Implementing Improved Synthetic Minority over Sampling Techniques for Imbalanced Learning

A. Bhuvaneswari[1] Dr. R. Manicka Chezian[2]
[1]Research Scholar [2]Associate Professor & Head of Department
[1,2]Department of Computer Science
[1,2]NGM College, Pollachi, India

Abstract— This paper exhibits a novel versatile manufactured (ISMOTE) examining approach for gaining from imbalanced informational collections. The fundamental thought of ISMOTE is to utilize a weighted appropriation for various minority class precedents as indicated by their dimension of trouble in realizing, where more manufactured information is created for minority class models that are harder to learn contrasted with those minority precedents that are less demanding to learn. Accordingly, the ISMOTE approach enhances learning as for the information disseminations in two different ways: (1) diminishing the predisposition presented by the class awkwardness, and (2) adaptively moving the grouping choice limit toward the troublesome precedents. Reproduction examinations on a few machine learning informational indexes demonstrate the viability of this technique crosswise over five assessment measurements.
Key words: Over Sampling, Cost Delicate Learning, Kernal Adaptive Subspace, Clustering, NN

I. INTRODUCTION

Gaining from imbalanced informational collections is a moderately new test for a considerable lot of the present information mining applications.

From applications in Web mining to content classification to biomedical information investigation [1], this test shows itself in two basic structures: minority premiums and uncommon cases. Minority premiums emerge in areas where uncommon articles (minority class tests) are of incredible premium, and it is the goal of the machine learning calculation to distinguish these minority class models as precisely as would be prudent. For example, in budgetary building, it is imperative to recognize false Mastercard exercises in a pool of vast exchanges [2] [3]. Uncommon examples, then again, worries about circumstances where information speaking to a specific occasion is constrained contrasted with different disseminations [4] [5], for example, the recognition of oil slicks from satellite pictures [6]. One should take note of that numerous imbalanced learning issues are caused by a blend of these two elements. For example, in biomedical information investigation, the information tests for various types of malignant growths are regularly extremely restricted (uncommon occasions) contrasted with ordinary non-dangerous cases; in this manner, the proportion of the minority class to the lion's share class can be noteworthy (at a proportion of 1 to 1000 or significantly more [4][7][8]). Then again, it is fundamental to foresee the nearness of malignant growths, or further arrange diverse kinds of tumors as precise as feasible for prior and legitimate treatment (minority premiums).

As a rule, imbalanced learning happens at whatever point a few sorts of information dissemination fundamentally rule the example space contrasted with other information dispersions. In this paper, we center on the two-class grouping issue for imbalanced informational collections, a point of real concentration in late research exercises in the exploration network. As of late, hypothetical examination and handy applications for this issue have pulled in a developing consideration from both scholarly community and industry. This is reflected by the foundation of a few noteworthy work-shops and uncommon issue gatherings, 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 investigations (ACM SIGKDD Explorations'04) [11].

The best in class investigate techniques to deal with imbalanced learning issues can be classified into the accompanying five noteworthy headings:

1) Sampling systems. This strategy expects to create various oversampling as well as under sampling systems to compensate for imbalanced dispersions in the first informational collections. For example, in [12] the cost bends strategy was utilized to think about the communication of both oversampling and under sampling with choice tree based learning calculations. Inspecting techniques with the coordination of probabilistic appraisals, pruning, and information preprocessing were examined for choice tree learning in [13]. Furthermore, in [14], "JOUS-Boost" was proposed to deal with imbalanced information learning by incorporating versatile boosting with jittering examining strategies.
2) Synthetic information age. This methodology plans to over-come awkwardness in the first informational indexes by misleadingly generating information tests. The SMOTE calculation [15], creates a self-assertive number of manufactured minority precedents to move the classifier learning predisposition toward the minority class. Destroyed Boost, an augmentation work dependent on this thought, was proposed in [16], in which the manufactured system was incorporated with versatile boosting procedures to change the technique for refreshing weights to all the more likely make up for skewed appropriations. With the end goal to guarantee ideal grouping precision for minority and greater part class, DataBoostIM calculation was proposed in [17] where manufactured information precedents are created for both minority and larger part classes using "seed" tests.
3) Cost-delicate learning. Rather than making adjusted information appropriations by inspecting methodologies or engineered information age strategies, cost-touchy

