Data Oversampling with Structure Preserving Variational Learning

Indu Solomon
International Institute of Information Technology Bangalore,
Bengaluru, India
indu.solomon@iiitb.ac.in

Senthilnath Jayavelu
Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR)
Singapore
J_Senthilnath@i2r.a-star.edu.sg

Md Meftahul Ferdaus
Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR)
Singapore
ferdaus_meftahul@i2r.a-star.edu.sg

Uttam Kumar
International Institute of Information Technology Bangalore,
Bengaluru, India
uttam@iiitb.ac.in

ABSTRACT

Traditional oversampling methods are well explored for binary and multi-class imbalanced datasets. In most cases, the data space is adapted for oversampling the imbalanced classes. It leads to various issues like poor modelling of the structure of the data, resulting in data overlapping between minority and majority classes that lead to poor classification performance of minority class(es). To overcome these limitations, we propose a novel data oversampling architecture called Structure Preserving Variational Learning (SPVL). This technique captures an uncorrelated distribution among classes in the latent space using an encoder-decoder framework. Hence, minority samples are generated in the latent space, preserving the structure of the data distribution. The improved latent space distribution (oversampled training data) is evaluated by training an MLP classifier and testing with unseen test dataset. The proposed SPVL method is applied to various benchmark datasets with i) binary and multi-class imbalance data, ii) high-dimensional data and, iii) large or small-scale data. Extensive experimental results demonstrated that the proposed SPVL technique outperforms the state-of-the-art counterparts.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

class imbalance, latent space, structure preserving, oversampling, classification

ACM Reference Format:
Indu Solomon, Senthilnath Jayavelu, Md Meftahul Ferdaus, and Uttam Kumar. 2022. Data Oversampling with Structure Preserving Variational Learning. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557575

https://doi.org/10.1145/3511808.3557575

1 INTRODUCTION

Class imbalance issue is a chronic disorder which is concerning the machine learning (ML) community and this problem arises when all classes in a dataset do not have an equal number of training samples. In most class imbalance problems, the class of interest is the minority class(es). However, most of the ML algorithms perform poorly in classifying minority class(es) due to their inductive bias towards the majority class. Class imbalance problem occurs regularly in real-world applications such as credit fraud detection, semiconductor defect detection and medical diagnosis [6]. In some cases, incorrect decision-boundary in ML classifiers may cause catastrophic consequences such as misclassifying faulty devices as healthy or unfit patients as fit.

To tackle the class imbalance problem, there exists in literature, data-level [2, 4, 9, 10], algorithm-level methods [3, 16], and ensemble learning methods [1, 15, 18]. In algorithm-level methods, the classifiers are punished more heavily for the wrong classification of the minority class than the majority class. In data-level methods, data are either re-sampled randomly from the original data or samples are synthetically generated using oversampling techniques. Another oversampling technique, structure preserving oversampling (SPO) proposed in [2] generates synthetic samples into the void area in the data space rather than the area closely tied to the minority class samples. These techniques go hand in hand with classical ML models and tabular data. However, for the deep architectures designed for vision/speech/text data these oversampling techniques fail to generate meaningful samples [8].

In recent times, latent space generative models like generative adversarial networks (GANs) [7], and variational autoencoders (VAEs) [11] are successfully used to extract useful information from high dimensional data. These deep learning-based generative models can generate new samples which are close to original data.

However, when these models are used directly as an oversampling technique in class imbalance problems, they suffer from different challenges, namely, mode collapse, blurred generated images and training instability. A hybrid of discriminative VAE (DVAE) and GAN called discriminative variational autoencoding adversarial network (DVAAN) is proposed in [8] to deal with class imbalance problems. DVE model imposes a prior on latent two-component mixture distribution, thus the model depends upon the true labels in the generative phase. Moreover, the assumption of two component latent distribution makes it suitable for binary data imbalance. In BAGAN [13], the generative model learns useful features from majority classes and uses these to generate images for minority classes. The generator in the GAN is initialized with the encoder module of an autoencoder that enables learning an accurate class-conditioning in the latent space [13]. The three-player adversarial game-based Generative Adversarial Minority Oversampling (GAMO) technique has been developed [14] to deal with class imbalance problems, where domain constraint convex generators were considered to handle mode collapse. However, GAMO exhibits some limitations, for instance, the generation of blurred images, and the generation of minority classes near decision boundaries that are difficult to classify [14]. DeepSMOTE [5] consists of an encoder-decoder framework, a SMOTE-based oversampling method, and a loss function with a reconstruction loss and a penalty term. DeepSMOTE backbone architecture is based on a DCGAN and is intended to address the class imbalance issue of vision data.

