Expert Systems With Applications 175 (2021) 114750

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection

Zhenchuan Li, Mian Huang *, Guanjun Liu, Changjun Jiang

Key Laboratory of the Ministry of Education for Embedded System and Service Computing, Shanghai Electronic Transactions and Information Service Collaborative Innovation Center, Department of Computer Science, Tongji University, Caoan Road 4800, Jiading District, Shanghai 201804, PR China

ARTICLE INFO

Keywords:
Class imbalance
Data overlap
Dynamic weighted entropy
Fraud detection

ABSTRACT

Class imbalance with overlap is a very challenging problem in electronic fraud transaction detection. Fraudsters have racked their brains to make a fraud transaction as similar as a genuine one in order to avoid being found. Therefore, lots of data of fraud transactions overlap with genuine transactions so that it is hard to distinguish them. However, most attention has been focused on class imbalance rather than overlapping issues for machine-learning-based methods of fraud transaction detection. This paper proposes a novel hybrid method to handle the problem of class imbalance with overlap based on a divide-and-conquer idea. Firstly, an anomaly detection model is trained on the minority samples for excluding both a few outliers of minority class and lots of majority samples from the original dataset. Then the remaining samples form an overlapping subset that has a low imbalance ratio and a reduced learning interference from both minority class and majority class than the original dataset. After that, this difficult overlapping subset is dealt with a non-linear classifier in order to distinguish them well. To achieve good properties of the overlapping subset, we propose a novel assessment criterion, Dynamic Weighted Entropy (DWE), to evaluate its quality. It is a specially designed trade-off between the number of excluded outliers of minority class and the ratio of class imbalance of overlapping subset. With the help of DWE, time consumption on searching good hyper-parameters is dramatically declined. Extensive experiments on Kaggle fraud detection dataset and a large real electronic transaction dataset demonstrate that our method significantly outperforms state-of-the-art ones.

1. Introduction

FinTech combining finance and technology has become a popular research area (Gomber, Koch, & Siering, 2017). Artificial Intelligence (AI) innovates new and intelligent FinTech for higher quality service; meanwhile, FinTech provides a broad platform and application scenarios for AI research and innovation. Electronic transaction fraud detection has attracted attention in Fintech. Recognizing fraud transactions is very challenging (Dal Pozzolo, Caelen, Le Borgne, Waterschoot, & Bontempi, 2014; Abdallah, Maarof, & Zainal, 2016). One important reason is the problem of class imbalance, i.e., the ratio of legal and fraud transaction samples is very large so that a machine-learning-based detector is in favor of legal transactions. Especially, the overlap problem between imbalanced classes makes fraud transactions be hardly identified. Class imbalance (Japkowicz & Stephen, 2002) occurs when there is a large difference in sample size between different classes of data, which is obvious in electronic transaction records that the number of fraud transactions is much less than that of normal legitimate transactions. Overlap (Denil, 2010) usually refers to the problem that samples of different classes occur in the same data space region which increases the difficulty of learning a classifier that can distinguish samples of different classes in the overlapping region. Since fraudsters leave no stone unturned to imitate the transaction behaviors of real cardholders to make the fraud detection system inefficient, fraud transactions and legitimate transactions will be intertwined in some data space region and cause the overlap problem. Some studies (Prati, Batista, & Monard, 2004; Stefanowski, 2013) have already noticed the overlap problem when dealing with imbalanced datasets. Das, Datta, and Chaudhuri (2018) detailedly analyzed the correlation of three principal properties (i.e. Imbalance Radio, overlap between classes, and size of dataset) of class imbalance problem which together increase the difficulty of recognizing different samples from different classes. They found that even if a dataset has a high Imbalance Radio, classifiers can achieve good performance without the appearance of data overlap. But as the

* Corresponding author.
E-mail addresses: chuanzhen_li@126.com (Z. Li), net-cn@163.com (M. Huang), liuguanjun@tongji.edu.cn (G. Liu), cjjiang@tongji.edu.cn (C. Jiang).

https://doi.org/10.1016/j.eswa.2021.114750
Received 25 July 2020; Received in revised form 4 January 2021; Accepted 16 February 2021
Available online 25 February 2021

degree of overlap of the data set increases, the performance of classifiers deteriorates, especially with the Imbalance Radio increasing simultaneously. Therefore, overlap problem has to be taken into consideration when building classifiers with imbalanced data set.

Although the class imbalance problem has been studied for decades, the most efforts are made to deal with the problem of how to balance the sample numbers of different classes. However, the inherent characteristics of the imbalanced datasets such as overlap do not get enough attention. There are a few studies considering class imbalance and overlap simultaneously. These methods can be summarized as a uniform framework, as shown in Fig. 1. Firstly, an original data set is divided into two parts, overlapping subset and non-overlapping subset, mostly based on the neighborhood based method (such as *k*-Nearest Neighbor (*k*NN) and its variations). Then an under-sampling method (e.g. Tomek links Kubat et al., 1997) is applied to remove some samples of majority class in the overlapping subset to make the decision boundary clearer and biased towards minority samples. Finally, the processed samples are used for the learning of a classifier to achieve a good performance of detecting minority samples.

However, when this framework is applied for detecting fraud transactions in real environment, it has some drawbacks. First, for finding the overlapping samples in a dataset, the neighborhood based methods need to calculate the distance of every pair of samples. But a real application has hundreds of millions of transactions with lots of features every day, so that these methods have to face the curse of dimensionality and have high computational complexity. Second, the overlapping subset are handled by the under-sampling method and some majority samples will be removed for the bias of achieving better accuracy of minority samples classification. However, for electronic transactions data set, fraud transactions (minority class) are deeply overlapped with legitimate transactions (majority class), the under-sampling method has to remove plenty of legitimate transactions to obtain a good decision boundary. Obviously, this leads to the loss of the important information of those removed samples and causes the increase of mis-recognition of legitimate transactions. Finally, the performance of neighborhood-based methods and undersampling methods in this framework can only be evaluated until the finish of the whole model learning. And any change of the hyper-parameters of neighborhood-based methods and under-sampling methods requires a new training of detection model and a re-verification of the performance of the new model. So the hyper-parameter setting of this framework is also a time- and resource-consuming process.

