# KernelADASYN: Kernel Based Adaptive Synthetic Data Generation for Imbalanced Learning

Bo Tang, *Student Member, IEEE,* and Haibo He, *Senior Member, IEEE*

***Abstract*—In imbalanced learning, most standard classification algorithms usually fail to properly represent data distribution and provide unfavorable classification performance. More specifically, the decision rule of minority class is usually weaker than majority class, leading to many misclassification of expensive minority class data. Motivated by our previous work ADASYN [1], this paper presents a novel kernel based adaptive synthetic over-sampling approach, named KernelADASYN, for imbalanced data classification problems. The idea is to construct an adaptive over-sampling distribution to generate synthetic minority class data. The adaptive over-sampling distribution is first estimated with kernel density estimation methods and is further weighted by the difficulty level for different minority class data. The classification performance of our proposed adaptive over-sampling approach is evaluated on several real-life benchmarks, specifically on medical and healthcare applications. The experimental results show the competitive classification performance for many real-life imbalanced data classification problems.**

***Index Terms*—Imbalanced learning, adaptive over-sampling, kernel density estimation, pattern recognition, medical and healthcare data learning**

## I. Introduction

In recent years, the imbalanced learning [2] has drawn a significant interest in numerous scientific areas, such as marketing, data mining, and bio-medical science. The imbalanced data indicates that there is a large difference on the sizes of data for different classes. Most standard learning algorithms have been proposed for balanced data without considering the issue of imbalanced data. For imbalanced data, these classic machine learning algorithms always fail to properly represent the data distribution and thus perform poorly for learning. More specifically, because the data distributions of minority class are always under-represented in these standard algorithms, the decision rules of these standard classifiers are very weaker for the minority class than the majority class. In practice, the minority class with limited data is more interesting and expensive than the majority class. One representative example is the well-known cancerous patients predication problem: there are more "Healthy" data samples than "Cancerous" data samples, such as the Mammography dataset [3] in which there are $10,923$ "Healthy" data samples and only $260$ "Cancerous" data samples. We expect that there is low error rate to predict a cancerous patient to be a healthy people, because it is always much more costly than classifying a healthy people as cancerous in medical industry. However, for this or the other more extremely imbalanced data set, the standard classifiers definitely provide imbalanced classification performance, with close to a $100$ percent classification accuracy for the majority class and with a near $0$ percent classification accuracy for the minority class. Thus, we require a classifier to have balanced classification performance even for an imbalanced data set, i.e., it should provide good enough accuracy for the minority class data without severely losing the accuracy of the majority class data.

The fundamental difficulty of these standard classification algorithms for imbalanced data is hardly to represent the distribution of the minority class data because of the domination of the majority class data. To solve this deficiency in these standard classification algorithms, in last decades researchers have proposed to incorporate sampling (e.g., oversampling and undersampling), cost-sensitive, and kernel-based learning into these algorithms to improve classification accuracy of the minority class. A comprehensive and systematic survey on imbalanced learning can be found in [2].

Generally speaking, sampling method operates the original data sets and aims to balance the data distribution using various under-sampling and over-sampling mechanisms. For instance, the random under-sampling approach is one of the most simple and straightforward techniques where only a set of majority class data is randomly selected to train classifiers with the original minority class data. A similar idea to the random under-sampling approach is the random over-sampling approach where a set of the minority class data is randomly selected and replicated to train classifiers with the original majority class data. While those under-sampling and over-sampling techniques can somehow address the issues of uneven data distribution in imbalanced data sets, several problems are also raised to hinder the learning [4][5]. Instead of using random sampling methods, the informed under-sampling methods and the synthetic over-sampling with data generation methods are more powerful for learning [6][7][8]. EasyEnsemble and BalanceCascade are two examples of the informed under-sampling method that have been proposed to address the issue of information loss in random under-sampling methods [6]. The synthetic minority over-sampling technique (SMOTE) is a well-known informed over-sampling method which aims to generate synthetic minority class data in feature space to balance the imbalanced data sets [8]. A extension of SMOTE, named SMOTEBoost, is proposed in [9] where SMOTE is integrated with adaptive boosting algorithms to update the weights of minority class data. In our previous work [1], ADASYN, an adaptive synthetic sampling approach, is proposed for imbalanced learning using a weighted distribu-

B. Tang and H. He. are with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA. Email: {btang, he}@ele.uri.edu

978-1-4799-7492-4/15/$31.00 

tion to strengthen the decision boundary toward the minority class data. In [10], kernel density estimation methods have been applied to estimate the probability density distribution of minority class to sample additional minority data samples.

