IMPROVED ADAPTIVE SEMI-UNSUPERVISED WEIGHTED OVERSAMPLING (IA-SUWO) USING SPARSITY FACTOR FOR IMBALANCED DATASETS

HASEEB ALI

A thesis submitted in
fulfillment of the requirement for the award of the
Degree of Masters of Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

DECEMBER 2019

In the name of Allah, Most Gracious, Most Merciful.
I praise and thank Allah.

Special thanks for my beloved father Muhammad Sarfraz Bashir.

For dearest,
Basit Ali, Zahra Qasim, Quratulain, Saima Bashir.
(Brother, brother, sister, sister, aunt)

For their love, support, enthusiasm, encouragement and motivation.

For my supervisor,
Assoc. Prof. Dr. Mohd. Najib bin Mohd. Salleh
For his incredible help, patience, understanding and support.

For all postgraduate members, fellow friends and house mates.

This thesis is dedicated to all of you.

ACKNOWLEDGEMENT

In the name of Allah, the most gracious, the most merciful. With the deepest sense of gratitude and humility, I praise and thank Allah for His blessings uncounted in my life and for His willing, I was able to complete this research successfully. This dissertation would not have been possible without the guidance, help and support of many people contributed and extended their valuable assistance in the preparation and completion of this research. I take this opportunity to express my profound sense of gratitude and respect to all those people.

First and foremost, I would like to express my sincere gratitude to my supervisor, Ass. Prof. Dr. Mohd. Najib bin Mohd. Salleh for his support in possible way, invaluable guidance, useful advice, patience, understanding and encouragement for me to the final level throughout the accomplishment of this research. His enthusiasm and optimism coupled with knowledge and experience, this evidence really rewarding for me. His feedback, editorial comments and suggestions were also invaluable for writing this thesis. I really appreciate it.

In preparing this research, my gratitude is extended to Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research under the Postgraduate Incentive Research Grant.

A special thanks to my beloved family, for their continuous prayer, encouragement, love, support, patience, and care whenever I needed during these challenging days. I dedicate this work to all of you. My overwhelming gratitude to all my friends who have been together with me, thanks for love, care, concern, and support.

Thanks to all staff in Faculty of Computer Science and Information Technology, Center for Graduate Studies, and Research Management Centre (RMC) for their support, cooperation and contribution all the way. Lastly, it is a pleasure to thank all those who have helped either directly or indirectly. Thank you.

ABSTRACT

The imbalanced data problem is common in data mining nowadays due to the skewed nature of data, which impact the classification process negatively in machine learning. For preprocessing, oversampling techniques significantly benefitted the imbalanced domain, in which artificial data is generated in minority class to enhance the number of samples and balance the distribution of samples in both classes. However, existing oversampling techniques encounter through overfitting and over-generalization problems which lessen the classifier performance. Although many clustering based oversampling techniques significantly overcome these problems but most of these techniques are not able to produce the appropriate number of synthetic samples in minority clusters. This study proposed an improved Adaptive Semi-unsupervised Weighted Oversampling (IA-SUWO) technique, using the sparsity factor which determine the sparse minority samples in each minority cluster. This technique consider the sparse minority samples which are far from the decision boundary. These samples also carry the important information for learning of minority class, if these samples are also considered for oversampling, imbalance ratio will be more reduce also it could enhance the learnability of the classifiers. The outcomes of the proposed approach have been compared with existing oversampling techniques such as SMOTE, Borderline-SMOTE, Safe-level SMOTE, and standard A-SUWO technique in terms of accuracy. As aforementioned, the comparative analysis revealed that the proposed oversampling approach performance increased in average by 5% from 85% to 90% than the existing comparative techniques.