To handle these drawbacks, this paper proposes a hybrid method with a Dynamic Weighted Entropy measurement for improving the efficiency of the whole model learning, and the divide-and-conquer idea is used. Our work is summarized as follows:

- We propose a Divide-and-Conquer Strategy based hybrid framework to address the class imbalance problem with overlap. This method includes two steps:
  - *Divide* step: An unsupervised anomaly detection model (such as isolation Forest (iForest) (Liu, Ting, & Zhou, 2008), One-Class SVM (OCSVM) (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001) and Auto-Encoder (AE) Zhou & Paffenroth, 2017) is to learn the principle profile of the minority samples (fraud transactions). With the help of this profile, almost all of the minority samples and some majority samples (legitimate transactions) heavily overlapped with these minority ones are put together and form the overlapping subset, and the rest majority samples and the rest minority samples (anomalies of fraud transactions) form the non-overlapping subset. This is a novel idea of using anomaly detection models to detect anomalies based on the profile of minority class, which is different from the conventional idea of anomaly detection models using a majority sample profile and regarding anomalies as minority samples.
  - *Conquer* step: The two different subsets are handled independently in this step. Since only few anomalies of minority samples are divided into the non-overlapping subset, they can be classified as the majority ones for simplicity. For the overlapping subset, a powerful supervised classifier (such as Random Forest (RF) (Breiman, 2001) and Artificial Neural Networks (ANN) (Basheer & Hajmeer, 2000)) is adopted to learn from this highly overlapped subset which however has a much lower Imbalance Radio than the original dataset, and thus it is easier for the classifier to focus on how to distinguish those heavily overlapped samples without the interference from the easily recognized samples.
- For improving the efficiency of the whole model learning, we propose a novel assessment measure, Dynamic Weighted Entropy (DWE), to evaluate the quality of the overlapping and non-overlapping subsets obtained in the *Divide* step by the anomaly detection model with some hyper-parameters. DWE considers both the number of excluded anomalies of minority samples and the Imbalance Radio of the overlapping subset, which should be carefully traded off for good overall performance of our proposed method. With the help of DWE, it is more efficient to select the optimal hyper-parameters of the anomaly detection model. The parameter selecting of our model can be regarded as the decision making under uncertain environment, which is comprehensively studied in Rodger (2019) from the perspective of methodology for reducing strategic decision-making entropy. Coincidentally, the basic idea of our DWE has something in common with it, readers can refer it for more inspirations.
- We conduct abundant experiments on the public fraud detection dataset from Kaggle [1] and our real electronic transaction dataset from a financial company in China. The positive correlation between DWE and the overall performance of our proposed method is verified. And the results of contrast experiments demonstrate that our method outperforms the state-of-the-art ones for the problem of class imbalance with overlap.

This paper is organized as follows: Section 1 summarizes the weaknesses of the methods for handling class imbalance problem with overlap and introduces our hybrid method. Section 2 reviews the research about overlap, class imbalance problems, and the methods for handling the class imbalance problem with overlap. In Section 3, the proposed hybrid framework with DWE is introduced in detail. Experiment setting and result analysis are illustrated in Section 4. Finally, the conclusion is provided in Section 5.

2. Related works

2.1. Relationship of overlap and class imbalance

When it comes to class imbalance, most of the attention is focused on the huge difference in the number of samples in different classes. As shown in Fig. 2 (a), the dataset has a high Imbalance Radio value with no overlap of different classes, it may not cause a problem for some classifiers with the loss function based on the max-magin criterion. But if the overlap appears with the class imbalance, as shown in Fig. 2 (b), it is much more difficult to deal with. Most of the studies of handling class imbalance problem pay attention on how to balance the sample size of different classes for improving the performance of classifiers. However, the performance of these methods have no obvious improvement to the problem of class imbalance with overlap if they just focus on the Imbalance Radio value but do not deeply analysis the characteristic of the dataset.

Fortunately, some studies has already taken notice of the weakness of these methods. Prati et al. (2004) is one of the earliest studies on this

[1] https://www.kaggle.com/mlg-ulb/creditcardfraud

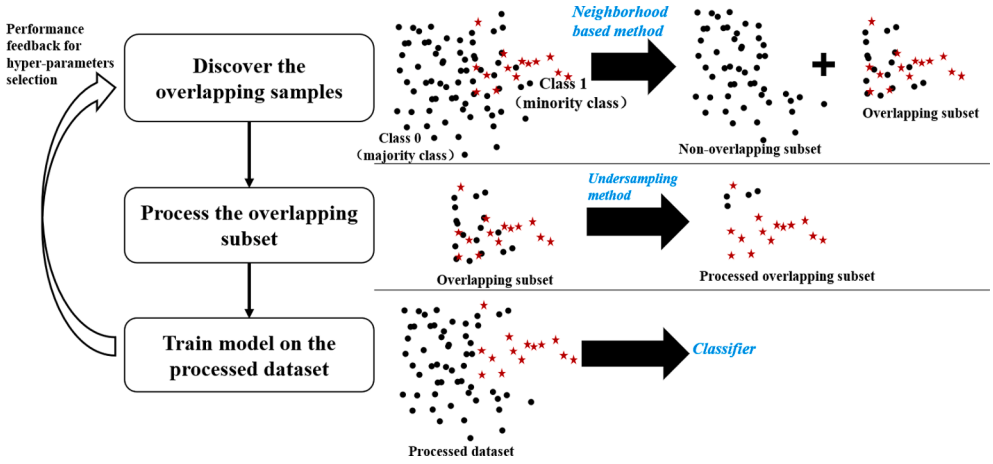

Fig. 1. The uniform framework for handling class imbalance and overlap problem.

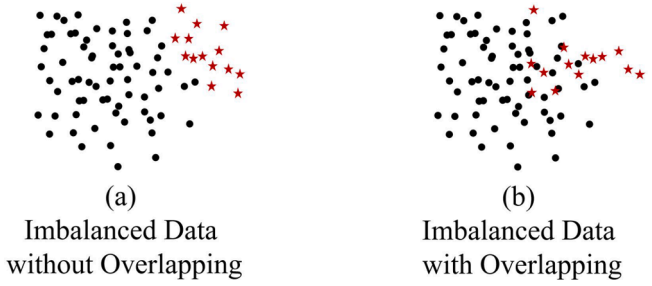

Fig. 2. The Datasets with Different Overlapping Degree but the same Imbalance Radio.