Cost-sensitive learning method uses cost-matrix to represent the misclassification cost of the minority and majority class data. In other words, if the misclassification of the minority class data is more expensive than the majority class data, a higher value for the minority class is assigned in cost-matrix. A general framework of the cost-sensitive learning, named Metacost, is proposed in [11]. In [12], the cost-sensitive learning is applied to neural network models for imbalanced learning, in which the output threshold is adjusted for the expensive minority class data, thereby improving the classification accuracy for the minority class.

Kernel-based method studies kernel classifiers for imbalanced learning problems. Support vector machine (SVM), the representative kernel-based learning approach, incorporating with sampling and ensemble methods has been well-studied [13] [14] [15]. In [16], a kernel-boundary alignment (KBA) algorithm is proposed by modifying the kernel matrix according to the imbalanced data distribution. In [17], a kernel classifier is constructed on the basis of orthogonal forward selection (OFS) to optimize the model generalization for learning.

The imbalanced learning methods actually provide strategies to adjust the learning process of a base classifier. Hence, the classification performance of an imbalanced learning method relates to the learning capacity of base classifiers. In addition to the classic classifiers, such as support vector machine [18], neural network [19], etc., many other powerful classification methods have also been proposed recently [20][21][22][23][24]. To address the "curse of dimensionality" in high-dimensional data spaces, feature reduction methods are usually applied to reduce the influence of irrelevant dimensions [25][26][27][28][29][30][31].

In this paper, we propose a kernel based adaptive synthetic data generation method for imbalanced learning. We first apply kernel density estimation (KDE) method to estimate the density distribution of minority class which can be considered as a probability density distribution used to sample new minority class data. Compared to the existing synthetic minority class data generation methods, such as SMOTE, the synthetic data generated in our probability density estimation method can be approximately considered as the distribution of the original minority class data. To strengthen the decision boundary toward the difficult-to-learn minority class data, we adopt our previous idea of ADASYN approach. The estimated distribution with KDE method is weighted according to the difficulty level of the minority class data. Via simulations and experiments on real-life data sets, we demonstrate the effectiveness of our proposed kernel based adaptive over-sampling approach on several real-life benchmarks which have various imbalanced data distributions.

The rest of this paper is organized as follows: In Section II, we present our kernel based adaptive over-sampling method based on kernel density estimation in detail. In Section III, we test the proposed adaptive over-sampling method incorporated into naive Bayes and decision tree classifiers on various real-life imbalanced data sets. Finally, a discussion and conclusion are provided in section IV.

## II. Proposed Algorithm

Motivated by the success of data generation methods, including SMOTE [8], SMOTEBoost [9], and ADASYN [1], we propose KernelADASYN over-sampling approach, an adaptive over-sampling method based on kernel density estimation to address the problem in imbalanced learning. We first apply kernel density estimation methods to estimate the probability density distribution of the minority class which provides us a reliable sampling distribution to generate new minority class data. To adaptively shift the decision boundary toward the difficult minority class data, we adopt the idea of ADASYN to weight the sampling distribution of minority class. For any imbalanced learning problem, suppose that we have training data set $D$ with $N$ data samples $\{\mathbf{x}_i, y_i\}$, $i = 1, 2, \cdots, N$, where $\mathbf{x}_i$ is the $i$-th data sample in $D$ and $y_i = \{+1, -1\}$ is the class membership associated with $\mathbf{x}_i$. Defining the number of minority and majority class data as $N^+$ and $N^-$, respectively, we have $N^+ < N^-$ and $N^- + N^+ = N$. The goal of our method is to generate informed minority class data, i.e., $N^+$ is increased, to improve overall prediction performance.

### A. Kernel Density Estimation

The estimation of probability density function (PDF) from a set of observed data has been well studied in statistics and machine learning. Here, we use kernel estimator to estimate the minority class's PDF. Given the minority class data set $\{\mathbf{x}_i, y_i\}$, $i \in I_{+1}$, where $I_{+1}$ is the index set of minority class and $y_i = +1$ for each $i \in I_{+1}$, the general kernel estimator with kernel $K$ is defined by

$$\hat{p}(\mathbf{x}) = \frac{1}{N^+ h} \sum_{i \in I_{+1}} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{1}$$

where $h$ is the bandwidth of each kernel, a smoothing parameter in kernel estimator. The kernel function $K$ is non-negative in feature space $\mathcal{X}$ and satisfies the following condition

$$\int_{\mathbf{x} \in \mathcal{X}} K(\mathbf{x}) d\mathbf{x} = 1 \tag{2}$$

Such kernel $K$ indicates that the estimated density estimation $\hat{p}(\mathbf{x})$ in Eq. (1) is a probability density. In our present work, we use Gaussian kernel to estimate the density distribution of the minority class which is given by

$$\begin{aligned} G_i(\mathbf{x}) &= K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \\ &= \frac{1}{(\sqrt{2\pi}h)^n} \exp\left(-\frac{1}{2}\frac{(\mathbf{x} - \mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i)}{h}\right) \end{aligned} \tag{3}$$

Hence, the estimated PDF $\hat{p}(\mathbf{x})$ is written as

$$\hat{p}(\mathbf{x}) = \frac{1}{N^+ h} \sum_{i \in I_{+1}} G_i(\mathbf{x}) \tag{4}$$

Because we have

$$\sum_{i \in I_{+1}} \frac{1}{N^+} = 1 \tag{5}$$

the PDF estimation in Eq. (4) from minority class data can be interpreted as Gaussian mixture model with $N^+$ components. Here, each component has the same weight $1/N^+$ and the same standard variance $h$. According to the estimated probability distribution, new minority class data can be easily sampled.