ABSTRAK

Masalah data yang tidak seimbang adalah umum dalam perlombongan data pada masa kini disebabkan oleh sifat semulajadi data, di mana ianya memberi impak negatif terhadap proses pengkelasan dalam pembelajaran mesin. Bagi pra-pemprosesan, teknik *oversampling* memberi manfaat secara signifikan kepada domain yang tidak seimbang, di mana data tiruan dijana dalam kelas minoriti untuk meningkatkan bilangan sampel dan mengimbangi pembahagian sampel dalam kedua-dua kelas. Walau bagaimanapun, teknik *oversampling* yang sedia ada menghadapi masalah pemadanan berlebihan dan *over-generalization* yang mengurangkan prestasi pengkelasan. Teknik *oversampling* berasaskan pengklusteran dapat mengatasi masalah ini, namun kebanyakan teknik ini tidak dapat menghasilkan bilangan sampel sintetik yang sesuai dalam kluster minoriti. Oleh itu, kajian ini mencadangkan teknik *Adaptive Semi-Unsupervised Weighted Oversampling* (IA-SUWO) yang lebih baik, dengan menggunakan faktor perenggangan yang menentukan sampel minoriti kecil dalam setiap kluster minoriti. Teknik ini mengambilkira sampel perenggangan minoriti yang jauh dari sempadan keputusan. Sampel-sampel ini juga membawa maklumat penting untuk mempelajari kelas minoriti, jika sampel-sampel ini juga dipertimbangkan untuk *oversampling*, nisbah ketidakseimbangan akan lebih berkurang juga dapat meningkatkan kemampuan pengkelasan. Hasil pendekatan yang dicadangkan telah dibandingkan dengan teknik *oversampling* yang sedia ada seperti teknik *SMOTE, Borderline-SMOTE, Safe-level SMOTE,* dan teknik A-SUWO dari segi ketepatan. Seperti yang dinyatakan di atas, analisis perbandingan menunjukkan bahawa pendekatan *oversampling* yang dicadangkan meningkat secara purata sebanyak 5% dari 85% ke 90% daripada teknik perbandingan sedia ada.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF SYMBOLS AND ABBREVIATIONS

IRUS	-	Inverse Random Undersampling
ACOSAMPLING	-	Ant Colony Optimization Sampling
Fast-CBUS	-	Fast Clustering Based Undersampling
DBSMOTE	-	Density based Synthetic Minority Oversampling Technique
SMOTE	-	Synthetic Minority Oversampling Technique
MSMOTE	-	Modified Synthetic Minority Oversampling Technique
MWMOTE	-	Majority Weight Minority Oversampling Technique
RWO-SAMPLING	-	Random Walk Oversampling
EMOTE	-	Enhanced Minority Oversampling Technique
GMM	-	Gaussian Mixture Model
SIMO	-	Synthetic Informative Minority Oversampling
SMOR	-	Synthetic Minority Oversampling Regression
RBO	-	Radial Based Oversampling
A-SUWO	-	Adaptive Semi-Unsupervised Weighted Oversampling
SOMO	-	Self-Organizing Map Oversampling
NBBag	-	Neighborhood Balanced Bagging
NBBag	-	Neighborhood Balanced Bagging
PSO	-	Practical Swarm Optimization
RBF	-	Radial Basis Function
IPF	-	Iterative–Partitioning Filter
EE	-	Easy Ensemble
CHC	-	Heterogeneous Cataclysmic Mutation
CBIS	-	Cluster Based Instance Selection
RUSboost	-	Random Undersampling Boosting
CUSBoost	-	Clustering based Undersampling Boosting

MLP	-	Multilayer Perceptron
RF	-	Random Forest
LR	-	Logistic Regression
ACO	-	Ant Colony Optimization
SVM	-	Support Vector Machine
MTD	-	Mega-Trend Diffusion
EFSVM	-	Entropy-Based Fuzzy Support Vector Machine
OAO	-	One-Against-One
OAA	-	One-Against-All
NCL	-	Negative Correlation Learning
EOS	-	Entropy Based Oversampling
EHS	-	Entropy Based Hybrid Sampling
EID	-	Entropy-Based Imbalance Degree
NB	-	Naïve Bayes
KNN	-	K-Nearest Neighbor
NN	-	Neural Network
LDA	-	Linear Discriminant Analysis

LIST OF APPENDICES

LIST OF PUBLICATIONS

(i) **H. Ali,** M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq (2019), Imbalance class problems in data mining: A review, Indonesian Journal of Electrical Engineering and Computer Science. vol. 14, no. 3, pp. 1552–1563.

(ii) **H. Ali,** M. N. M. Salleh, K. Hussain, A. Ahmad, U. Ayaz, M. Arshad, N. Rashid, M. khan (2019). A review on data preprocessing methods for class imbalance problem, International Journal of Engineering & Technology, vol. 8, no. 3, pp. 390–397.

(iii) **H. Ali,** M. N. M. Salleh, K. Hussain (2019), Improved Adaptive Semi-Unsupervised Weighted Oversampling using Sparsity Factor for Imbalanced Datasets, International Journal of Advanced Computer Science and Applications, vol. 10, no. 11.