problem. It finds the strong correlation between the overlap degree of different classes based on a systematic experiments with a set of artificial datasets. García, Mollineda, Sánchez, Alejo, and Sotoca (2007) and García, Sánchez, and Mollineda (2007) conduct many experiments to analyze the behavior of some widely-used classifiers on some artificial imbalanced datasets with overlap. They demonstrate that the class imbalance may not pose a problem by itself and the problem of class imbalance with overlap should be treated carefully because the characteristic of local subsets may different from the whole dataset. They also analyze the impacts of overall Imbalance Radio, local Imbalance Radio and the size of overlapping region on the performance of KNN model (García, Mollineda, & Sánchez, 2008). Denil and Trappenberg (2010) conducts some experiments to verify the interdependent effects of overlap and class imbalance. Besides, it finds that more training data may lead to the performance degradation of the SVM model when overlap is present with class imbalance. Stefanowski (2013) presents a discussion about the main properties (i.e. small disjuncts, overlapping between classes and rare examples) of dataset that lead to the difficulty of classifier learning combined with the class imbalance. Mercier et al. (2018) proposes a new overlapping measure (degOver) that takes the Imbalance Radio and the structure of dataset into consideration, and demonstrates more powerful classifiers (e.g. MLP and CART) can achieve better performance on complex dataset than some simple classifiers (e.g. KNN and linear SVM). Das et al. (2018) present a comprehensive study on data irregularities and their interrelations including class imbalance, small disjuncts, class skew and missing features. They also summarize the most recent and notable methods for handling data irregularities. Finally, some notable future research directions are presented to inspire better model design. Some reviews (Fernández, García, & Herrera, 2011; López, Fernández, Moreno-Torres, & Herrera, 2012; Ali, Shamsuddin, & Ralescu, 2015; Branco, Torgo, & Ribeiro, 2016; Haixiang et al., 2017) point out that the overlap of classes should be payed more attention to.

2.2. Handling class imbalance with overlap

Apart from the analysis of the interrelationship between overlap and class imbalance, some methods are also proposed to deal with the difficult problem of class imbalance with overlap. Most of those methods can be summarized as the process shown in Fig. 1. The main differences of those methods are the applied models for discovering and dealing with the overlapping subset.

Some researches employ the neighborhood-based method for preprocessing the original dataset with class overlap and imbalance problems. Alejo, Sotoca, García, and Valdovinos (2011) and Alejo, Valdovinos, García, and Pacheco-Sanchez (2013) present a method combining the data editing techniques and a balanced MSE loss function for MLP. The data editing method based on the variants of KNN is applied to delete the overlapping samples of the majority classes for producing smooth decision boundaries. The MLP based on the balanced MSE loss function is trained on the edited dataset. Vuttipittayamongkol and Elyan (2020) present four *k*NN-based methods to identify and remove majority samples from the overlapping region. They also try to make the minority samples as visible as possible and the excessive elimination of majority samples as few as possible. Fernandes and de Carvalho (2019) discusses the drawback of removing majority samples in the overlapping region and proposes an evolutionary ensemble-based method with ensembles of classifiers to handle this problem. However, the basic classifier is still trained on the preprocessed dataset in which the majority samples are eliminated in the overlapping region. And the neighborhood-based minimum spanning tree is applied to discover the overlapping region. Although these neighborhood-based methods can discover the overlapping subset directly and accurately, their expensive calculation cost make them hardly applicable when the actual data is high-dimensional and has a huge amount of samples.

There are also a few methods that applies a clustering based approaches for preprocessing the imbalanced dataset with class overlap. Some researches (Das, Krishnan, & Cook, 2014; Bunkhumpornpat & Sinapiromsaran, 2017) present a clustering-based undersampling (ClusBUS) technique to handle the problem of class imbalance with overlap. At the same time, the density-based clustering model, DBSCAN, is applied to discover and remove majority samples from the overlapping region. Rubbo and Silva (2018) focuses on improving the performance of KNN model on the datasets with class imbalance and overlap problem, and the self-organizing maps (SOM) model is applied to cluster samples according to their similarity for discovering the overlapping samples and eliminating them selectively. Vuttipittayamongkol, Elyan, Petrovski, and Jayne (2018) proposes an Overlap-Based

Undersampling (OBU) method in which the soft clustering algorithm (e.g. Fuzzy c-means) is applied to find the overlapping region and remove those majority samples in the entire overlapping region. However, the excessive eliminations of samples possibly lead to the serious loss of information.

Some hybrid methods are proposed too. Vorraboot, Rasmequan, Chinnasarn, and Lursinsap (2015) present a hybrid method that divides an original dataset into non-overlapping subset, borderline dataset and overlapping subset, and apply different responsive classification algorithms for these subsets. Lee and Kim (2018) presents an overlap-sensitive margin (OSM) based method for handling this problem. A modified fuzzy support vector machine is applied for separating an original dataset into soft- and hard-overlap subsets. Then the two obtained subsets are classified using the support vector machine and 1-nearest neighbor algorithm, respectively.

These methods use an undersampling method on the overlapping subset for achieving a clear boundary biasing to the minority class, but it is hard to make sure which majority sample should be eliminated so that they lose information uncontrollably.

3. The proposed method

3.1. The proposed hybrid method

To overcome the problem of class imbalance with overlap in fraud detection, we propose a novel hybrid method. As shown in Fig. 3, inspired by the idea of Divide-and-Conquer, our hybrid framework consists of two steps, *Divide* step and *Conquer* step.

Divide step: The original dataset is divided into overlapping subset and non-overlapping subset in this step. An unsupervised anomaly detection model is applied to construct the principle profile of the minority samples without the participation of majority samples. Almost all of the minority samples and some majority samples deeply overlapped with these minority ones conform to the learned profile, so they are put into the same subset as the overlapping subset. The rest majority samples and few minority samples do not conform to the learned profile, so they are regarded as the "anomalies" and are put into another subset as the non-overlapping subset. The neighborhood-based model in many existing methods needs to calculate the distance between each pair of samples in the minority and majority classes in order to obtain the overlapping subset; but our anomaly detection model in our framework is different from it: our model is trained on the minority samples with a much smaller sample size than the whole dataset. So, our framework is much more efficient when obtaining the overlapping subset.

There already have many anomaly detection models, such as OCSVM (Schölkopf et al., 2001), iForest (Liu et al., 2008) and deep auto-encoder (Zhou & Paffenroth, 2017). Each of them has its own specialty and can construct different profiles of the minority samples based on different hyper-parameter settings. The hyper-parameters should be set carefully for obtaining high-quality subsets, which is very important to the model learning in the next step.

Conquer step: In this step, different subsets are treated based on their different characteristics. For the non-overlapping subset, there are only few anomalies of the minority samples and most of them are the majority samples. Hence, all samples in non-overlapping subsets are regarded as the minority ones for simplicity in this paper. For the overlapping subset, because the samples in it are overlapped heavily, it is hard to distinguish them accurately. So a powerful supervised classifier is applied for handling this difficult problem. Fortunately, although some majority samples are included in the overlapping subset, the Imbalance Radio of this subset is lower than that of the original dataset, because a large number of majority samples are put into the non-overlapping subset, and the anomalies of minority samples are also excluded. So, the interference from both minority class and majority class for the classifier learning is reduced. This is very beneficial for the classifier to focus on learning how to distinguish these different classes of samples and achieve optimal performance. Additionally, the classifier can be selected according to the difficulty of the overlapping subset.