learning adopts an alternate strategy to address this issue: It utilizes a cost-framework for various sorts of mistakes or occurrence to encourage gaining from imbalanced informational indexes. In other words, cost-touchy learning does not alter the imbalanced information dispersion specifically; rather, it focuses on this issue by utilizing distinctive cost-lattices that depict the expense for misclassifying a specific information test. A hypothetical examination on ideal cost-touchy mastering for twofold grouping issues was considered in [18]. In [19] as opposed to utilizing misclassification costs, an example weighting strategy was utilized to actuate cost-delicate trees and exhibited better execution. In [20], Metacost, a general cost-touchy learning system was proposed. By wrapping an expense limiting system, Metacost can make any self-assertive classifier cost-touchy as per distinctive necessities. In [21], cost-touchy neural system models were researched for imbalanced characterization issues. A limit moving procedure was utilized in this strategy to alter the yield edge toward economical classes, with the end goal that staggering expense (costly) examples are probably not going to be misclassified.

4) Active learning. Dynamic learning systems are conventionally used to take care of issues identified with unlabeled preparing information. As of late, different methodologies on dynamic gaining from imbalanced informational indexes have been proposed in writing [1] [22] [23][24].

Specifically, a functioning learning strategy dependent on help vector machines (SVM) was proposed in [23][24]. Rather than looking through the whole preparing information space, this strategy can viably choose instructive occasions from an irregular arrangement of preparing populaces, accordingly fundamentally lessening the computational cost when managing vast imbalanced informational collections. In [25], dynamic learning was utilized to think about the class awkwardness issues of word sense disambiguation (WSD) applications. Different techniques including max-certainty and min-mistake were examined as the ceasing criteria for the proposed dynamic learning strategies.

5) Kernel-based strategies. Piece based strategies have likewise been utilized to consider the imbalanced learning issue. By coordinating the regularized symmetrical weighted minimum squares (ROWLS) estimator, a piece classifier development calculation dependent on symmetrical forward determination (OFS) was proposed in [26]to enhance the model speculation for gaining from two-class imbalanced informational indexes. In [27], a part limit arrangement (KBA) calculation dependent on changing the bit grid as per the imbalanced information conveyance was proposed to tackle this issue. Hypothetical investigations notwithstanding exact examinations were utilized to exhibit the adequacy of this strategy.

In this paper, propose a versatile manufactured (ISMOTE) inspecting way to deal with location this issue. ISMOTE depends on the possibility of adaptively creating minority information tests as indicated by their dispersions: more manufactured information is produced for minority class tests that are harder to learn contrasted with those minority tests that are less demanding to learn. The ISMOTE technique cannot just lessen the learning predisposition presented by the first lopsidedness information conveyance, however can likewise adaptively move the choice limit to concentrate on those hard to learn tests.

The rest of this paper is composed as pursue. Segment II presents the ISMOTE calculation in detail, and examines the significant focal points of this technique contrasted with ordinary engineered approaches for imbalanced learning issues. In area III, we test the execution of ISMOTE on different machine learning test seats. Different assessment measurements are utilized to evaluate the execution of this technique against existing strategies. At long last, an end is introduced in Section IV.

II. ISMOTE ALGORITHM

In the minority class for the training set used in 10-fold cross-validation. The minority class was over-sampled at 100%, 200%, 300%, 400% and 500% of its original size. The graphs show that the tree sizes for minority over-sampling with replacement at higher degrees of replication are much greater than those for ISMOTE, and the minority class recognition of the minority over-sampling with replacement technique at higher degrees of replication isn't as good as ISMOTE

SMOTE (T, N, k) Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbors k Output: (N/100) * T synthetic minority class samples