CHAPTER 1

INTRODUCTION

1.1 Overview

Amount of data is increasing day by day along with disparate distributions in many real-time applications. In a dataset, if the quantity of specimens present in one class is more than other class, then this dataset is said to be highly disparate in nature (Wang and Yao, 2012; Chawla *et al.*, 2004). The major class is used to identify any imbalanced dataset that has more number of specimens, whereas the minor class contains less number of specimens (Wang and Yao, 2012). Oftenly, major class expresses the specimens as negative and minor class expresses the specimens as positive (He and Garcia 2009; Van and Khoshgoftaar, 2009). The amount of majority class specimens dominates the minority class specimens by the class's ratios which can be 100 with 1 and 1000 with 1, etc. The dataset having only two classes is known as a binary class, whereas the dataset containing more than two classes is known as multi-class, and both the binary and multi-class datasets suffer from imbalance data problems.

Many real-world domains include imbalance dataset problems, like detecting unreliable telecommunication customers, word pronunciations learning, marking of oil spills in the images of satellite radar, information retrieval, text classification, filtering tasks, the revelation of fake telephone calls and most importantly the medical diagnosis (Raskutti and Kowalczyk, 2004; Wu and Chang 2003). An example of real-world domain is bank transactions which is shown in Figure 1.1, presents the highly imbalanced dataset. In which majority class remarkably dominates the minority class.

In such circumstances, mostly the majority classes bias the classifiers towards themselves and the classifier presents the rates of minority classes classification poorly, eventually, a classifier addresses entirely as majority class and ignores the minority class. To solve problems affiliated with the class imbalance, various techniques have been proposed in the literature (Seiffert *et al.*, 2008).

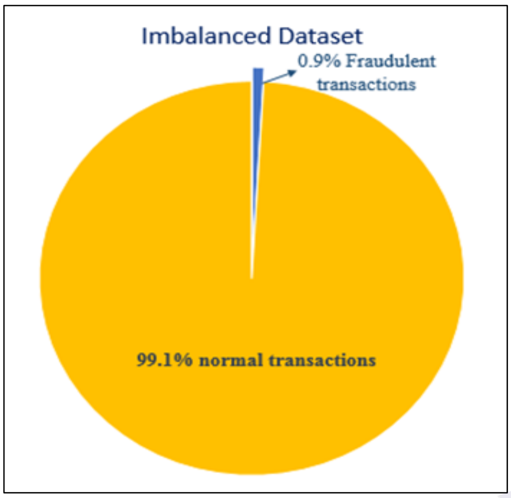

Figure 1.1: Examples of bank transactions imbalanced dataset

Many issues at a time need attention such as multiple classes problem, binary class problems, cost of misclassified class, class overlapping, insignificant disjoints, and size of the imbalanced datasets. In Tan *et al.*, (2003) claims that problems of binary classes related to imbalance data received higher attention than the multi-class imbalance problems. In multi-class imbalance problems, the number of majority and minority classes can be more than one like, one minority class and many majority classes or one majority class and many minority classes (Tan *et al.*, 2003). Despite of the binary class or multi-class imbalances problem, whenever the data is disparate in nature, it will be seriously more daring to proceed with the minority class (Wang and Yao, 2012). This research is considering only binary class imbalance problem due to reason that most of the practitioners used binary class datasets and mostly classifiers are developed for binary class classification purpose.

Owing to the importance of this issue, to solve these problems, there are significant contributions made in developing techniques. These propositions can be categorized into three types according on how they are proceeding with class imbalance. External or data level approach, which is a preprocessing phase of data to rebalance the class distributions to decrease the disparate distribution effect in classification process (Batista *et al.*, 2004; Chawla *et al.*, 2004). Second, the internal or algorithmic level approach creates or modifies the existing algorithms and takes consequences of minor class into consideration (Quinlan, 1991; Wu and Chang, 2005; Zadrozny and Elkan, 2001). Lastly the third one, cost-sensitive approach, that may unite data level and algorithmic level approaches to integrate a variety of misclassification cost for every class in the learning phase (Chawla *et al.,* 2008; Freitas *et al.,* 2007). Other than these basic approaches ensemble of classifiers, feature selection and clustering methods along resampling methods shows significant results for the imbalanced data problems.

In external or data level, resampling is performed in datasets before the classification process to balance the data externally. For example, the specimens of majority class are randomly removed, and specimens of minority class are increased by duplicating to balance the ratio, or in the ideal case, no specimen is created or deleted but the choice of specimens to create or eliminate is informed (Chawla *et al.*, 2002). The preprocessing of data in resampling methods is more effective to balance the class data before the learning process. Many achievements have been made using hybrid sampling techniques. In the algorithmic approach, minority class is taken into consideration and the learner is not allowed to bias for the majority class to overcome the overall cost of misclassification (Joshi *et al.*, 2001). Various methods like variants of SVM, feature selection, one class learning and clustering in algorithmic level are useful to resolve imbalanced problems and misclassification cost. In the cost-sensitive method, it considers all types of costs and mostly focuses on misclassification cost to minimize the total cost in order to make classifier nonbiased (Ling and Sheng, 2008).