3.2. The Proposed DWE

There are two key factors that influence the overall performance of our method. First, the minority samples in the non-overlapping subset are viewed as "anomalies" and thus are treated as the majority samples, which will directly influence the overall recall value of minority class if they are actually not anomalies. Therefore, our method should reduce such miscalculation. The other one is about the size of majority samples put into the overlapping subset. Too many majority samples will lead to a high Imbalance Radio of the overlapping subset, while a high Imbalance Radio can increase the difficulty of learning a classifier and decrease the final performance. These two factors are sensitive to the setting of hyper-parameters of the anomaly detection model, because the numbers of the anomalies are directly controlled by the learned boundary of the anomaly detection model. It is significant and necessary to give an effective indicator to measure the quality of the overlapping subset and guide the hyper-parameters selection for achieve better

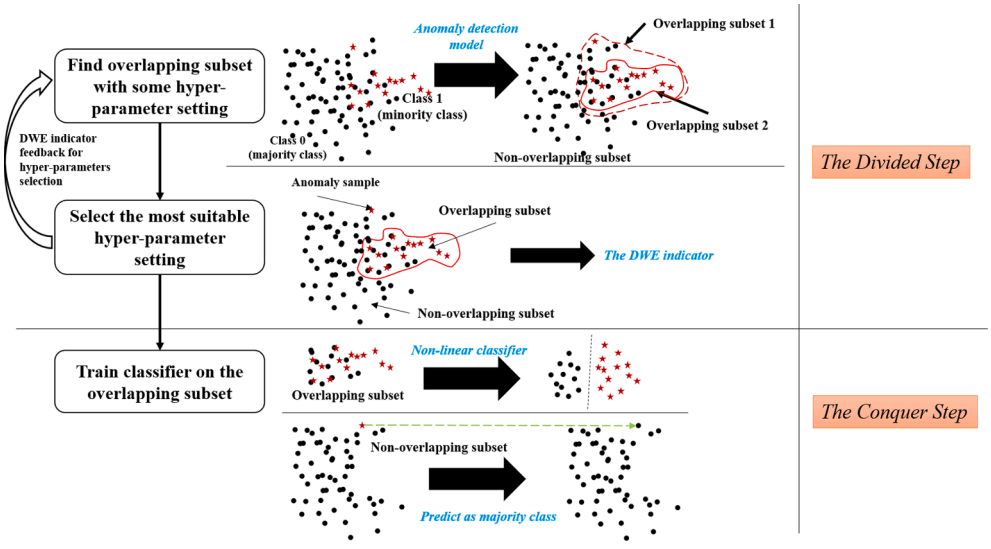

Fig. 3. The proposed hybrid framework.

overall performance.

We propose an indicator for evaluating the obtained overlapping subset through the anomaly detection model, named Dynamic Weighted Entropy (DWE), that considers key factors mentioned above. Our DWE is designed as:

$$G_{DWE}(\theta) = W_{snr} \cdot H \tag{1}$$

where θ denotes the hyper-parameters of the anomaly detection model, H is the information entropy of the overlapping subset, and W_{snr} is the dynamic weight of H based on the SNR (Signal-to-Noise Ratio Johnson, 2006) of minority class. The H and W_{snr} can be formalized as follows:

$$H = -\sum_{i=0}^{k}\left(p_i \lg\left(p_i\right)\right) \tag{2}$$

$$W_{snr} = \lg\frac{n_{all} - n_{outliers}}{n_{outliers}} = \lg\left(\frac{n_{all}}{n_{outliers}} - 1\right) \tag{3}$$

where $i \in \{0, 1, .., k\}$ and k represents different classes in an overlapping subset, p_i is the probability of any sample belonging to class i, n_{all} is the number of all minority samples and $n_{outliers}$ presents the number of outliers of minority class.

Algorithm 1: Our Hybrid Algorithm for Fraud Detection

Input: Training dataset with $X = \{x_1, x_2, \cdots, x_m\}$ includes n_{mi} minority samples X_{mi} and n_{maj} majority samples X_{maj} where $n_{mi} + n_{maj} = m$.
Parameter:$\theta_{ADM} = \{\theta^1_{ADM}, \theta^2_{ADM}, \cdots, \theta^r_{ADM}\}$: The hyper-parameter set of the anomaly detection model (ADM) for the best hyper-parameter selection.
θ_{clf}: The hyper-parameters of the non-liner classifier.
Output: The learned anomaly detection model M_{ADM} with the best hyper-parameter $\widetilde{\theta}_{ADM}$, and the trained non-linear classifier M_{clf}.
1.Let $i = 0, V_{GWE} = 0$ and $\widetilde{\theta}_{ADM} = \theta^1_{ADM}$;
Divide step:
2.**While** $i < r$ **do**
3. train anomaly detection model with θ^i_{ADM} on X_{mi};
4. divide X into overlapping subset and non-overlapping subset;
5. calculate H by Formula (2);
6. calculate W_{snr} by Formula (3);
7. $G_{DWE} \leftarrow W_{snr} \cdot H$;
8. **if** $V_{GWE} < G_{DWE}$ **then**
9. $V_{GWE} \leftarrow G_{DWE}$;
10. $\widetilde{\theta}_{ADM} \leftarrow \theta^i_{ADM}$;
11. **end if**
12. **end while**
13. obtain M_{ADM} with $\widetilde{\theta}_{ADM}$;
14. obtain overlapping subset and non-overlapping subset with M_{ADM};
Conquer step:
15. train a anomaly detection model on X_{mi} and obtain M_{ADM} with $\widetilde{\theta}_{ADM}$;
16. train a non-liner classifier on overlapping subset and obtain M_{clf} with θ_{clf};
17. **return** M_{ADM}, M_{clf}

The information entropy H can be used to measure the average amount of information of a dataset (Lee & Xiang, 2000). We apply it to measure the average information of overlapping subset. If there are too many majority samples divided into the overlapping region, the $p_{majority}$ will close to 1 and the $p_{minority}$ will close to 0 such that the value of H will close to the minimum value 0. Once the number of majority samples is close to that of minority samples, H will be close to the maximum value 1. Meanwhile, the bigger H, the more balanced overlapping subset, which is benefit to the classifier learning in the *Conquer* step.