### *B. Adaptive Synthetic Over-Sampling Distribution*

Notice that Eq. (4) can be used as a basic sampling distribution to generate new minority class data using kernel density estimation. However, note that it assigns the same weight for each minority class data, which may hinder the discriminative capacity for minority class data. Instead of using an uniform distribution over all minority class data, we extend the ADASYN method and measure the distribution of weights for different minority class data according to their difficulty level of learning [1].

For each minority class data $\mathbf{x}_i$, $i \in I_{+1}$, we find its $k$ nearest neighbors based on the Euclidean distance and count the number of majority class data in its $k$ nearest neighbors as $k^+$. We calculate the ratio $r_i$ for each minority class as

$$r_i = \frac{k^+}{k}, \quad i \in I_{+1} \tag{6}$$

To ensure the $r_i$ to be a density distribution, we normalize the ratio according to

$$\hat{r}_i = \frac{r_i}{\sum_{i \in I_{+1}} r_i} \tag{7}$$

Hence, we use the following weighted kernel density estimation as over-sampling distribution for minority class data generation

$$\hat{p}(\mathbf{x}) = \frac{1}{N^+ h} \sum_{i \in I_{+1}} \hat{r}_i G_i(\mathbf{x}) \tag{8}$$

We present our proposed algorithm in detail in Algorithm 1. According to the sampling distribution in Eq. (8), we sample new minority class data as $D_t = \{\mathbf{x}_i, +1\}$, $i = 1, 2, \cdots, \Delta$. $\Delta$ is the number of new minority class data, which is determined by balanced coefficient $\beta$ according to Eq. (9). Similar to ADASYN, Algorithm 1 not only provides a balanced representation of data distribution, but also helps to shift decision boundary of classifier toward these difficult-to-learn minority class data. Unlike the ADASYN, however, our new algorithm samples the new minority class data according to a probability distribution. In such manner, the synthetic data can be considered to be sampled by the same distribution as the minority class data observed in the original training data set. Compared to the SMOTEBoost and DataBoost-IM which update the distribution function with performance evaluation, the classifier is more efficient as it is based on data distribution characteristics [1].

Fig. 1 illustrates our adaptive over-sampling process for an imbalanced data set and the synthetic minority class data using our approach compared with SMOTE and ADASYN over-sampling approaches. In this imbalanced data set, the majority class data satisfy single Gaussian distribution, while the minority class data satisfy Gaussian mixture distribution with two components. More specifically, we have $N^+ = 100$ minority class data and $N^- = 500$ majority class data. Fig. 1 shows the significance of our approach in two-fold: firstly, the synthetic minority class data are sampled with a probability distribution; secondly, the synthetic minority class data are mainly sampled around decision boundary and force the classifier focus on those difficult-to-learn minority class data. Combing with the synthetic minority class data generated by three individual over-sampling approaches, we train three decision-tree classifiers and use 200 testing data to evaluate the classification performance. Fig. 2 shows the classification error for different $\beta$ values using our proposed approach in comparison to SMOTE and ADASYN approach. These results are averaged over 100 runs. In Fig. 2, $\beta = 0$ means no additional synthetic minority class data is sampled, and $\beta = 1$ means $N^- - N^+$ synthetic minority class data are sampled and hence the new data set is fully balanced. We further use naive Bayes classifier as base classifier to evaluate performance in Fig. 3. Both Fig. 2 and Fig. 3 demonstrate the performance improvement of our proposed over-sampling approach.