1) (∗ If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTE d.∗)
2) if N < 100
3) then Randomize the T minority class samples
4) T = (N/100) ∗ T
5) N = 100
6) endif
7) N = (int)(N/100) (∗ The amount of SMOTE is assumed to be in integral multiples of 100.
8) k = Number of nearest neighbours
9) numattrs = Number of attributes
10) Sample[ ][ ]: array for original minority class samples
11) new index: keeps a count of number of synthetic samples generated, initialized to
12) Synthetic[ ][ ]: array for synthetic samples (∗ Compute k nearest neighbours for each minority class sample only.)
13) for i ← 1 to T
14) Compute k nearest neighbours for i, and save the indices in the nnarray
15) Populate(N, i, nnarray)
16) endfor Populate(N, i, nnarray) (∗ Function to generate the synthetic samples. ∗)
17) while N 6= 0
18) Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbours of i.
19) for attr ← 1 to numattrs
20) Compute: dif = Sample[nnarray[nn]][attr] − Sample[i][attr]
21) Compute: gap = random number between 0 and 1
22) Synthetic[newindex][attr] = Sample[i][attr] + gap ∗ dif

23) endfor
24) new index++ 25. N = N − 1 26. endwhile 27. return (∗ End of Populate. ∗) End of Pseudo-Code.

The key thought of ISMOTE calculation is to utilize a thickness.

Conveyance rˆ as a rule to consequently choose the i number of manufactured examples that should be produced for every minority information model. Physically, rˆ is an estimation i of the conveyance of weights for various minority class models as indicated by their dimension of trouble in learning.

The subsequent dataset post ISMOTE won't just give a reasonable portrayal of the information conveyance (as indicated by the coveted parity level characterized by the β coefficient), however it will likewise drive the learning calculation to concentrate on those hard to learn precedents. This is a noteworthy distinction contrasted with the SMOTE [15] calculation, in which parallel quantities of engineered tests are created for every minority information precedent. Our goal here is like those in SMOTEBoost [16] and DataBoost-IM [17] calculations: giving distinctive weights to various minority precedents to make up for the skewed disseminations. In any case, the methodology utilized in ISMOTE is more productive since both SMOTEBoost and DataBoost-IM depend on the assessment of speculation execution to refresh the circulation work, though our calculation adaptively refreshes the conveyance dependent on the information appropriation qualities. Subsequently, there is no speculation assessment required for creating manufactured information tests in our calculation.

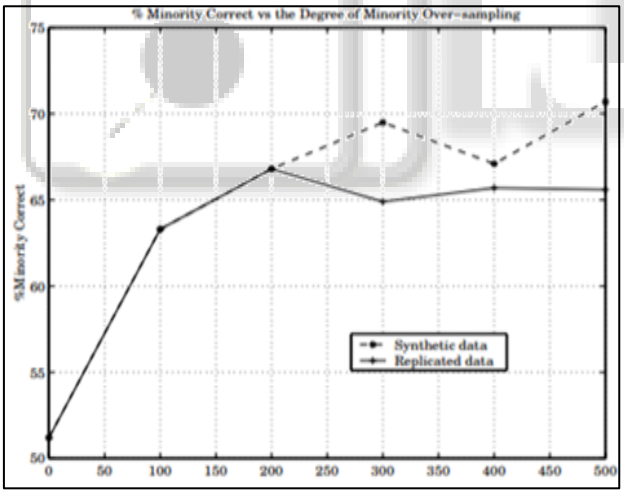

Fig. 1: Synthetic Minority over Sampling Technique

Fig. 1 demonstrates the order mistake execution for vary ent β coefficients for a counterfeit two-class imbalanced dataset. The preparation informational index incorporates 50 minority class models and 200 lion's share class precedents, and the testing informational collection incorporates 200 models. All information models are produced by multidimensional Gaussian dispersions with various mean and covariance network parameters. These outcomes depend on the normal of 100 keeps running with a choice tree as the base classifier. In Fig. 1, β = 0 relates to the order mistake dependent on the first imbalanced dataset, while β = 1 speaks to a completely adjusted informational collection produced by the ISMOTE .

III. Simulation Analysis and Discussions

A. Data Set Analysis

This paper test the calculation on different certifiable machine learning informational collections as condensed in Table 1. Every one of these informational collections are accessible from the UCI Machine Learning Repository [28]. Likewise, since our enthusiasm here is to test the taking in abilities from two-class imbalanced issues, we made changes on a few of the first informational indexes as indicated by different abstract outcomes from comparable analyses [17] [29]. A short portrayal of such adjustments is talked about as pursues.