Data-level methods or preprocessing techniques are more versatile and can be applied globally, however the algorithmic approaches are more confined to specific classifiers. On the other hand, the cost-sensitive methods are problem-specific, also require to be implemented by the classifier (Galar *et al.*, 2012). There is plenty of work performed by the research community on the preprocessing of data to overcome the imbalanced class issues. Those researchers who do not have much expertise in machine learning usually practice the preprocessing approach which is easier to be employed in single or ensemble models than modifying any learning algorithm (Galar *et al.*, 2012).

The preprocessing of data can be done by resampling of data externally in order to balance the distribution of instances in the majority and minority classes. The most common approaches to balance the ratio are undersampling and oversampling (Chawla *et al.*, 2002). The random undersampling method, which eliminates the specimens of majority class randomly and generates a subset of the primary dataset in a way to balance the ratio. It may lead to the loss of potential data due to eliminating that can be used in the induction process. The random oversampling method increases the number of specimens in the minority class by replicating the existing specimens randomly and generates a superset of the primary data. But it can enhance the chances of overfitting due to replication (Galar *et al.,* 2012). Various oversampling techniques are proposed to overcome the problems faced in oversampling of data (Lin *et al.*, 2017). Synthetic minority over-sampling technique (SMOTE) proposed by Chawla *et al.*, (2002), was the first approach proposed after random oversampling in which new specimens are created in minority class by interpolation of many minority class specimens which reside together. This method avoids the overfitting problem but possibly it creates noisy and borderline specimens which may create problems. Despite its weakness, this method is popular and frequently used by the research community and numerous modifications are made through this method (Vanhoeyveld and Martens, 2018). Although, many achievements have been made until now using hybrid, or novel proposed sampling techniques, still there are some issues that need to be solved; such as, over-generalization, overlapping of minority instances with majority instances, imbalance data within-class which is in between the different sub-clusters of minority class, some of the sub-clusters have a large number of samples and do not need to be oversampled more and some of the clusters having a small number of samples which can be neglected by classifier need more weight to be oversampled (Nekooeimehr and Lai-Yuen, 2016).

So that, external or data-level methods can be composed of random or informed. The random method only determine the samples to be duplicated or eliminated (Chawla *et al.,* 2004). In informal method it consider the distribution of samples, informed methods take into account the critical area of the input space, like safe areas according to Bunkhumpornpat *et al.*, (2009), sparse areas Nickerson *et al.*, (2001) or areas which are closer to decision boundary (Han *et al.*, 2005). The clustering based oversampling techniques partitions the input space first and then applies sampling method on the dataset to adjust the size of different clusters to show good potential. Consequently, the generation of noise can be avoided and informed methods tackle the imbalances problem within classes significantly.

It is noteworthy that it is not essential and compulsory to balance the number of samples in minority and majority classes exactly, all of the resampling techniques allow to re-sample to any desired ratio. Different sample ratios are recommended by Zhong *et al.,* (2013) for different data size. It depends on the type of examples in minority class and how they impact on learning classifier from imbalanced data (Napierala and Stefanowski, 2016). It has been tried to decide the best sampling rates automatically by several authors for different problem settings and their imbalanced ratios (IR).

1.2 Problem Statement

In existing oversampling techniques, generally these techniques generate new synthetic samples in minority class by using two approaches, first approach is that these techniques generate synthetic samples between the candidate minority sample and its NN-nearest neighbors (Chawla *et al.*, 2002; Han *et al.*, 2005; He *et al.*, 2008). Second approach is they generate synthetic samples between a candidate sample and its NN-nearest neighbors from the same sub-cluster (Barua *et al.*, 2014). These both approaches led to the generation of overlapping of generated samples with the majority class samples. As in the first approach, NN-nearest neighbor might be far from this selected sample and the generated synthetic sample fall into the incorrect region and overlapped with the majority class samples or act as the noise. While in the second approach, it generate synthetic sample from the same cluster which may include the majority class samples within it and arise the over-generalization problem. Overlapping of synthetic instances can deteriorate the performance of the classifiers significantly (Beyan and Fisher, 2015) (Barua *et al.,* 2014).

Moreover, MWMOTE technique (Barua *et al.,* 2014), oversampled the data according to the weights assigned to the samples of minority clusters based on the Euclidean distance of minority cluster from the majority samples. However, it neglects or assigns no weight to small concepts of the minority samples which are far from the majority samples even if they contain important information. These small concepts referred to as within-class imbalance. It is important to oversample all the minority sub-cluster to overcome the within-class imbalance problem, neglecting these small concepts of a minority class or within-class imbalances, it will bias the classifier towards oversampled ones.