**Algorithm 1:** KernelADASYN Algorithm

**Input**:
(1) Training data set $D$ with $N$ data samples $\{\mathbf{x}_i, y_i\}$, $i = 1, 2, \cdots, N$, where $\mathbf{x}_i$ is the $i$-th data sample in $D$ and $y_i = \{1, -1\}$ is the class membership associated with $\mathbf{x}_i$. $I_{+1}$ and $I_{-1}$ are the index set of the minority and majority class data, respectively, in data set $D$;
(3) $k$: Number of nearest neighbors under consideration;
(4) $h$: Bandwidth of Gaussian kernel function;
**Procedure**:
(1) Calculate the number of synthetic minority class data according to

$$\Delta = (N^- - N^+) \times \beta \tag{9}$$

where $\beta \in [0,\ 1]$ specifies the balance level with new synthetic minority class data. $\beta = 0$ means no additional minority class data is needed, and $\beta = 1$ means there is the same number of minority class data and majority class data after this generation procedure;
(2) For each minority class data $\mathbf{x}_i$, $i \in I_{+1}$, calculate the ratio $r_i$ according to Eq. (6) and normalize it according to Eq. (7);
(3) Build the weighted over-sampling distribution with kernel density estimation $\hat{p}_{\mathbf{x}}$:

$$\hat{p}(\mathbf{x}) = \frac{1}{N^+ h} \sum_{i \in I_{+1}} \hat{r}_i \frac{1}{(\sqrt{2\pi} h)^n} \exp\left(-\frac{1}{2} \frac{|\mathbf{x} - \mathbf{x}_i|}{h}\right) \tag{10}$$

where $|\mathbf{x}| = \mathbf{x}^T \mathbf{x}$ measures the distance of two feature vectors;
(4) Sample $\Delta$ minority class data $D_t = \{\mathbf{x}_i, +1\}$, $i = 1, 2, \cdots, \Delta$, according to the distribution $\hat{p}_{\mathbf{x}}$ and append them into training data set $D' = D \cup D_t$;
**Output**: Training data set $D'$ with synthetic minority class data.

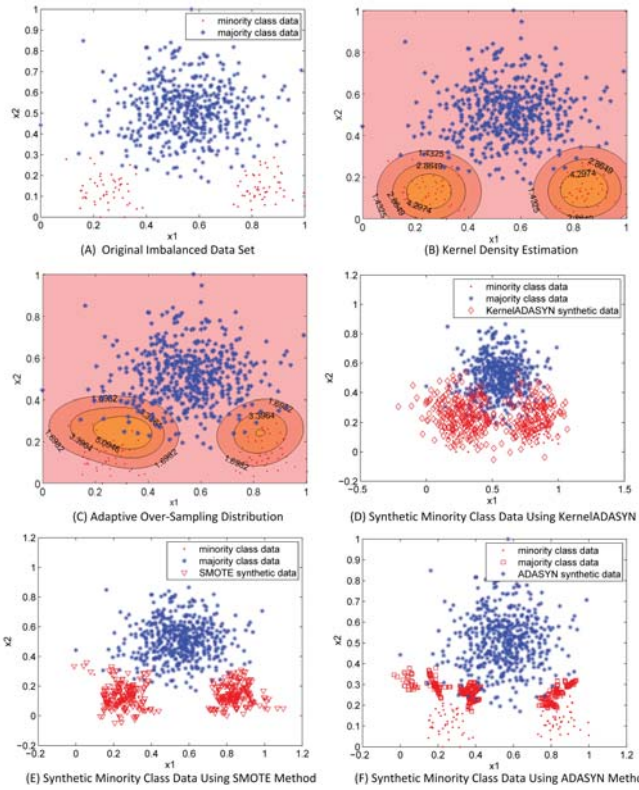


Fig. 1: The process of our KernelADASYN approach in comparison to SMOTE and ADASYN. (A). The original two-class imbalanced data set; (B). Kernel-based density estimation from minority class data. The contour values indicate the estimated probability of minority class using kernel density estimation method; (C). Adaptive over-sampling distribution, a probability distribution weighted from kernel-based density estimation in (B); (D). A fully balanced data set after data generation using KernelADASYN; (E). Data generation using SMOTE; (F). Data generation using ADASYN.

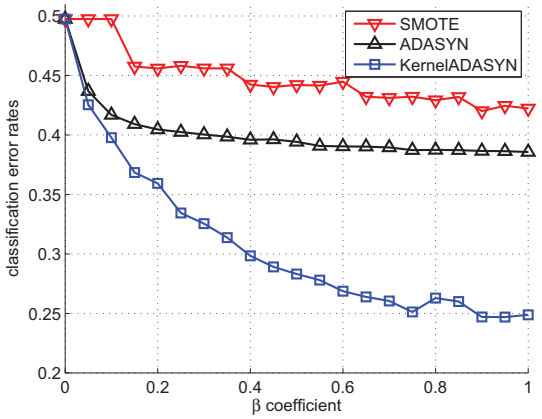


Fig. 2: The performance comparison for different coefficients $\beta$ using decision tree as base classifier.