To overcome the above shortcomings, an encoder-decoder framework coupled with a multi-class structure preserving oversampling architecture is proposed in this paper. The proposed architecture produces a nearly distinct class distribution in the latent space. Thus the generated balanced dataset provides sufficient variability to attain a better-generalized model. Our experimental results demonstrate that the proposed method performs better than existing state-of-the-art methods across six selected datasets with moderate to very high imbalance. The main contributions of this paper are as follows:

- Structure preserving variational learning (SPVL) architecture, in which an encoder-decoder framework of VAE is coupled to a multi-class structure preserving oversampling (MSPO).
- A latent space oversampling strategy.
- The proposed architecture can be applied on both structured and unstructured datasets, where the datasets are of high dimension, high imbalance and large scale.

2 SPVL ARCHITECTURE

The proposed SPVL architecture captures an uncorrelated distribution among classes in the latent space utilizing an encoder-decoder framework of a VAE and applies an MSPO technique in the latent space for data oversampling. An MLP classifier in the latent space is trained for the evaluation of the oversampled data. Figure 1 shows the block diagram of the architecture.

Multi-class Latent Space Oversampling (MSPO).

In our proposed architecture, the data abstraction property of the encoder-decoder framework of VAE is used for bringing down the dimensionality of raw data. The setup assumes that the input

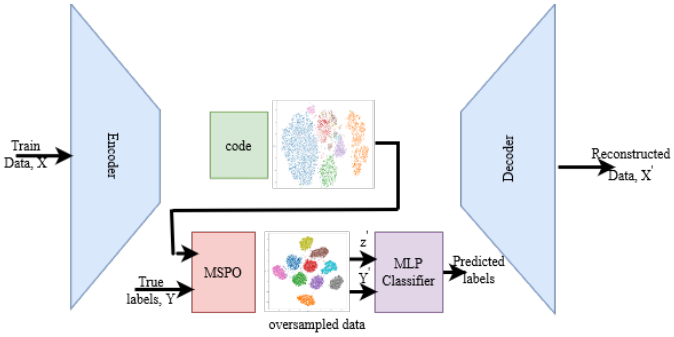


Figure 1: Block diagram of training phase of the proposed model

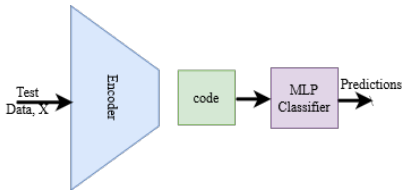


Figure 2: Testing phase of the proposed model

data X belong to a very high dimensional space $X \in \mathbb{R}^d$ and the data X is sampled from an unknown probability distribution $P(X)$.

The MPSO technique efficiently captures the covariance structure of the normally distributed latent space data and prevents synthetic sample invasion into other classes' space. Our architecture operates on the Eigen spectrum of the minority class(es) to perform multi-class oversampling.

Given the multi-class latent space data with k classes, let P_i be the number of samples in i^{th} minority class. and N is the number of samples in the majority classes. Total number of samples to be generated for i^{th} minority class is $(N - P_i)$. Sample covariance and Eigen decomposition for i^{th} minority class, W_{L_i} and $D_i = V_i^T W_{L_i} V_i$ are computed. Multiplying D_i equation on both sides by $D_i^{-1/2}$ gives,

$$I = (V_i D_i^{-1/2})^T W_{L_i} (V_i D_i^{-1/2}) = F_i^T W_{L_i} F_i \tag{1}$$

where $F_i = V_i D_i^{-1/2}$ is a scaled transformation for the i^{th} minority class and it has unit covariance structure. The property of unit covariance structure helps in generating synthetic minority samples by sampling from a zero mean unit variance distribution and mapping them back to the covariance structure of i^{th} minority class by inverse transforming with F_i^{-1}.