These problems was overcome by Nekooeimehr and Lai-Yuen (2016) which proposed, Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO). This technique significantly avoids over-generalization by using semi-unsupervised clustering approach and solve within-class imbalance problem by considering small concepts and misclassification rate of minority samples. However, standard A-SUWO assigns weights to minority samples for oversampling according to the Euclidean distance of minority samples (in each minority sub-cluster) from the majority samples. Which means it oversample only those minority samples in each minority sub-clusters that are near the decision boundary or closer to the majority class samples and neglect those minority samples that are far from the majority samples or behind these samples. These minority samples also carry the important information about the minority class which can improve the learnability of the classifier, also if these samples can be used for oversampling it will increase the number of synthetic samples which significantly reduce between the class and within-class imbalance problem.

Hence, if sampling weights also corresponds to minority samples according to their sparsity (density) factor along with the closeness factor which is measured by the Euclidean distance, can yield more synthetic samples (Douzas *et al.*, 2018) which reduce the between the class imbalances problem. To overcome the within-class imbalance problem, the sampling weight depends on how dense a single cluster as compared to how dense the all sub-clusters are on average. In this way, sub-cluster with sparse minority samples will be oversampled and the number of samples in all sub-clusters become equivalent. The equivalent number of samples in each cluster will reduce the within-class imbalance problem and improve the accuracy for the classification of minority samples (Douzas and Bacao, 2017).

1.3 Aim of Study

The aim of this study is to make an improved preprocessing technique that makes a classifier non-biased and give maximum accurate results of a dataset provided for the learning process. As general classification algorithms are developed to tackle balanced class distributions, imbalanced data learning continues to be a challenging and common problem for all classifiers. Although several approaches have been proposed but most of them generate unnecessary noise, overfitting by oversampling methods or cannot detect hard-to-learn instances (Tang and He, 2015; Douzas *et al.*, 2018). Thus the predictive capacity of classification algorithms is inadequate by imbalanced class data. Many algorithms with the aim to maximize the classification accuracy are biased towards majority class because a higher classification accuracy can be achieved by a classifier when it does not predict even a single minority class sample properly. Applications on real-world domains including fraud detection in banking and rare medical diagnoses exhibit naturally minority class. Our proposed improved method will enhance the minority class samples generation with avoiding over-generalization and within-class imbalance problem.

1.4 Objectives of Study

This study embarks on the following objectives:

i) To propose an improved Adaptive Semi-Unsupervised Weighted Oversampling (IA-SUWO) for imbalanced datasets by introducing the sparsity factor in it.

ii) To develop the improved IA-SUWO technique by embedding the sparsity factor in standard A-SUWO.

iii) To evaluate the performance of the IA-SUWO technique based on precision, F-measure and ROC, and benchmark the result with standard A-SUWO technique and other oversampling techniques.

1.5 Research Questions

This study consider the following research questions:

i) What is the sparsity factor and how it will improve this technique?

ii) How this sparsity factor work for the imbalances problem?

iii) Outcomes of the proposed technique is improved or not? If improved then how much?

1.6 Scope of Study

This research is focused on following solutions to the problem.

i) Analyze the selection of samples in the dataset for oversampling and improve the selection process. 3 datasets used are for this study named as, Iris, Wine and Glass.

ii) Determine the previous synthetic sample generation schemes and improve the synthetic sample generation by using improved Adaptive Semi-Unsupervised Weighted Oversampling.

iii) Validate the performance of the proposed IA-SUWO technique with the standard A-SUWO technique and other oversampling techniques names as, SMOTE, Safe-level SMOTE, Borderline-SMOTE for solving imbalanced data problems before and after the classification by using four classifiers: Naïve Bayer, K-Nearest Neighbor (KNN), Logistic regression and Neural Network.

1.7 Significance of Study

An immense benefit of this research work is that it provides a method that frequently outperforms the available widely used oversampling approaches such as random oversampling and SMOTE. It contributes to the diagnosis of diseases, prevention of credit card frauds and detection of abnormalities in environmental observations. A common problem in the classification process, imbalanced data is demonstrated naturally in many important real-world applications. This research study proposed the oversampling method that can be applied to any dataset and independently of the chosen classifier, its potential influence will be significant. The effectiveness of the algorithm is proficient in focusing more on overall accuracy and reducing misclassification.