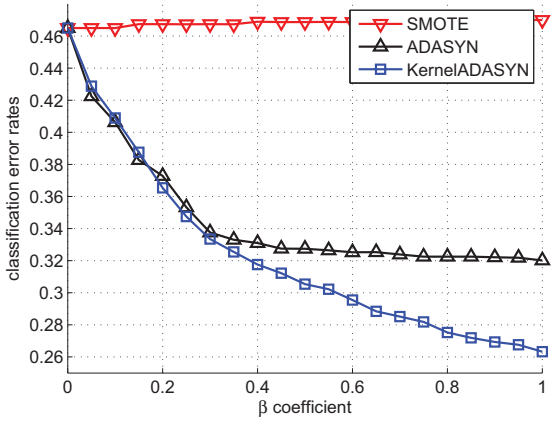


Fig. 3: The performance comparison for different coefficients $\beta$ using naive Bayes as base classifier.

## III. Experiments and Analysis

### A. Evaluation Metrics

Because of the special attention on the minority class data, it is inappropriate to only use *accuracy* to evaluate the performance for imbalanced learning. For instance, in an imbalanced data set where the minority class data make up $1\%$ of the whole dataset, one can obtain the accurancy of $99\%$ if one simply classifies all the test data as the majority class. Therefore, the accuracy metric may be misleading in imbalanced learning as the misclassification cost for minority and majority class data is different. To address this issue in imbalanced learning, researchers have further proposed a number of alternate metrics to fairly assess the performance. In our experiments, we use the *accuracy*, *precision*, *recall*, *F-Measure*, and *G-Mean*. Considering the two-class classification problem, let the minority class as the "positive" class and the majority class as the "negative" class. These metrics can be defined as followings:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{11}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$\text{F-Measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{14}$$

$$\text{G-Mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \times \frac{\text{TN}}{\text{TN} + \text{FP}}} \tag{15}$$

where TP (true positive) means the number of positive data that are classified correctly, TN (true negative) means the number of positive data that are classified correctly, FP (false positive) means the number of negative data that are misclassified as positive, and FN (false negative) means the number of positive data that are misclassified as negative.

From above equations, one can see that the recall measure indicates how often a positive class data is correctly predicted as positive class data, and the precision measure indicates how often a data that is predicted as a positive class is actually positive [32]. Hence, the precision actually measures the exactness, while recall measures the completeness [2]. The goal of imbalanced learning is to improve the recall performance without hurting the precision performance. However, this goal is usually conflicting. F-Measure combines the recall and precision to provide a measurement of functionality of a classifier [2]. G-Mean measure evaluates the degree of inductive bias with respect to a ratio of positive accuracy and negative accuracy. We also note that there are other metrics existing for performance evaluation in imbalanced learning, such as ROC [33] and H-Measurement [34]. ROC curves assess the performance by examining the relation between TP rate and FP rate. ROC analysis is established by adjusting a classification threshold to obtain different TP rate and FP rate.

### B. Real-Life Data Sets

We study the performance of our proposed algorithm on seven real-world datasets, specifically on medical and healthcare applications, including Pima Indians diabetes dataset from National Institute of Diabetes and Digestive and Kidney Diseases [35], Indian liver patient dataset (ILPD) [36], Parkinsons dataset [37], Vertebral Column dataset [38], breast cancer dataset [39], breast tissue dataset [40] and SPECT heart dataset [41]. The characteristics of these datasets are summarized in Table I. The data set *Pima Indians Diabetes* aims to predict whether the patient has diabetes or not. In this data set, all patients are all females at least 21 years old of Pima Indian heritage. With Indian Liver Patient Dataset (*ILPD*) the aim is to predict liver patient according to the age, gender, total proteins, etc. The data set *Parkinsons* contains the speech signal information to study general voice disorders. The main goal is to discriminate healthy people from those who have Parkinson's disease. The data set *Vertebral Column* has six biomechanical features used to classify orthopaedic patients into two classes (normal or abnormal). Both the data set of *Breast Cancer* and *Breast Tissue* aim to predict the patient is normal or abnormal according to the measurements. The data set *SPECT* contains cardiac Single Proton Emission Computed Tomography (SPECT) images and aims to classify each image to be normal or abnormal. All of these datasets are two-class dataset, and we consider the minority class data as positive data and majority class data as negative data. By defining *imbalance ratio* (Imb. Rat.) as the ratio of the number of positive class data to the number of negative class data, Table I also shows that the imbalanced ratio of these datasets ranges from $0.25$ to $0.54$.

### C. Experimental Results and Analysis

We test the classification performance of our proposed KernelADASYN algorithm for these seven real-life imbalanced data sets with decision tree and naive Bayes as base classifiers. We present the performance of our proposed algorithm with the assessment metrics including accuracy, recall, precision, F-Measure and G-Mean, compared to the SMOTE and the ADASYN algorithms in Table II and Table III, using the decision tree and the naive Bayes classifier as the base classifier, respectively. We also evaluate the performance of the decision tree and the naive Bayes classifiers trained from the original imbalanced data sets without introducing synthetic minority class data. The results in Table II and Table III are averaged over 100 runs. At each run, we randomly select half of the whole data set as the training data and use the remaining half as test data. For SMOTE, ADASYN and our proposed KernelADASYN algorithm, we set the number of nearest neighbors $k = 5$ and the coefficient $\beta = 1$ to form fully balanced data sets.