However, although H can merely measure the property of overlapping subset, it has no ability of evaluating the loss of minority samples. This paper applies the dynamic SNR of the minority class, W_{snr}, to measure the impact of the decision boundary on the minority samples. W_{snr} is the ratio of signal level to the noise (represents the outlier of minority class) level, which is dynamically changed on different settings of hyper-parameters of the anomaly detection model. The more minority samples are regarded as outliers, the lower the W_{snr} is. But, a lower W_{snr} means that more minority samples are regarded as the outliers and thus put into the non-overlapping subset, such that these outliers will be classified as the majority ones. This will cause a lower recall and is not our expectation. It is expected that only few minority samples regarded as outliers with limited information loss and a small part of majority samples form the overlapping subset in order to maintain a low Imbalance Radio. Hence, we apply W_{snr} as the dynamic weight of H to restrict them for achieving the maximum value which needs that W_{snr} and H are as big as possible.

The pseudo code of our hybrid method is presented in Algorithm 1.

4. Experiments

4.1. Datasets

We conduct extensive experiments on real-world datasets:

4.1.1. Kaggle dataset

The Credit Card Fraud Detection dataset from Kaggle consists of anonymized credit card transactions labeled as fraudulent or genuine. These transactions are generated in two days in September 2013 by European cardholders. This dataset has only 492 fraud transactions out of all 284807 transactions causing a very high Imbalance Radio value 577.8. These transactions has 30 features (i.e. Time, V1, V2, …, V28 and Amount). Besides the Time and Amount, all other features are transformed into numerical values by a PCA method for the privacy protection. More details about these datasets are shown in Table 1.

4.1.2. Private dataset

Our private dataset is from a financial company in China. It contains up to 3.5 million transactions from April to June in 2017 labeled by professional investigators of the financial company. Table 2 shows the details of the transactions in each month, and the class imbalance problem is also very serious. The features of this dataset include transaction time, transaction amount, transaction type, merchant type, currency type, card type and so on. To make full use of this dataset, we divide it into 9 groups in chronological order and each group consists of transactions from 10 consecutive days. Then, every 4 consecutive groups form a new dataset for model training (with the first 3 groups) and testing (with the last group). Thus the original dataset is divided into 6 private datasets. Table 1 shows more details about these datasets.

4.2. Experimental methods

Baseline Models For dataset from Kaggle and our private datasets, the Random Forest (RF) model is adopted as the baseline model because it is widely used in many literatures. The random undersampling method is applied for balance the dataset before model training. The hyper-parameter settings of RF models are shown in Table 4.

The Proposed Method The proposed hybrid method in this paper is a

Table 1
The Information of Datasets.

Datasets	Samples	Features	Fraud/ Legitimate	Imbalance Radio	Overlap Degree
Kaggle	284807	30	492/284315	577.8	0.391
Private1	1625103	43	23027/ 1602076	69.6	0.295
Private2	1687306	43	33305/ 1654001	49.6	0.289
Private3	1669601	43	34098/ 1635503	47.9	0.281
Private4	1555098	43	29955/ 1525143	50.9	0.273
Private5	1500516	43	30607/ 1469909	48.0	0.284
Private6	1455989	43	29106/ 1426883	49.0	0.266

Table 2
The Private Dataset.

Date	Samples	Features	Fraud Rate
2017–04	1,243,035	43	1.07%
2017–05	1,216,299	43	2.22%
2017–06	1,042,714	43	2.39%

universal framework. The anomaly detection model in the *Divide* step and the non-linear classifier in the *Conquer* step can be flexibly changed according to the difficulty of dataset. For a comprehensive verification of the proposed hybrid method, some commonly used anomaly detection models (i.e. OCSVM, iForest, AE) are applied to divided the original dataset into overlapping subset and non-overlapping subset in the *Divide* step. And the RF and ANN are applied as the non-linear classifier in the *Conquer* step trained on the overlapping subset.

4.2.1. Compared methods

We compare our hybrid method with some of commonly used methods for class imbalance problem and with state-of-the-art methods for handling class imbalance problem with overlap.

1. Tomek links (Kubat et al., 1997): It is a popular baseline under-sampling method for imbalanced learning, which removes the noisy and borderline examples from the majority class. RF is also as its classifier.
2. SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002): It is a famous over-sampling method for imbalanced dataset. It generates new minority samples by linear interpolation between adjacent points to balance different classes. The RF model is used as the classifier on the balanced dataset.
3. OC-SVM (Schölkopf et al., 2001): It is a single class learning method that trained on the minority samples only without considering the majority samples. It is particularly suitable for severe imbalance problem.
4. OSM (Lee & Kim, 2018): It is an overlap-sensitive margin (OSM) classifier based on a modified fuzzy support vector machine and *k*NN algorithm to address imbalanced and overlapping data sets.
5. NB-Tomek (Vuttipittayamongkol & Elyan, 2020): This method is proposed to eliminate majority class instances from the overlapping region and minimize the information loss of excessive data elimination.

Evaluation Metrics The conventional confusion matrix of binary classification is shown in Table 3. For evaluating the performance of classifiers, Accuracy, Precision and Recall are the mostly used basic measures. However, they have some weakness of biasing to the majority class when the classifier learning on an imbalanced dataset. F_1 score is the weighted harmonic mean of precision and recall that is a comprehensive and balancing measure. Area under precision-recall curve (AUC_PR) (Saito & Rehmsmeier, 2015) is another metric very suitable for measuring classifier's performance on imbalanced datasets because of its susceptibility of classifiers to imbalanced data sets.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Table 3
The Confusion Matrix.

	True Fraud	True Legitimate
Predicted Fraud	TP	FP
Predicted Legitimate	FN	TN

Table 4
Key hyper-parameters of models for different datasets.

ModelsDatasets	Kaggle	Private
RF	$tree\_number = 50$, $max\_deep = 5$	$tree\_number = 100$, $max\_deep = 7$
AE	It has 4 fully connected layers: encoder-1:16 + elu encoder-2:8 + tanh decoder-1:8 + elu decoder-2:16 + tanh	It has 6 fully connected layers: encoder-1:32 + elu encoder-2:16 + tanh encoder-3:8 + elu decoder-1:8 + elu decoder-2:16 + tanh decoder-3:32 + elu
ANN	It has 3 fully connected layers: FC-1:32 + elu FC-2:64 + elu FC-3:16 + softmax	It has 5 fully connected layers: FC-1:64 + elu FC-2:128 + elu FC-3:128 + elu FC-4:64 + elu FC-5:32 + softmax

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{7}$$

4.3. Experiment environment and parameter settings

Hardware/Software Setup The machine used for all experiments has an Intel Xeon Phi 7290 CPU with 72 cores at 1.5 GHz and 125 GB RAM on Ubuntu 18.04. The developing environment is based on the Python-3.6.8, Tensorflow-1.4.0, Scikit-learn-0.24 and Imbalanced-learn-0.7 (Lemaítre, Nogueira, & Aridas, 2017). The commonly used models (i.e. RF, iForest, CO-SVM, *k*NN) are implemented through Scikit-learn, and the Imbalanced-learn is adopted to implement Tomek Links, SMOTE and random undersampling. The neural network models (AE and ANN) are implemented based on Tensorflow. Other models (OSM and NB-Tomek) are implemented with python code according to their corresponding articles.