REFERENCES

Aamir, M., Wahid, F., Mahdin, H., & Nawi, N. M. (2019). An efficient normalized restricted Boltzmann machine for solving multiclass classification problems. *International Journal of Advanced Computer Science and Applications*, *10*(8), 416–426.

Addabbo, A. D., & Maglietta, R. (2015). Parallel selective sampling method for imbalanced and large data classification. *Pattern Recognition Letters*, *62*, 61–67.

Agrawal, A., Viktor, H. L., & Paquet, E. (2015). SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling. *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, *01*, 226–234.

Atif, M., Kittler, J., & Yan, F. (2012). Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, *45*(10), 3738–3750.

Avenue, M., Hill, M., Cohen, W. W., Of, C., & Pruning, R. (1995). Fast Effective Rule Induction. *Proceedings of the Twelfth International Conference*, 115–123.

Babu, S., & Ananthanarayanan, N. R. (2017). EMOTE: Enhanced Minority Oversampling Technique. *Journal of Intelligent and Fuzzy Systems*, *33*(1), 67–78.

Barandela, R., Sánchez, J. S., & Valdovinos, R. M. (2003). New Applications of Ensembles of Classifiers. *Pattern Analysis and Applications*, *6*(3), 245–256.

Barua, S., Islam, M., Yao, X., & Murase, K. (2014). MWMOTE — Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering*, *26*(2), 405–425.

Batista, G., Prati, R., & Monard, M., (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter - Special Issue on Learning from Imbalanced Datasets, 6*(1), 20–29.

Beyan, C., & Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition, 48*(5), 1653–1672.

Błaszczyński, J., Deckert, M., Stefanowski, J., & Wilk, S. (2010). Integrating selective pre-processing of imbalanced data with Ivotes ensemble. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6086 LNAI*, 148–157.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE : Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem. *Advances in Knowledge Discovery and Data Mining, PAKDD 2009*, 475–482.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: Density-based synthetic minority over-sampling technique. *Applied Intelligence*, *36*(3), 664–684.

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.

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, *17*(2), 225–252.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *Proceeding. Knowledge Base Discovery Databases.*, 107–119.

Chawla, N. V, Japkowicz, N., & Drive, P. (2004). Editorial : Special Issue on Learning from Imbalanced Data Sets. *Sigkdd Explorations*, *6*(1), 2000–2004.

Chen, X., Kang, Q., Zhou, M., & Wei, Z. (2016). A novel under-sampling algorithm based on Iterative-Partitioning Filters for imbalanced classification. *IEEE International Conference on Automation Science and Engineering*, *2016-Novem*, 490–494.

Cieslak, D. A., Chawla, N. V., & Striegel, A. (2006). Combating imbalance in network intrusion datasets. *2006 IEEE International Conference on Granular Computing*, 732–737.

Datta, S., & Das, S. (2015). Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Networks*, *70*, 39–52.

Del Castillo, M. D. (2004). A multistrategy approach for digital text categorization from imbalanced documents. *ACM SIGKDD Explorations Newsletter*, *6*(1), 70.

Devi, D., Biswas, S. K., & Purkayastha, B. (2019). Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. *Connection Science*, *31*(2), 105–142.

Dong, Y., & Wang, X. (2011). A new over-sampling approach: Random-SMOTE for learning from imbalanced datasets. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7091 LNAI*, 343–352.

Douzas, G., & Bacao, F. (2017). Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications*, *82*, 40–52.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1–20.

Fan, Q., Wang, Z., Li, D., Gao, D., & Zha, H. (2017). Entropy-based fuzzy support vector machine for imbalanced datasets. *Knowledge-Based Systems*, *115*, 87–99.

Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, *53*(1), 226–233.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Freitas, A., da Costa Pereira, A., & Brazdil, P. (2007). Cost-Sensitive Decision Trees Applied to Medical Data. *DaWaK*, *4654*, 303–312.

Freund, Y, & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proc. of the 13th Int. Conf.*, 148–156.

Freund, Yoav, & Schapire, R. E. (1997). A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Galar, M., Fern, A., Barrenechea, E., & Bustince, H. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE transactions on systems, man, and cybernetics—part c: applications and reviews*, *42*(4), 463–484.

Gao, M., Hong, X., Chen, S., & Harris, C. J. (2011). Neurocomputing A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. *Neurocomputing*, *74*(17), 3456–3466.

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.

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE : A New Over-Sampling Method in. *Springer-Verlag Berlin Heidelberg*, 878–879.

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *Annals of Statistics*, *26*(2), 451–471.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, 1322–1328.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hu, Shengguo, Yanfeng Liang, Ying He, L. M. (2009). MSMOTE : Improving Classification Performance when Training Data is imbalanced. *Second International Workshop on Computer Science and Engineering*, 627–631.