Distance criteria is used to make sure that the generated synthetic sample b_i is close to the i^{th} minority class samples. The latent space dataset is updated by adding the newly generated synthetic samples with the existing minority samples. Class labels are generated corresponding to i^{th} minority class and the class label vector Y is updated accordingly. A multilayer perceptron classifier performs the classification task. The classifier is trained with the oversampled dataset Z', and the class label Y', (Figure 1). In the testing phase of SPVL architecture, the unseen test data is fed to the trained encoder to transfer to the latent space and fed to the trained classifier for classification (Figure 2). The overall loss function for

the SPVL model is given by,

$$\mathcal{L}_{SPVL} = \lambda_1 \mathcal{L}_{ED} + \lambda_2 \mathcal{L}_{MSPO} + \lambda_3 \mathcal{L}_C \tag{2}$$

where,

$$\mathcal{L}_{ED}(\theta, \phi) = \sum_{i \in X} -\mathbb{E}_{z \sim Q_\theta(z|x_i)} [logP_\phi(x_i|z)] + KL(Q_\theta(z|x_i)||p(z)) \tag{3}$$

$$\mathcal{L}_{MSPO}(z_{Li}, z_{ON}, b_i) = \sum_i \mathfrak{D}(z_{Li}.b_i) - \sum_i \mathfrak{D}(z_{ON}, b_i) \tag{4}$$

$$\mathcal{L}_C(f_\theta, (Z', Y')) = -\mathbb{E}_{(Z'_i, Y'_i) \sim (Z', Y')} \sum_i \sum_{k=1}^{K} \mathbb{1}_{[k=y_i]} \log(f_\theta(Z'_i)) \tag{5}$$

In equation 3, the first term is a reconstruction loss and the second loss term acts as a regularizer. Reconstruction loss can be cross-entropy loss or squared error loss, and the selection of reconstruction loss is data-dependent. In equation 4, $\mathfrak{D}()$ is a distance function and it makes sure that the generated samples are closer to the i^{th} minority cluster and far from the other class(es)'s clusters. Equation 5 is cross-entropy loss.

Algorithm 1 SPVL

Input: X-training data (tabular/image)
Parameter: E-Encoder, D-Decoder, $MSPO$-Oversampling, C-Classifier, K-classes, Z-latent space samples, Y-class labels, Z'- oversampled latent data, Y'-updated labels.
Output: θ_E-trained encoder, θ_c-latent space classifier parameters.
Training:
  for N steps **do**
    for n_1 steps **do**
      $Z \leftarrow E(X)$
      $X' \leftarrow D(Z)$
      Update the weights of E and D
    end for
    $Z' \leftarrow MSPO(Z)$
    Update the samples of Z' using eq 4
    Updated class labels $\leftarrow Y'$
    for n_2 steps **do**
      $\hat{Y} \leftarrow C(Z', Y')$
      Update weights of C
    end for
  end for
Testing:
Input: X: unseen test data(tabular/image)
Output: $\hat{Y}$: predictions
  $Z \leftarrow E(X)$
  $\hat{Y} \leftarrow C(Z)$

3 EXPERIMENTAL RESULTS

We evaluate the performance of our proposed SPVL method with six benchmark datasets. Among six datasets, two datasets are image datasets and the remaining four are tabular datasets. The details are given in Table 1. Two of the tabular datasets are credit score datasets and the remaining two are semiconductor process datasets for fault detection.

Dataset	Type	Class	Features	IF
MNIST	Image	multi-class	784	0.96
Fashion-MNIST	Image	multi-class	784	0.96
German credit	Tabular	Binary	24	0.40
Australian credit	Tabular	Binary	14	0.11
Wafer	Tabular	Binary	152	0.97
Secom	Tabular	Binary	590	0.87

Table 1: Datasets selected for SPVL evaluation

The dataset imbalance factor (IF) is given by,

$$IF = 1 - \frac{k}{n_k} \min_j n_j \tag{6}$$

where k and n_k are the total number of classes and the total number of samples respectively, n_j is the total number of samples in class j. The imbalance factor (IF) varies between 0 (balanced) and 1(extreme imbalance).

Traditional and generative techniques selected for conducting this comparative study on the benchmark datasets are baseline classifier (Q), random oversampling + classifier (RO+Q), SMOTE + classifier (SMOTE +Q), borderline SMOTE + classifier (BSMOTE + Q), generative adversarial oversampling (GAMO) and Deep SMOTE. To ensure a fair comparison, all the selected techniques are run in the same environment. The classification performance of oversampled datasets is evaluated in terms of Average Class Specific Accuracy (ACSA) and Geometric Mean (GM). ACSA and GM for the dataset are given by,

$$ACSA = \frac{1}{k} \sum_{i=1}^{k} \frac{P_i}{T_i} \ \% \qquad GM = \sqrt[k]{\prod_{i=1}^{k} \frac{P_i}{T_i}} \ \% \tag{7}$$