This dataset is utilized to group a given outline as one of four sorts of vehicles [30]. This dataset has a sum of 846 information models and 4 classes (Opel, Saab, transport and van). Every precedent is spoken to by 18 qualities. We pick "Van" as the minority class and fall the rest of the classes into one dominant part class. This gives us an imbalanced two-class dataset, with 199 minority class precedents and 647 larger part class models.

Pima Indian Diabetes dataset. This is a two-class informational index and is utilized to anticipate positive diabetes cases. It incorporates an aggregate of 768 cases with 8 characteristics. We utilize the positive cases as the minority class, which give us 268 minority class cases and 500 lion's share class cases.

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

Table 1: Data Set Characteristics Used In This Paper

1) Vowel Acknowledgment Dataset

This is a discourse acknowledgment dataset used to order unique vowels. The first dataset incorporates 990 models and 11 classes. Every model is represented by 10 characteristics. Since every vowel in the first informational index has 10 precedents, we pick the primary vowel as the minority class and crumple the rest to be the lion's share class, which gives 90 and 900 minority and dominant part models, individually.

2) Ionosphere Dataset

This informational collection incorporates 351 precedents with 2 classes (great radar returns versus terrible radar returns). Every model is spoken to by 34 numeric traits. We pick the "awful radar" cases as minority class and "great radar" occurrence as the dominant part class, which gives us 126 minority class precedents and 225 larger part class models.

3) Abalone Dataset

This informational collection is utilized to anticipate the time of abalone from physical estimations. The first informational collection incorporates 4177 precedents and 29 classes, and every model is spoken to by 8 qualities. We pick class "18" as the minority class and class "9" as the larger part class as

proposed in [17]. What's more, we additionally evacuated the discrete component (include "sex") in our present reproduction. This gives us 42 minority class models and 689 dominant part class precedents; each spoken to by 7 numerical characteristics.

B. Assessment Measurements for Imbalanced Informational Indexes

Rather than utilizing the general characterization precision as a solitary assessment rule, we utilize an arrangement of appraisal measurements identified with recipient working attributes (ROC) charts [31] to assess the execution of ISMOTE calculation. We utilize ROC based assessment measurements on the grounds that under the imbalanced learning condition, customary generally speaking grouping precision will be unable to give an exhaustive appraisal of the watched learning calculation [17] [31] [32] [33] [6] [34] [16]. Let {p, n} be the positive and negative testing precedents and {Y, N } be the grouping results given by a learning calculation for positive and negative forecasts. A portrayal of order execution can be defined by a disarray lattice (possibility table) as represented in Fig. 2. We pursued the proposals of [15] [34] and utilize the minority class as the positive class and greater part class as the negative class.

C. Simulation Analyses

As another learning strategy, ISMOTE can be additionally reached out to deal with imbalanced learning in various situations, subsequently conceivably advantage an extensive variety of certifiable applications for gaining from imbalanced informational indexes. We give a brief talk on conceivable future research headings in this Section.

Right off the bat of all, in our current investigation, we contrasted the ISMOTE calculation with a solitary choice tree and SMTOE calculation [15] for execution appraisal. This is principally on the grounds that these techniques are single-show based learning calculations. Factually, gathering based learning calculations can enhance the exactness and power of learning execution, subsequently as a future research bearing, the ISMOTE calculation can be stretched out for mix with outfit based learning calculations. To do this, one should utilize a bootstrap examining system to test the first preparing informational collections, and afterward insert ISMOTE to each inspected set to prepare a theory. At long last, a weighted blend casting a ballot rules like AdaBoost.M1 [35] [36] can be utilized to join all choices from various theories for the last anticipated yields. In such circumstance, it is fascinating to see the execution of such supported ISMOTE calculation with those of SMOTEBoost [16], DataBoost-IM [17] and other outfit based imbalanced learning calculations.