Jerzy, B., & Stefanowski, J. (2014). Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, *150*, 529–542.

Jin, R., & Zhang, J. (2007). Multi-class learning by smoothed boosting. *Machine Learning*, *67*(3), 207–227.

Joshi, M. V., Kumar, V. and R. C. A. (2001). Evaluating Boosting Algorithms to Classify Rare Classes : Comparison and Improvements. *In First IEEE International Conference on Data Mining*, 257–264.

Kang, Q., Chen, X. S., Li, S. S., & Zhou, M. C. (2017). A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification. *IEEE Transactions on Cybernetics*, *47*(12), 4263–4274.

Kittler, J., Hater, M., & Duin, R. P. W. (1998). Combining classifiers. *Proceedings - International Conference on Pattern Recognition*, *20*(3), 226–239.

Koziarski, M., Krawczyk, B., & Woźniak, M. (2019). Radial-Based oversampling for noisy imbalanced data classification. *Neurocomputing*, (2019).

Krawczyk, B., Woźniak, M., & Herrera, F. (2015). Weighted one-class classification for different types of minority class examples in imbalanced data. *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining, Proceedings*, 337–344.

Li, L., He, H., & Li, J. (2019). Entropy-based Sampling Approaches for Multi-class Imbalanced Problems. *IEEE Transactions on Knowledge and Data Engineering*, 1–12.

Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, *409–410*, 17–26.

Ling, C. X., & Sheng, V. S. (2008). Cost-Sensitive Learning and the Class Imbalance Problem. *Encyclopedia of Machine Learning*, 231–235.

Ling, C. X., Sheng, V. S., & Yang, Q. (2006). Test strategies for cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1055–1067.

Liu, J., Zhou, X., Li, D., Li, X., Dong, Z., & Wang, S. (2005). *Advanced Data Mining and Applications*. *Lecture Notes in Artificial Intelligence*.

Liu, Y., Yu, X., Xiangji, J., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing and Management*, *47*(4), 617–631.

Liu, Z., & Wu, D. (2019). Unsupervised Ensemble Learning for Class Imbalance Problems. *Proceedings 2018 Chinese Automation Congress, CAC 2018*, 3593–3600.

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class imbalance problem in data mining: review. *International Journal of Computer Science and Network*, *2*(1), 83–87.

Majid, A., Ali, S., Iqbal, M., & Kausar, N. (2014). Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. *Computer Methods and Programs in Biomedicine*, *113*(3), 792–808.

Maldonado, S., & López, J. (2014). Imbalanced data classification using second-order cone programming support vector machines. *Pattern Recognition*, *47*(5), 2070–2079.

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, *257*, 331–341.

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. *Icml-1999*, (January), 258–267.

Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*(3), 563–597.

Nejatian, S., Parvin, H., & Faraji, E. (2018). Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. *Neurocomputing*, *276*, 55–66.

Nekooeimehr, I., & Lai-Yuen, S. K. (2016). Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Systems with Applications*, *46*, 405–416.

Nickerson, A. S., Japkowicz, N., & Milios, E. (2001). Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets. *In Proceedings of the Eighth International Workshop on AI and Statitsics*, (2001), 5.

Ofek, N., Rokach, L., Stern, R., & Shabtai, A. (2017). Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing*, *243*, 88–102.

Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, *106*, 15–29.

Popel, M. H., Hasib, K. M., Habib, S. A., & Shah, F. M. (2018). A Hybrid Under-Sampling Method to Classify Imbalanced Data candidates ' declaration. *2018 21st International Conference of Computer and Information Technology (ICCIT)*, (May 2018), 1–7.

Quinlan, J. R. (1991). Improved estimated for the acccuracy of small disjuncts. *Machine Learn*, *6*(1991), 93–98.

Raskutti, B., & Kowalczyk, A. (2004). Extreme re-balancing for SVMs: a case study. *ACM SIGKDD Explorations Newsletter*, *6*(1), 60–69.

Rayhan, F., Ahmed, S., Mahbub, A., Jani, R., Shatabda, S., & Farid, D. M. (2018). CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification. *2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2017*, 1–5.

Richhariya, B., & Tanveer, M. (2018). A robust fuzzy least squares twin support vector machine for class imbalance learning. *Applied Soft Computing Journal*, *71*, 418–432.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, *5*(Jan), 101–141.

Rout, N., Mishra, D., & Mallick, M. K. (2018). Handling Imbalanced Data: A Survey. *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications,Springer.*, (January), 431–443.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE – IPF :

Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184–203.