Parmeter Settings We provide the hyper-parameter settings of the used networks (AE and ANN) in our proposed method and the networks are differnet for the Kaggle and the private datasets. Their architectures are shown in Table 4: These neural networks are all trained based on the Adam optimizer (Kingma & Ba, 2014) with suggested parameters. And the parameter settings for model training are as follows:

4.3.1. For Kaggle dataset

The batch size of 64 was used and these models are trained for 50 epochs. The learning rate schedule of AE model was initialized at $\lambda = 1e-6$ with a step size of 10 epochs and $\gamma = 0.1$. And the λ is set $1e-5$, the γ is set 0.1 for the ANN model.

4.3.2. For Private datasets

The batch size of 128 was used and these models are trained for 100 epochs. The learning rate schedule of AE model was initialized at $\lambda = 1e-5$ with a step size of 10 epochs and $\gamma = 0.1$. And the λ is set $1e-4$, the γ is set 0.1 for the ANN model.

For the compared methods, we use their suggested hyper-parameter settings according to their corresponding studies. The grid search method is applied to search the best hyper-parameters of the models (OCSVM and iForest) used in the *Conquer* step of our proposed method.

4.4. Experimental results and analysis

Effectiveness of DWE We first conduct an experiment to verify the effectiveness of the proposed indicator, DWE. Since DWE is just related to the properties of the overlapping and non-overlapping subsets, the OCSVM is applied as the anomaly detection model in the *Divide* step of our proposed method to divide the original dataset with different hyper-parameter settings. RF is applied as the non-linear classifier in the

Conquer step trained for distinguishing samples from different classes. Finally, the F_1 value of the whole model of our hybrid method is obtained so that the relationship between the F_1 value and the DWE value can be analyzed. As shown in Fig. 4, it is obvious that the F_1 value is proportional to DWE, and almost all the highest F_1 values are achieved with the biggest DWE. It can be concluded that the proposed DWE has the ability to measure the quality of the obtained overlapping and non-overlapping subsets. Note that the bigger the DWE value, the better the performance of the proposed hybrid method.

Effectiveness of Our Hybrid Method This experiment is designed to verify the effectiveness of the proposed hybrid method compared with some state-of-the-art methods introduced above. The proposed hybrid method is implemented with some specific anomaly detection models (i.e. OCSVM, iForest and AE) combining with different non-linear classifiers (i.e. RF and ANN). The hyper-parameters of these specific anomaly detection models are selected with the help of our proposed DWE indicator. Tables 5 and 6 show the experimental results. First, almost all of the models based on the proposed hybrid method achieve better results than those state-of-the-art methods. Then, it is obvious that the implemented hybrid methods with the ANN model as the non-linear classifier can achieve better performance than that with the RF model because of the more powerful feature learning ability of the ANN model. Compared with different hybrid methods with the same non-linear classifier, it can be seen that our hybrids method in which AE is as the anomaly detection model can achieve the best performance on all datasets, because of its strong ability on learning the distribution of dataset. Our hybrid methods with RF or iForest have almost the same performance on all experiment datasets. It can be concluded that our proposed hybrid method has the superiority than the state-of-the-art methods for handling the class imbalance problem with overlap. And the more powerful anomaly detection model and the non-classifier, the better overall performance can be achieved.

Time Consumption Analysis The time consumption of every model in the experiments is recorded. Fig. 5 presents the average time consumption of each model in all the experiments. All of the neighborhood-based method (i.e. Tomek Links and NB-Tomek) have the longest time consumption because their trainings need all samples of a dataset. When the OCSVM or SVM models are trained with all samples of every dataset, their time consumption is also very large. However, all of our hybrid models maintain relative low time consumptions, because they are