Also, ISMOTE can be summed up to numerous class imbalanced learning issues too. Albeit two-class imbalanced order issues rule the exploration exercises in the present research network, this isn't a confinement to our technique. To stretch out the ISMOTE thought to multi-class issues, one first needs to ascertain and sort the level of class awkwardness for each class as for the most huge class, ys ∈ Y = {1, ..., C}, which is characterized as the class personality mark with the biggest number of precedents. At that point for all classes that fulfill the condition d < dth, the ISMOTE calculation is executed to adjust them as per their own information dispersion attributes. In this circumstance, the refresh of ri in condition (3) can be altered to reflect distinctive needs in various applications. For example, in the event that one might want to adjust the precedents in class yk , (yk ∈ {1, ..., C} and yk = ys), at that point the meaning of I in condition (3) can be characterized as the quantity of models in the closest neighbors having a place with class ys, or having a place with every single different class aside from yk (like changing the count of the closest neighbors to a Boolean sort work: having a place with yk or not having a place with yk).

Moreover, the ISMOTE calculation can likewise be changed to encourage gradual learning applications. Most present imbalanced learning calculations expect that agent information tests are accessible amid the preparation procedure. In any case, in some genuine applications, for example, portable sensor systems, Web mining, observation, country security, and correspondence systems, preparing information may constantly end up accessible in little lumps over some stretch of time. In this circumstance, a learning calculation ought to have the ability to amass past involvement and utilize this information to take in extra new data to help forecast and future basic leadership forms. The ISMOTE calculation can possibly be adjusted to such a gradual learning situation. To do this, one should progressively refresh the ri dissemination at whatever point another piece of information tests is gotten. This can be practiced by a web based learning and assessment process.

IV. RESULTS & DISCUSSION

To measure the efficacy of the proposed ISMOTE and compare its performance and compare its differences with SMOTE, Borderline-SMOTE. The difference between the two We collected twenty real-world data sets from the UCI website and the KEEL website for two statistical tests. Abalone signed rank test and paired t-tests.

For the assessment of imbalances in two categories, we often refer to the lesser categories as Positive class and the larger categories to negative classes. The confusion matrix is a very typical evaluation method. We show it in Table 1, the column represents the real category label, the real category label, and the row represents the other label predicted by the classifier. TP (True Positive) is a small number of categories that are correctly classified by the classifier. FN ((False Negative) is a few categories that are incorrectly misclassified by the classifier. FP (False Positive) is the most misclassified classifier by the classifier. TN (True Negative) is correctly classified by the classifier Most of the other places. In addition to using fusion matrix, there are several composite performance indicators calculated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is the correct proportion of classifiers in all instances. In general, the higher the Accuracy, the better the performance of the measured algorithm. But it does not apply to category imbalances because the number of Positive class instances is less than the number of Negative class instances.

$$TP_{rate} = \frac{TP}{TP + FN} \text{ (also known as Recall)}$$

Recall is the correct proportion of the classifier in all Positive class instances. I.e. A small class of Accuracy.

$$FP_{rate} = \frac{FP}{TN + FP}$$

$FP_{rate}$is the proportion of classifier errors in all Negative class instances? A common example is a false alarm. The higher the FP_{rate}, the higher the number of false alarms may occur.

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

Precision is the correct proportion of the classifier class in all instances where the classifier is judged to be positives. The choice between Recall and Recall is that the positives instance is very expensive to be judged by the classifier. If so, it is better to use Recall.

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

AUC (the Area Under the Curve) is the area between the ROC curve and the coordinate axis. AUC stands for randomly selecting a positive instance and randomly selecting a negative instance, and then the classifier will then use this classifier to predict the correct ratio of this positive instance will be higher than the rate at which the classifier will predict the wrong instance prediction error. AUC is an indicator that is often used to measure the performance of classifiers. The larger the AUC value, the representative points

$$G - mean = \sqrt{PositiveAccuracy * NegativeAccuracy} = \sqrt{\frac{TP}{Tp + FN} * \frac{TN}{Tn + FP}}$$

G-mean reflects the ability of the classifier to balance the two, reflecting the ability of the classifier to balance the two. G-mean is a more comprehensive indicator of the performance of the classifier, because it considers both the classifier is Accuracy for the positive class instance and Accuracy for the Negative class instance. Therefore, the larger the G-mean indicator, the larger the indicator for the classifier, and the better the ability of the classifier to correctly judge the two types of instances.

$$F - measure = \frac{(1 + \beta^2) * Recall * Precision}{\beta^2 * Recall + Precision}$$

F-measure parameters β is a user-adjustable parameter used to weigh Recall & the importance of the Precision two indicators, but often set to 1, representing Recall is as important as Precision (our experiment also sets β to 1). F-measure is a simultaneous consideration. It is a value that considers both Precision and Recall. If both Precision and Recall are high, F-measure will be very high. Therefore F- measure can be used as a measure of the strength of the classifier in dealing with the problem of unevenness. Can be used as a measure of the strength of the classifier in dealing with the problem of unevenness.