where k is the number of classes, P_i is the correctly predicted samples of i^{th} class and T_i is the total number of samples in i^{th} class. Image datasets selected for the experiments are MNIST [12] and Fashion-MNIST [17], and they contain 60000 training images and 10000 test images and each image is of 28×28 dimensions. These are balanced multi-class datasets. However, the imbalance is introduced artificially by selecting disparate number of training samples from each class in such a way that the imbalance factor for the dataset is as high as 0.96. We have randomly selected {4000, 2000, 1000, 750, 500, 350, 200, 100, 60, 40} samples from the classes {0,1,2…,9} respectively. The test dataset for MNIST and Fashion-MNIST was formed by randomly selecting 100 samples from each {0,1,2…,9} class (train and test settings are same as GAMO's [14] experimental setup). The experimental results from Table 2 show that the SPVL algorithm has outperformed the selected traditional and generative variants of oversampling techniques. Figure 3 shows a comparison of t-SNE representations of imbalanced distribution and balanced distribution of GAMO, DeepSMOTE and SPVL oversampling techniques respectively. Figure 3d clearly shows that the SPVL technique is capable of capturing the class boundaries and preventing class overlaps. The generated images by SPVL oversampling technique for MNIST dataset shown in Figure 4 display good visual quality.

We have evaluated the performance of proposed SPVL model on four tabular datasets, out of four, two are credit score data and the remaining two are semiconductor process data. These tabular

	Q		RO+Q		SMOTE+Q		BSMOTE+Q		GAMO		Deep SMOTE		**SPVL(Ours)**	
Datasets	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM	ACSA	GM
German	65.4	64.5	65.3	64.9	68.4	68.3	66.9	66.9	70.1	69.8	73.7	73.6	**75.3**	**75.3**
Australian	84.8	84.5	86.2	86.2	87.4	87.4	87.1	87.1	87.6	87.6	87.7	87.7	**90.3**	**90.5**
Wafer	97.0	97.0	97.5	97.4	97.4	97.3	97.3	97.2	98.6	98.6	98.2	98.2	**99.1**	**99.1**
Secom	53.2	30.6	62.6	60.1	57.7	48.1	60.5	57.6	61.6	59.0	64.3	64.2	**68.1**	**67.9**
MNIST	88.0	87.3	88.3	87.0	89.0	87.9	88.7	88.0	88.9	87.9	89.1	88.2	**89.9**	**89.5**
FMNIST	81.3	77.5	80.4	77.7	80.9	79.3	80.3	77.9	80.5	78.6	80.2	76.9	**82.6**	**80.6**

Table 2: Performance Comparison of SPVL with other models for six benchmark datasets

datasets are inherently imbalanced with moderate to high imbalance factor. Comparison results for the tabular datasets from Table 2 show that the proposed SPVL performs better than the selected traditional and generative oversampling techniques.

Ablation study is conducted on MNIST (image) and wafer (tabular) datasets and the results are tabulated in Table 3. This study is conducted by discarding the multi-class structure preserving oversampling technique from the proposed SPVL architecture and incorporating other oversampling techniques on the latent space data namely no oversampling, SMOTE, ADASYN, random oversampling and Borderline SMOTE. The average class specific accuracy (ACSA) and geometric mean (GM) for the corresponding datasets are tabulated. These experiments conducted on tabular and image data indicates the effectiveness of the proposed technique on structured and unstructured data.

	MNIST dataset		WAFER dataset	
Method	ACSA(%)	GM(%)	ACSA(%)	GM(%)
VAE+MLP	85.5	84.1	93.9	93.8
VAE+SMOTE+MLP	88.3	87.5	96.1	96.0
VAE+ADASYN+MLP	86.1	85.2	96.1	96.0
VAE+RANDOM+MLP	86.6	85.3	96.6	96.6
VAE+BoS+MLP	86.6	85.5	94.7	94.6
SPVL(ours)	**89.8**	**89.1**	**99.1**	**99.1**