For each data set in Table II and Table III, we highlight the best performance in each performance evaluation metric. Moreover, we count the total number of winning time for each method across different evaluation metrics. Based on the comparison on these seven test benchmarks, our proposed oversampling method can achieve competitive classification performance compared to the SMOTE and ADASYN approaches. Considering the overall winning times for different evaluation metrics, our approach outperforms the other methods. More specifically, in terms of overall accuracy, F-Measure and G-Mean metrics, our approach can improve the classification performance for both minority and majority class data without hurting one class for the other.

## IV. Conclusion

In this paper, we developed a new adaptive over-sampling algorithm, KernelADASYN, based on kernel density estimation for imbalanced data classification problems. The adaptive over-sampling approach can sample synthetic minority class data with a constructed probability distribution. Moreover, by adjusting the weights for different minority class data, it can force the classifier to focus on these difficult-to-learn minority class data, thereby improving the classification performance. Experimental results on artificial data set and real-life imbalanced data classification problems, specifically on medical and healthcare applications, with various performance evaluation metrics show the robustness and the effectiveness of this method.

## V. Acknowledgement

This research was partially supported by National Science Foundation (NSF) under grant ECCS 1053717 and CCF 1439011, and the Army Research Office under grant W911NF-12-1-0378.

## References

[1] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[3] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process," *Medical Physics*, vol. 34, no. 11, pp. 4164–4172, 2007.

TABLE I: The characteristics of seven real-life imbalanced data sets

| Dataset | # of features | # of data | minority class | majority class | # of minority class data | # of majority class data |
|---|---|---|---|---|---|---|
| *Pima Indians Diabetes* | 8 | 768 | tested positive | tested negative | 268 | 500 |
| *ILPD* | 10 | 583 | liver patient | normal | 167 | 416 |
| *Parkinsons* | 22 | 195 | abnormal | normal | 48 | 147 |
| *Vertebral Column* | 6 | 310 | abnormal | normal | 100 | 210 |
| *Breast Cancer* | 9 | 699 | malignant | benign | 241 | 458 |
| *Breast Tissue* | 9 | 699 | carcinoma | others | 21 | 85 |
| *SPECT* | 9 | 699 | heart abnormal | heart normal | 55 | 212 |

TABLE II: Evaluation Metrics and Performance Comparison Using Decision Tree as Base Classifier.

| Datasets | Methods | OA | Precision | Recall | F-Measure | G-Mean |
|---|---|---|---|---|---|---|
| *Pima Indians Diabetes* | Decision Tree | 0.7049 | **0.5803** | 0.5642 | 0.5696 | 0.6612 |
| | SMOTE | 0.6750 | 0.5284 | **0.6627** | 0.5867 | 0.6707 |
| | ADASYN | 0.6799 | 0.5374 | 0.5881 | 0.5606 | 0.6535 |
| | KernelADASYN | **0.7089** | 0.5731 | 0.6560 | **0.6112** | **0.6950** |
| *ILPD* | Decision Tree | 0.6515 | 0.3858 | 0.3687 | 0.3741 | 0.5274 |
| | SMOTE | 0.6326 | 0.3797 | 0.4518 | 0.4116 | 0.5629 |
| | ADASYN | 0.6357 | 0.3841 | 0.4542 | 0.4139 | 0.5642 |
| | KernelADASYN | **0.6571** | **0.4004** | **0.4598** | **0.4187** | **0.5722** |
| *Parkinsons* | Decision Tree | 0.8237 | 0.6611 | 0.6417 | 0.6398 | 0.7478 |
| | SMOTE | 0.8361 | 0.6627 | **0.7500** | 0.6938 | 0.8012 |
| | ADASYN | 0.8216 | 0.6456 | 0.6958 | 0.6611 | 0.7723 |
| | KernelADASYN | **0.8433** | **0.6672** | **0.7500** | **0.7000** | **0.8059** |
| *Vertebral Column* | Decision Tree | 0.7903 | **0.6711** | 0.7040 | 0.6837 | 0.7631 |
| | SMOTE | 0.7845 | 0.6491 | 0.7480 | 0.6910 | 0.7720 |
| | ADASYN | **0.8013** | 0.6703 | 0.7800 | **0.7195** | 0.7944 |
| | KernelADASYN | 0.7935 | 0.6559 | **0.7820** | 0.7194 | **0.7984** |
| *Breast Cancer* | Decision Tree | 0.9367 | 0.9117 | 0.9058 | 0.9077 | 0.9287 |
| | SMOTE | 0.9476 | **0.9171** | 0.9333 | 0.9244 | 0.9439 |
| | ADASYN | 0.9401 | 0.9080 | 0.9200 | 0.9133 | 0.9349 |
| | KernelADASYN | **0.9481** | 0.9115 | **0.9408** | **0.9258** | **0.9463** |
| *Breast Tissue* | Decision Tree | 0.8962 | 0.7179 | **0.8600** | 0.7644 | 0.8766 |
| | SMOTE | 0.9115 | **0.8530** | 0.6900 | 0.7391 | 0.8074 |
| | ADASYN | **0.9288** | 0.8172 | 0.8400 | **0.8196** | **0.8907** |
| | KernelADASYN | 0.8788 | 0.7313 | 0.7200 | 0.7038 | 0.8070 |
| *SPECT* | Decision Tree | 0.8143 | 0.5605 | 0.5037 | 0.5270 | 0.6695 |
| | SMOTE | 0.7820 | 0.4759 | 0.5593 | 0.5105 | 0.6830 |
| | ADASYN | 0.7902 | 0.4978 | **0.5595** | 0.5227 | **0.6880** |
| | KernelADASYN | **0.8174** | **0.5513** | 0.5185 | **0.5280** | 0.6755 |
| *Winning Time* | Decision Tree | 0 | 2 | 1 | 0 | 0 |
| | SMOTE | 0 | 2 | 2 | 0 | 0 |
| | ADASYN | 2 | 0 | 1 | 2 | 2 |
| | KernelADASYN | **5** | **3** | **4** | **5** | **5** |