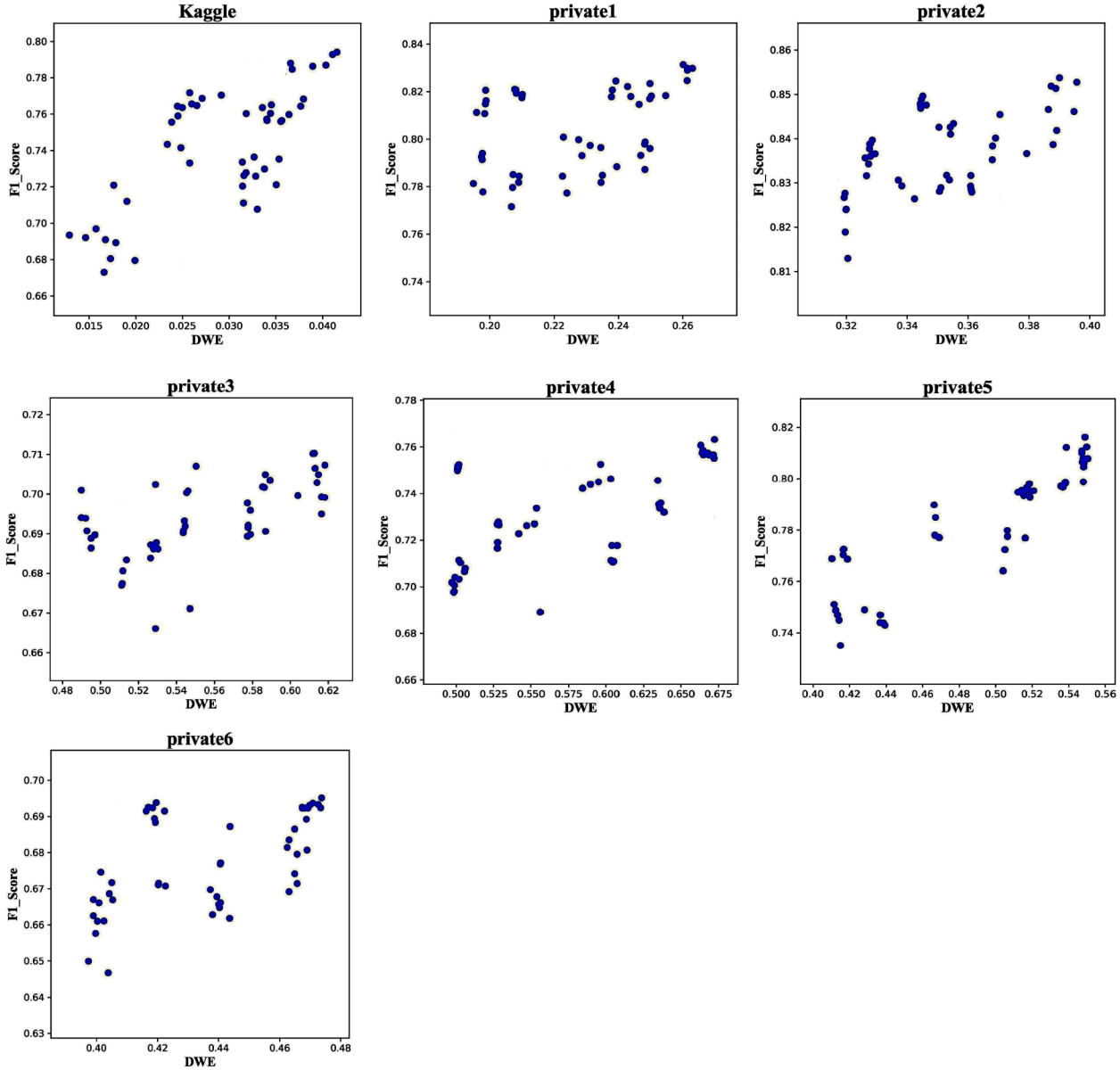

Fig. 4. The F_1 values of our hybrid method with different DWE values.

Table 5
Comparison of different methods based on F_1.

ModelsDatasets	Kaggle	Private1	Private2	Private3	Private4	Private5	Private6
RF	0.67	0.68	0.69	0.64	0.60	0.68	0.69
Tomek Links	0.60	0.72	0.71	0.67	0.63	0.67	0.70
SMOTE	0.64	0.72	0.74	0.66	0.62	0.69	0.68
OC-SVM	0.67	0.66	0.68	0.57	0.59	0.68	0.66
OSM	0.69	0.73	0.74	0.68	0.65	0.71	0.71
NB-Tomek	0.66	0.72	0.76	0.69	0.62	0.69	0.70
Hybrid(OCSVM + RF)	0.70	0.75	0.76	0.71	0.67	0.71	0.72
Hybrid(iForest + RF)	0.71	0.75	0.75	0.72	0.66	0.73	0.72
Hybrid(AE + RF)	0.72	0.76	0.76	0.71	0.69	0.75	0.74
Hybrid(OCSVM + ANN)	0.71	0.76	0.77	0.73	0.68	0.73	0.74
Hybrid(iForest + ANN)	0.71	0.75	0.76	0.73	0.67	0.74	0.73
Hybrid(AE + ANN)	**0.73**	**0.78**	**0.79**	**0.75**	**0.70**	**0.76**	**0.77**

Table 6
Comparison of different methods based on *AUC_PR*.

ModelsDatasets	Kaggle	Private1	Private2	Private3	Private4	Private5	Private6
RF	0.57	0.63	0.70	0.59	0.60	0.63	0.63
Tomek Links	0.50	0.71	0.68	0.61	0.63	0.64	0.67
SMOTE	0.54	0.68	0.71	0.63	0.62	0.62	0.64
OC-SVM	0.57	0.61	0.66	0.59	0.59	0.63	0.62
OSM	0.59	0.69	0.69	0.66	0.65	0.68	0.66
NB-Tomek	0.56	0.68	0.73	0.62	0.62	0.67	0.69
Hybrid(OCSVM + RF)	0.60	0.70	0.74	0.69	0.67	0.70	0.71
Hybrid(iForest + RF)	0.61	0.72	0.72	0.68	0.66	0.69	0.72
Hybrid(AE + RF)	0.62	0.71	0.75	0.72	0.69	0.72	0.72
Hybrid(OCSVM + ANN)	0.61	0.73	0.74	0.71	0.68	0.71	0.73
Hybrid(iForest + ANN)	0.61	0.73	0.73	0.69	0.67	0.72	0.73
Hybrid(AE + ANN)	**0.63**	**0.75**	**0.77**	**0.73**	**0.70**	**0.73**	**0.75**

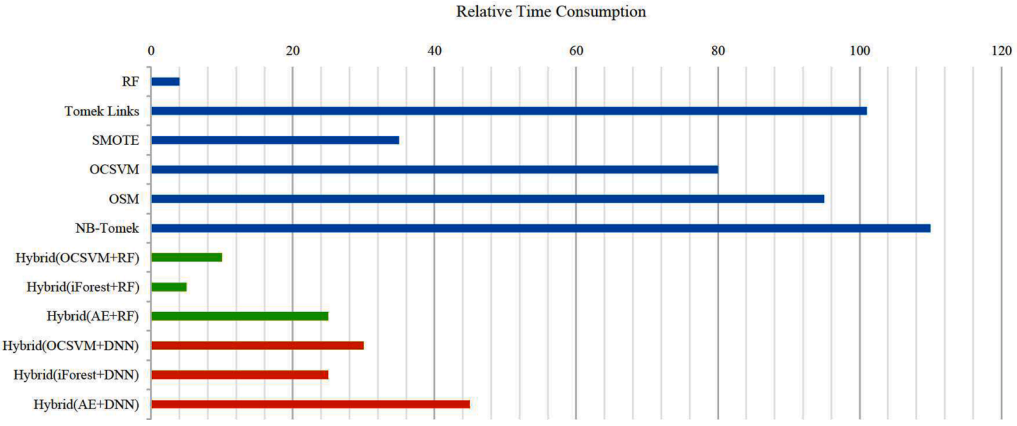

Fig. 5. The time consumption of different models.

trained on the minority samples with smaller sample size instead of a whole dataset.

Fig. 6 shows the time consumption on training the anomaly detection model and the non-linear classifier of our hybrid method. With the guidance of our DWE, the hyper-parameters of anomaly detection model can be easily selected and the huge time consumption of training the nonlinear classifiers repeatedly can be avoided. Especially when the time consumption of the non-linear classifier is larger than that of the anomaly detection model, at least half of the training time can be saved with the help of the proposed DWE indicator.

5. Conclusion

The existing methods for the class imbalance problem with overlap do not take the impact of huge sample size of a dataset in fraud detection into consideration, which results in high time consumption and low efficiency. Besides, the undersampling techniques applied in the existing methods may lead to the information losses and the degradation of overall performance. Hence, this paper proposes a novel hybrid method with the Dynamic Weighted Entropy (DWE) for handling this problem in fraud detection that focuses on improving the model efficiency, avoiding information losses and saving time. And the effectiveness of the method is verified by abundant experiments compared with some SOTA methods.