Table 3: Ablation study on MNIST and WAFER dataset

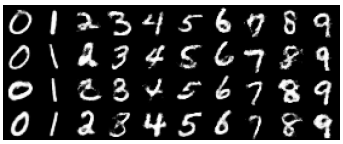
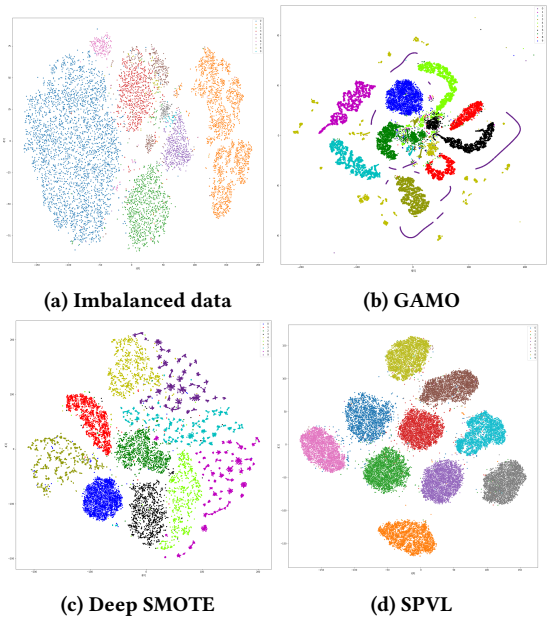


Figure 4: SPVL oversampled latent space generated images

Implementation of the proposed SPVL architecture is done in python programming language. The implementation of the model for wafer dataset is done in google colab notebook with Intel(R) Xeon(R) CPU @ 2.20 GHz with python version 3.7.12, NumPy version 1.19.5, TensorFlow version 2.7.0, Keras version 2.7.0, sklearn version 1.0.2. SPVL model implementation for the remaining five datasets (MNIST, Fashion MNIST, Australian credit, German credit, and Secom) is done in Python Jupyter notebook with AMD Ryzen 5 CPU @ 2.10 GHz with python version 3.7.4, NumPy version 1.19.5, TensorFlow version 2.5.1, Keras version 2.5.0, sklearn version 0.23.2.

4 CONCLUSIONS

In this paper, we propose a novel SPVL architecture, which uses the encoder-decoder framework of VAE with multi-class structure preserving oversampling. The proposed SPVL oversampling approach



(a) Imbalanced data **(b) GAMO**

(c) Deep SMOTE **(d) SPVL**

Figure 3: t-SNE representation of (a) Imbalanced data distribution ($IF = 0.96$), (b)GAMO balanced distribution, (c) Deep SMOTE balanced distribution, (d) SPVL balanced distribution

is simple, powerful and capable of capturing the distribution among classes in a separable manner and can preserve the class boundaries. The performance evaluation of SPVL technique is carried out by training an MLP classifier with the SPVL balanced dataset and testing with unseen dataset. The improved class distribution captured in the latent space has helped to obtain better classification performance in terms of performance metrics, namely, ACSA and GM for six selected benchmark datasets. Generated images exhibit superior visual quality.

5 ACKNOWLEDGMENTS

Indu Solomon and Uttam kumar are grateful to International Institute of Information Technology Bangalore (IIIT Bangalore), India for the infrastructure support and acknowledge Mphasis Cognitive Computing Centre of Excellence for the financial assistance under Grant No. 7111.

Senthilnath Jayavelu and Md M. Ferdaus acknowledge funding from the Accelerated Materials Development for Manufacturing Program at A*STAR via the AME Programmatic Fund by the Agency for Science, Technology and Research under Grant No. A1898b0043.

REFERENCES

[1] Tahira Alam, Chowdhury Farhan Ahmed, Sabit Anwar Zahin, Muhammad Asif Hossain Khan, and Maliha Tashfia Islam. 2018. An effective ensemble method for multi-class classification and regression for imbalanced data. In *Industrial Conference on Data Mining*. Springer, 59–74.

[2] Hong Cao, Xiao-Li Li, Yew-Kwong Woon, and See-Kiong Ng. 2011. SPO: Structure preserving oversampling for imbalanced time series classification. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 1008–1013.

[3] Cristiano L Castro and Antônio P Braga. 2013. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE transactions on neural networks and learning systems* 24, 6 (2013), 888–899.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. 2022. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[6] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[8] Ting Guo, Xingquan Zhu, Yang Wang, and Fang Chen. 2019. Discriminative sample generation for deep imbalanced learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), August 10-16 2019, Macao, China.*

[9] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*. Springer, 878–887.

[10] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328.

[11] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[12] Yann LeCun. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* (1998).

[13] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. 2018. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655* (2018).

[14] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. 2019. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1695–1704.

[15] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano—RUSBoost. 2010. A Hybrid Approach to Alleviating Class Imbalance‖ IEEE Transactions On Systems. *Man, And Cybernetics—Part A: Systems And Humans* 40, 1 (2010).

[16] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. 2010. Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[17] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[18] Jian Yin, Chunjing Gan, Kaiqi Zhao, Xuan Lin, Zhe Quan, and Zhi-Jie Wang. 2020. A novel model for imbalanced data classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6680–6687.