vol. 8, pp. 409- 439, 2007.

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

[16] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost: Improving Prediction of the Minority Class in Boosting," in *Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases*, pp. 107-119, Dubrovnik, Croatia, 2003.

[17] H. Guo and H. L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: the DataBoost-IM Approach," in *SIGKDD Explorations: Special issue on Learning from Imbalanced Datasets*, vol.6, issue 1, pp. 30 - 39, 2004.

[18] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI'01)*, pp. 973-978, 2001.

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

[20] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 155-164, San Diego, CA, 1999.

[21] Z. H. Zhou and X. Y. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63-77, 2006.

[22] N. Abe, "Invited talk: Sampling Approaches to Learning From Imbalanced Datasets: Active Learning, Cost Sensitive Learning and Beyond," in *ICML-KDD'03 Workshop: Learning from Imbalanced Data Sets*, 2003.
[23] S. Ertekin, J. Huang, and C. L. Giles, "Active Learning for Class Imbalance Problem," in *Proc. Annual Int. ACM SIGIR Conf. Research and development in information retrieval*, pp. 823 - 824, Amsterdam, Netherlands, 2007.
[24] S. Ertekin, J. Huang, L. Bottou, C. L. Giles, "Learning on the Border: Active Learning in Imbalanced Data Classification," in *CIKM'07*, November 6-8, 2007, Lisboa, Portugal.
[25] J. Zhu and E. Hovy, "Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem," in *Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 783-790, Prague, June 2007.
[26] X. Hong, S. Chen, and C. J. Harris, "A Kernel-Based Two-Class Classifier for Imbalanced Data Sets," *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 28-41, 2007.
[27] G. Wu and E. Y. Chang, "KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no.6, pp. 786-795, 2005.
[28] UCI Machine Learning Repository, [online], available: http://archive.ics.uci.edu/ml/
[29] F. Provost, T. Fawcett, and R. Kohavi, "The Case Against Accuracy Estimation for Comparing Induction Algorithms," in *Proc. Int. Conf. Machine Learning*, pp. 445-453 Madison, WI. Morgan Kauffmann, 1998
[30] J. P. Siebert, "Vehicle Recognition Using Rule Based Methods," Turing Institute Research Memorandum TIRM-87-018, March 1987.
[31] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," Technical Report HPL-2003-4, HP Labs, 2003.
[32] F. Provost and T. Fawcett, "Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distributions," in *Proc. Int. Conf. Knowledge Discovery and Data Mining*, Menlo Park, CS, AAAI Press, 43-48, 1997.
[33] M. A. Maloof, "Learning When Data Sets Are Imbalanced and When Cost Are Unequal and Unkown," in *ICML'03 Workshop on Learning from Imbalanced Data Sets II*, 2003
[34] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-sided Selection," in *Proc. Int. Conf. Machine Learning*, San Francisco, CA, Morgan Kaufmann, pp. 179-186, 1997.
[35] Y. Freund and R. E. Schapire, "Experiments With a New Boosting Algorithm," in *Proc. Int. Conf. Machine Learning (ICML'96)*, pp. 148-156, 1996.
[36] Y. Freund and R. E. Schapire, "Decision-theoretic Generalization of Online Learning and Application to Boosting," in *J. Computer and Syst. Sciences*, vol. 55, no. 1, pp. 119-139, 1997.