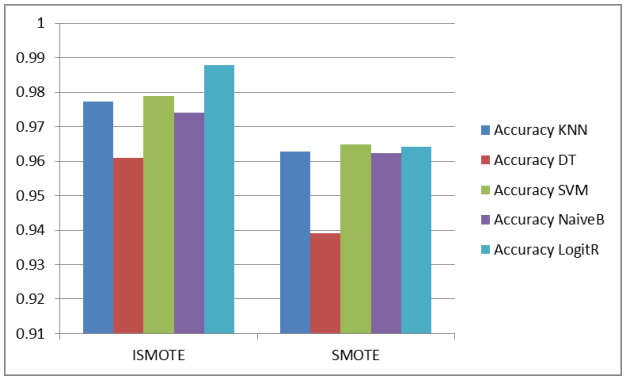

Fig. 2: Abalone

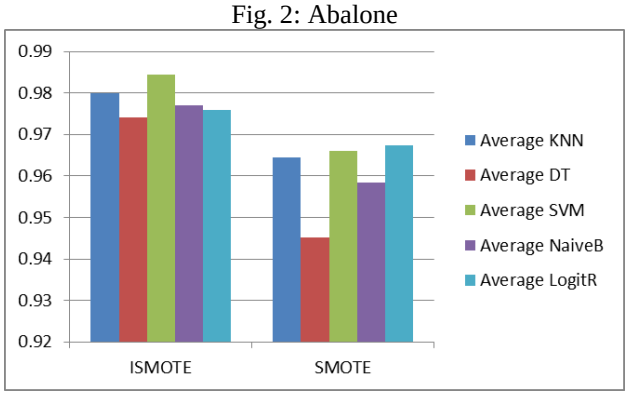

Fig. 3: Vehicle

V. CONCLUSIONS

In this paper, we propose a novel versatile learning calculation ISMOTE for imbalanced information order issues. In light of the first information appropriation, ISMOTE can adaptively create engineered information tests for the minority class to diminish the inclination presented by the imbalanced information dispersion. Furthermore, ISMOTE can likewise self-governing move the classifier choice limit to be more centered on those hard to learn models, along these lines enhancing learning execution. These two destinations are practiced by a dynamic change of weights and a versatile learning method as indicated by information conveyances. Reproduction results on five informational collections dependent on different assessment measurements demonstrate the viability of this strategy.

Imbalanced learning is a testing and dynamic research point in the man-made consciousness, machine learning, information mining and many related regions. We are as of now researching different issues, for example, numerous classes' imbalanced learning and gradual imbalanced learning. Inspired by the outcomes in this paper, we trust that ISMOTE may give a ground-breaking technique in this area.

REFERENCES

[1] F. Provost, "Machine Learning from Imbalanced Data Sets 101," Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets, MenloPark, CA, 2000.
[2] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolf, "Distributed DataMining 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-uniformclass 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 Guidefor 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 costsensitive," 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, June2007.
[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
[28] Distribution," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no.6, pp. 786-795, 2005.
[29] UCI Machine Learning Repository, [online], available: http://archive.ics.uci.edu/ml/
[30] 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
[31] J. P. Siebert, "Vehicle Recognition Using Rule Based Methods," Turing Institute Research Memorandum TIRM-87-018, March 1987.
[32] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," Technical Report HPL-2003-4, HP Labs, 2003.
[33] 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.
[34] 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

[35] 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.
[36] Y. Freund and R. E. Schapire, "Experiments With a New Boosting Algorithm," in Proc. Int. Conf. Machine Learning (ICML'96), pp. 148- 156, 1996.
[37] 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