In this research, we designed a new measurement, i.e. DWE, for assisting the hyper-parameters selection of our model. The parameter selection is actually a decision-making that will influence the performance of the model. From the perspective of decision-making, our

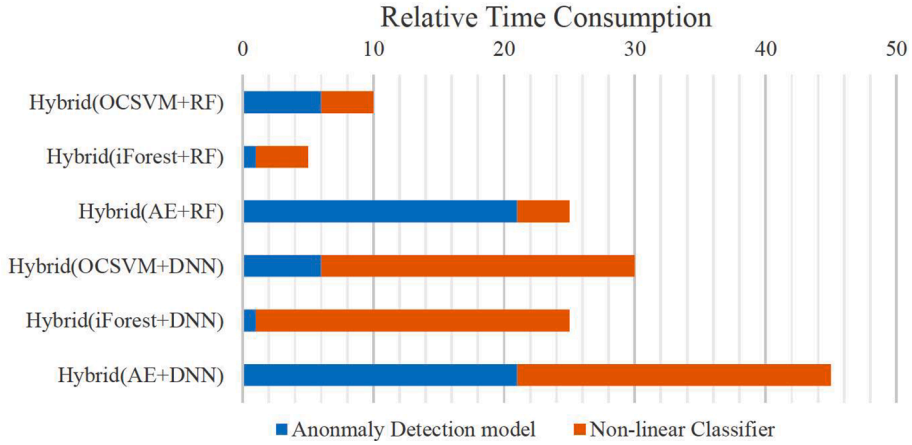

Fig. 6. The time consumption of anomaly detection and nonlinear classifier.

method minimizes the misclassification of minority samples and maximizes the balance of sample numbers of majority and minority in the overlapping subset, which has something in common with the idea of the research (Rodger, 2019) for Strategic Decision-Making. This inspires us for the future studies that we plan to make a comprehensively study about the methodology of decision meaking and introduce it to improve our research for handling the overlapping subset. We also think it is significant to combine the idea of decision-making with the selection of parameters of models, and we plan to do further research for some general methods of parameter selection from the perspective of decision-making.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This paper was supported in part by the National Key Research and Development Program of China (Grant No. 2018YFB2100801) and in part by the Fundamental Research Funds for the Central Universities of China (Grant No. 22120190198).

References

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68*, 90–113.
Alejo, R., Sotoca, J. M., García, V., & Valdovinos, R. M. (2011). Back propagation with balanced mse cost function and nearest neighbor editing for handling class overlap and class imbalance. In *International Work-Conference on Artificial Neural Networks* (pp. 199–206). Springer.
Alejo, R., Valdovinos, R. M., García, V., & Pacheco-Sanchez, J. H. (2013). A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters, 34*, 380–388.
Ali, A., Shamsuddin, S. M., Ralescu, A. L., et al. (2015). Classification with class imbalance problem: a review. *International Journal of Advanced Computer, 7*, 176–204.
Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43*, 3–31.
Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR), 49*, 1–50.
Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
Bunkhumpornpat, C., & Sinapiromsaran, K. (2017). Dbmute: density-based majority under-sampling technique. *Knowledge and Information Systems, 50*, 827–850.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications, 41*, 4915–4928.
Das, B., Krishnan, N. C., & Cook, D. J. (2014). Handling imbalanced and overlapping classes in smart environments prompting dataset. In *Data Mining for Service* (pp. 199–219). Springer.
Das, S., Datta, S., & Chaudhuri, B. B. (2018). Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition, 81*, 674–693.
Denil, M. (2010). *The effects of overlap and imbalance on svm classification*. Master's. Dalhousie University.
Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In *Canadian Conference on Artificial Intelligence* (pp. 220–231). Springer.
Fernandes, E. R., & de Carvalho, A. C. (2019). Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. *Information Sciences, 494*, 141–154.
Fernández, A., García, S., & Herrera, F. (2011). Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In *International Conference on Hybrid Artificial Intelligence Systems* (pp. 1–10). Springer.
García, V., Mollineda, R. A., & Sánchez, J. S. (2008). On the k-nn performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications, 11*, 269–280.
García, V., Mollineda, R. A., Sánchez, J. S., Alejo, R., & Sotoca, J. M. (2007). When overlapping unexpectedly alters the class imbalance effects. In *Iberian Conference on Pattern Recognition and Image Analysis* (pp. 499–506). Springer.
García, V., Sánchez, J., & Mollineda, R. (2007). An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In Iberoamerican Congress on Pattern Recognition (pp. 397–406). Springer.
Gomber, P., Koch, J.-A., & Siering, M. (2017). Digital finance and fintech: current research and future research directions. *Journal of Business Economics, 87*, 537–580.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239.
Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*, 429–449.
Johnson, D. H. (2006). Signal-to-noise ratio. *Scholarpedia, 1*, 2088.
Kingma, D.P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kubat, M., Matwin, S. et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In Icml (pp. 179–186). Nashville, USA volume 97.
Lee, H. K., & Kim, S. B. (2018). An overlap-sensitive margin classifier for imbalanced and overlapping data. *Expert Systems with Applications, 98*, 72–83.
Lee, W., & Xiang, D. (2000). Information-theoretic measures for anomaly detection. In *Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001* (pp. 130–143). IEEE.
Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research, 18*, 1–5. http://jmlr.org/papers/v18/16-365.
Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE.
López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications, 39*, 6585–6608.
Mercier, M., Santos, M. S., Abreu, P. H., Soares, C., Soares, J. P., & Santos, J. (2018). Analysing the footprint of classifiers in overlapped and imbalanced contexts. In *International Symposium on Intelligent Data Analysis* (pp. 200–212). Springer.
Prati, R. C., Batista, G. E., & Monard, M. C. (2004). Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Mexican International Conference on Artificial Intelligence* (pp. 312–321). Springer.
Rodger, J. A. (2019). Quantumis: A qualia consciousness awareness and information theory quale approach to reducing strategic decision-making entropy. *Entropy, 21*, 125.

Rubbo, M., & Silva, L. A. (2018). Prototype selection using self-organizing-maps and entropy for overlapped classes and imbalanced data. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS One, 10*, Article e0118432.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*, 1443–1471.

Stefanowski, J. (2013). Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In *Emerging paradigms in machine learning* (pp. 277–306). Springer.

Vorraboot, P., Rasmequan, S., Chinnasarn, K., & Lursinsap, C. (2015). Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. *Neurocomputing, 152*, 429–443.

Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences, 509*, 47–70.

Vuttipittayamongkol, P., Elyan, E., Petrovski, A., & Jayne, C. (2018). Overlap-based undersampling for improving imbalanced data classification. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 689–697). Springer.

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 665–674).