[4] D. Mease, A. J. Wyner, and A. Buja, "Boosted classification trees and class probability/quantile estimation," *The Journal of Machine Learning Research*, vol. 8, pp. 409–439, 2007.

[5] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on Learning from Imbalanced Datasets II*, vol. 11, 2003.

[6] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[7] S. Chen, H. He, and E. A. Garcia, "RAMOBoost: Ranked minority oversampling in boosting," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1624–1642, 2010.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTE-Boost: Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases: PKDD 2003*, pp. 107–119, 2003.

[10] M. Gao, X. Hong, S. Chen, C. J. Harris, and E. Khalaf, "PDFOS: PDF estimation based over-sampling for imbalanced two-class problems," *Neurocomputing*, vol. 138, pp. 248–259, 2014.

[11] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155–164, 1999.

[12] 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.

[13] F. Vilariño, P. Spyridonos, J. Vitrià, and P. Radeva, "Experiments with SVM and stratified sampling with an imbalanced problem: Detection of intestinal contractions," in *Pattern Recognition and Image Analysis*, pp. 783–791, 2005.

[14] P. Kang and S. Cho, "EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems," in *Neural Information Processing*, pp. 837–

TABLE III: Evaluation Metrics and Performance Comparison Using Naive Bayes as Base Classifier.