Schapire, R. E. (1990). The Strength of Weak Learnability (Extended Abstract). *Machine Learning*, *5*, 197–227.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, *37*(3), 297–336.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). A comparative study of data sampling and cost sensitive learning. *Proceedings - IEEE International Conference on Data Mining Workshops, ICDM Workshops 2008*, 46–52.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *40*(1), 185–197.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, *48*(5), 1623–1637.

Tan, A. C., Gilbert, D., & Deville, Y. (2003). Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics. International Conference on Genome Informatics*, *14*(July), 206–217.

Tang, B., & He, H. (2015). KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning. *2015 IEEE Congress on Evolutionary Computation, CEC 2015 - Proceedings*, 664–671.

Tianlun, Z., & Xi, Y. (2018). G-SMOTE: a gmm-based synthetic minority oversampling technique for imbalanced learning. *Arxiv:1810.10363v1, a Preprint*.

Tiwari, D. (2014). Handling Class Imbalance Problem Using Feature Selection. *International Journal of Advanced Research in Computer Science & Technology*, *2*(2), 516–520.

Tsai, C. F., Lin, W. C., Hu, Y. H., & Yao, G. T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, *477*, 47–54.

Van H, J., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. Data and Knowledge Engineering, 68(12), 1513–1542.

Vanhoeyveld, J., & Martens, D. (2018). Imbalanced classification in sparse and large behaviour datasets. Data Mining and Knowledge Discovery, 32(1), 25–82.

Vluymans, S., Fernández, A., Saeys, Y., Cornelis, C., & Herrera, F. (2018). Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: a fuzzy rough set approach. *Knowledge and Information Systems*, *56*(1), 55–84.

Voorhees, E. M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing and Management*, *22*(6), 465–476.

Wang, K., Makond, B., Chen, K., & Wang, K. (2014). A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing Journal*, *20*, 15–24.

Wang, Q., Luo, Z., Huang, J., Feng, Y., & Liu, Z. (2017). A novel ensemble method for imbalanced data learning. *Computational Intelligence and Neuroscience*, *2017*, 1–11.

Wang, S., Chen, H., & Yao, X. (2010). Negative Correlation Learning for Classification Ensembles. *Proceeding Internation Joint Conference Neural Network*, 2893–2900.

Wang, S., & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 324–331.

Wang, S., & Yao, X. (2012). Multiclass Imbalance Problems : Analysis and Potential Solutions. *IEEE transactions on systems, man, and cybernetics*, *42*(4), 1119–1130.

Wasikowski, M., Chen, X., & Member, S. (2010). Combating the Small Sample Class Imbalance Problem Using Feature Selection. I*EEE transactions on knowledge* *and data engineering*, *22*(10), 1388–1400.

Wu, G., & Chang, E. Y. (2005). KBA: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 786–795.

Wu, G., & Chang, E. Y. E. (2003). Class-boundary alignment for imbalanced dataset learning. *The Twentieth International Conference on Machine Learning (ICML), Workshop on Imbalanced Data Sets*, (1), 49–56.

Wu, X. (2005). 10 Challenging Problems in Data Mining Developing a Unifying Theory of Data Min- ing Scaling Up for High Dimensional Data and High Speed Data Streams, 1–9.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2008). *Top 10 algorithms in data mining*. *Knowledge and Information Systems* (Vol. 14).

Yan, R., Liu, Y., Jin, R., & Hauptmann, A. (2003). On predicting rare classes with svm ensembles in scene classification. *Landscape*, 21–24.

Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing*, *105*, 3–11.

Yu, H., Ni, J., & Zhao, J. (2013). ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing*, *101*, 309–318.

Zadrozny, B., & Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '01*, 204–213.

Zhai, J., Zhang, S., & Wang, C. (2017). The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers. *International Journal of Machine Learning and Cybernetics*, *8*(3), 1009–1017.

Zhang, H., & Li, M. (2014). RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, *20*(1), 99–116.

Zheng, Zhaohui, Xiaoyun Wu, R. S. (2004). Feature Selection for Text Categorization on Imbalanced Data. *Sigkdd Explorations.*, *6*(1), 80–89.

Zhong, W., Raahemi, B., & Liu, J. (2013). Classifying peer-to-peer applications using imbalanced concept-adapting very fast decision tree on IP data stream. *Peer-to-Peer Networking and Applications*, *6*(3), 233–246.

Zhu, T., Lin, Y., Liu, Y., Zhang, W., & Zhang, J. (2019). Minority oversampling for imbalanced ordinal regression. *Knowledge-Based Systems*, *166*, 140–155.