| Datasets | Methods | OA | Precision | Recall | F-Measure | G-Mean |
|---|---|---|---|---|---|---|
| *Pima Indians Diabetes* | Naive Bayes | 0.7346 | 0.6055 | 0.6928 | 0.6455 | 0.7236 |
| | SMOTE | 0.7361 | **0.6086** | 0.6896 | 0.6457 | 0.7238 |
| | ADASYN | 0.7259 | 0.5893 | **0.7152** | 0.6453 | 0.7227 |
| | KernelADASYN | **0.7372** | 0.5985 | 0.7039 | **0.6463** | **0.7241** |
| *ILPD* | Naive Bayes | 0.6086 | 0.4039 | 0.7831 | 0.5319 | 0.6473 |
| | SMOTE | 0.6041 | 0.4085 | 0.8614 | 0.5535 | 0.6554 |
| | ADASYN | 0.6017 | 0.4081 | **0.8711** | **0.5549** | 0.6537 |
| | KernelADASYN | **0.6196** | **0.4142** | 0.8000 | 0.5442 | **0.6585** |
| *Parkinsons* | Naive Bayes | 0.7412 | 0.4885 | 0.8833 | **0.6281** | 0.7822 |
| | SMOTE | 0.7196 | 0.4656 | 0.8792 | 0.6081 | 0.7650 |
| | ADASYN | 0.7165 | 0.4632 | **0.8958** | 0.6099 | 0.7666 |
| | KernelADASYN | **0.7773** | **0.5381** | 0.7542 | 0.6271 | **0.7878** |
| *Vertebral Column* | Naive Bayes | 0.7290 | 0.5521 | 0.8520 | 0.6695 | 0.7551 |
| | SMOTE | 0.7277 | 0.5500 | 0.8640 | 0.6716 | 0.7561 |
| | ADASYN | 0.7252 | 0.5469 | **0.8720** | 0.6717 | 0.7552 |
| | KernelADASYN | **0.7310** | **0.5532** | 0.8660 | **0.6747** | **0.7592** |
| *Breast Cancer* | Naive Bayes | 0.9553 | 0.9496 | 0.9192 | 0.9339 | 0.9462 |
| | SMOTE | 0.9585 | **0.9516** | 0.9267 | 0.9388 | 0.9505 |
| | ADASYN | **0.9670** | 0.9457 | **0.9600** | **0.9525** | **0.9653** |
| | KernelADASYN | 0.9630 | 0.9493 | 0.9433 | 0.9460 | 0.9581 |
| *Breast Tissue* | Naive Bayes | **0.8481** | **0.5723** | 0.9000 | 0.6961 | 0.8654 |
| | SMOTE | 0.8327 | 0.5708 | 0.9100 | 0.6879 | 0.8562 |
| | ADASYN | 0.8058 | 0.5169 | **0.9500** | 0.6617 | 0.8522 |
| | KernelADASYN | 0.8250 | 0.5490 | **0.9500** | **0.6953** | **0.8709** |
| *SPECT* | Naive Bayes | 0.7353 | 0.4241 | 0.8370 | 0.5619 | 0.7693 |
| | SMOTE | 0.7218 | 0.4132 | **0.8444** | 0.5534 | 0.7622 |
| | ADASYN | 0.6797 | 0.3735 | **0.8444** | 0.5168 | 0.7321 |
| | KernelADASYN | **0.7511** | **0.4415** | 0.8333 | **0.5762** | **0.7790** |
| *Winning Time* | Naive Bayes | 1 | 1 | 0 | 1 | 0 |
| | SMOTE | 0 | 2 | 1 | 0 | 0 |
| | ADASYN | 1 | 0 | **7** | 2 | 1 |
| | KernelADASYN | **5** | **4** | 1 | **4** | **6** |


846, 2006.

[15] Y. Liu, A. An, and X. Huang, "Boosting prediction accuracy on imbalanced datasets with SVM ensembles," in *Advances in Knowledge Discovery and Data Mining*, pp. 107–118, 2006.

[16] 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.

[17] 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.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 4. Springer New York.

[20] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] B. Tang, H. He, Q. Ding, and S. Kay, "A parametric classification rule based on the exponentially embedded family," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 367–377, Feb 2015.

[23] B. Tang, Q. Ding, H. He, and S. Kay, "Hybrid classification with partial models," in *IEEE International Joint Conference on Neural Networks*, pp. 3726–3731, July 2014.

[24] B. Tang and H. He, "ENN: Extended nearest neighbor method for pattern recognition," *IEEE Computational Intelligence Magazine*, 2015 in press.

[25] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[26] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[27] Y. Xia, H. Tong, W. Li, and L.-X. Zhu, "An adaptive estimation of dimension reduction space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 363–410, 2002.

[28] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613, 1998.

[29] J. Xu, G. Yang, Y. Yin, H. Man, and H. He, "Sparse-representation-based classification with structure-preserving dimension reduction," *Cognitive Computation*, vol. 6, no. 3, pp. 608–621, 2014.

[30] J. Xu, Y. Yin, H. Man, and H. He, "Feature selection based on sparse imputation," in *IEEE International Joint Conference on Neural Networks*, pp. 1–7, 2012.

[31] J. Xu, G. Yang, H. Man, and H. He, "L1 graph based on sparse coding for feature selection," in *Advances in Neural Networks*, pp. 594–601, 2013.

[32] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, 2013.

[33] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, pp. 1–38, 2004.

[34] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine learning*, vol. 77, no. 1, pp. 103–123, 2009.

[35] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 261, American Medical Informatics Association, 1988.

[36] B. V. Ramana, M. S. P. Babu, and N. Venkateswarlu, "A critical comparative study of liver patients from USA and India: an exploratory analysis," 2012.

[37] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties

for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.
[38] E. Berthonnaud, J. Dimnet, P. Roussouly, and H. Labelle, "Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters," *Journal of spinal disorders & techniques*, vol. 18, no. 1, pp. 40–47, 2005.
[39] R. MICHALSKI, "The multi-purpose incremental learning system AQ15 and its testing application to three medical domains," *Proceedings. of AAAI-86, Morgan Kaufmann*, pp. 1041–1045, 1986.
[40] J. Jossinet, "Variability of impedivity in normal and pathological breast tissue," *Medical and Biological Engineering and Computing*, vol. 34, no. 5, pp. 346–350, 1996.
[41] L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. S. Goodenday, "Knowledge discovery approach to automated cardiac SPECT diagnosis," *Artificial intelligence in medicine*, vol. 23, no. 2, pp. 149–169, 2001.