Classification Imbalanced Data Sets: A Survey

Shrouk El-Amir
Department of Computer Science, Faculty of Computers and Information, Zagazig University, Zagazig, Sharkia, Egypt

Heba El-Fiqi
Department of Computer Science, Faculty of Computers and Information, Zagazig University, Zagazig, Sharkia, Egypt

ABSTRACT
Unbalanced data, a snag often found in real-world applications, can seriously adversely affect machine learning algorithms ' classification efficiency. Various tries are made to classify unbalanced data sets. In order to face the imbalanced data sets snag, we should rebalance them artificially through machine learning classifiers by oversampling and/or under-sampling.

Keywords
Imbalance dataset, sampling, cost-sensitive learning, imbalance ratio

1. INTRODUCTION
imbalanced data sets are deemed a special case of data science snags. This shape of snags is based on an extension or restriction of the original data science problem. This problem is also considered an extended supervised learning model. This snag takes place when the distribution of class is extra lower than the other classes. Usually, we deal with two terms which are called (positive class /minority class) and (negative class /majority class). this type of data creates a new problem in the field of data mining in all the cosmos areas[1-4]because the standard machine learning algorithms deal with unbalanced data without sensitivity to the unbalanced distribution of the classes which lead to a bias towards the (negative class/majority class). Sometimes this problem is not actually an issue for some applications because the norm classification algorithms depend on the balanced of the distributions of the class .but it is an actual issue for real-world applications like text classification [5],telecommunications, finances, oil spills detection using radar[2],e-mail foldering[6],the fraudulent calls detection[3],medical diagnosis[7],etcetera because as a result of during this standing, the extra interest of the learning is concentrated on the minority classes instead of the majority classes which wants to be correctly identified in these applications.

2. IMBALANCE PROBLEM
The imbalanced datasets snag comes when the cases of one class(positive class/minority class) are fewer than the cases associated with the other classes(negative class/majority class). For instance, in medical diagnosis trouble where the disease conditions are completely rare according to natural cases, the central target is to discover infection folks with diseases[8].

Using the international measures of quality for the building of the model show that research on imbalanced a group of data class distribution issue is serious in data mining. Thither two points to consider: (1) the Imbalanced or suddenly change direction class distribution issue is Widespread in many varieties of fields of large significance in the data processing community. Announced applications contain medical diagnosis[7], oil spills detection using radar[2], the fraudulent calls detection[3], risk management, text classification et cetera.; and (2) most common classification learning applications are prepared to be unsuitable during facing the imbalanced class problem. These classification applications such as decision trees, support vector machines[9-11], backpropagation neural networks[9], Bayesian network, nearest neighbor[12] and also the recently mentioned associative classification applications[13, 14]. However, some machine learning algorithms have been modified to handle the imbalanced classification problem as Random Forest[15], Adaptive Boosting (Adaboost)[16, 17], Gradient Boosting [18], Entropy-based fuzzy support vector machine [19].

The different approaches for facing the imbalanced data issue in the literature can be divided into three groups: data level, algorithm level and cost-sensitive approaches[1]. At data level approaches(resampling techniques), the data is modified by resampling the main training data space for better balancing of the two classes [20, 21]. In the next part will we discuss these methods[22].

3. SAMPLING METHODS
we perform the preprocessing stage which is often performed before classifiers start the training process to obtain better input data for converting the imbalanced data to balanced data for preventing the skewed class distribution from biasing toward the majority class. Resampling techniques are divided into three elements (Over-sampling methods, Under-sampling methods, Hybrid methods)depending on what method will be used to balance the class distribution.

Over-sampling methods: to prevent the skewed class distribution from biasing toward the majority class. We create new minority class samples in a random way. This technique suffers from problem ,it may cause overfitting and produce an extra computational overhead. Oversampling is also categorized into two forms: Informative Oversampling and Random Oversampling. Informative Oversampling method industrial produces minority class instances based on a pre-determined criterion[23]. Random Oversampling is the method which copies positive instances from the original data set in a random way till the number of positive instances close to the number of negative instances.

Under-sampling methods: for converting the imbalanced data to balanced data. We delete some of the majority class samples in a random way. undersampling is also divided into two shapes: Informative undersampling and Random undersampling. Informative Oversampling method selects majority class instances based on a pre-determined criterion to make the data more balanced. Random undersampling is the method which deletes positive examples from the original data set in a random way till the number of positive examples close to the number of negative examples. This technique suffers from problem ,it may cause loss information because the deleted examples may contain useful information.

Hybrid methods: the integration between the under-sampling methods and the over-sampling methods.

There are numerous ways in which random sampling is improved, such as Tomek links[24], Condensed Nearest Neighbor Rule[25], One-sided selection , US-CNN + TL , Neighborhood Cleaning Rule (NCL), Undersampling Based on Clustering **(SBC)** and Class Purity Maximization **(CPM)** etc .

4. COST-SENSITIVE LEARNING

Regarding the algorithm level techniques, the normal learning algorithms are adjusted to focus on a decision threshold biased towards the positive class. Cost-sensitive techniques merge both the data level and algorithm level techniques by giving higher misclassification costs to positive instances and Reducing the cost in a comprehensive way. The misclassification costs have been ordinarily represented by a cost matrix C in which C (i, j) indicates the costs of classifying a sample belonging class i to class j[26]. The popular Cost-sensitive methods are Gradientboost [18]and Adaboost (Adaptive Boosting) [16, 17], which allocate weight to the samples through training. So, in each iteration, the weight of misclassified samples increases, and correctly classified decreases. Contained weights correction imposes on the learning process to be attending more on misclassified in subsequent iterations. In the case of imbalanced data, the minority class is incorrectly classified, so, the boosting will increase the accuracy of the results. Chao Chen[27] proposed two tracks based on the original RF to face the imbalanced datasets snag. The first way is called Weighted RF which merges class weights into the RF classifier. The second way is called Balanced RF, which combines the sampling technique and the ensemble idea. It deletes some of the majority instances from the major class and grows trees based on the new balanced datasets. DengjuYao[28]proposed an improved RF, which adopted a new sampling method to the original RF .he took some of the subsets from the majority class in a random way and made these taken subsets belong to the minority class and train the RF on the new balanced datasets.

5. COMPARISON

The two methods for handling unbalanced data sets are sampling and cost-sensitive. The quality of these methods depends on the type of data we use. The factors that affect these methods ' quality are 1) data set size, 2) class imbalance ratio in the dataset. Table 1 shows which method in the given cases will perform perfectly and worst:

Table 1. Table captions should be placed above the table

	Oversampling	**Undersampling**	**Cost-sensitive**
Large data size	Worst	Worst	Perfect
Small data size	Perfect	less perfect	Worst
drawbacks	Increases the Processing time.	Loss of data	Need to provide the value of misclassification cost

6. LITERATURE SURVEY

Joonho Gong and Hyunjoong Kim Proposed a new hybrid sampling strategy for improving the performance classifiers using the integration between Rose sampling and undersampling under a boosting form which described in paper RHSBoost: Improving classification performance in imbalance data[29].

Wenhao Xie, Gongqian Liang, Zhonghui Dong, Baoyu Tan, and Baosheng Zhang proposed a novel Random-SMOTE (AKN-Random-SMOTE) algorithm which is based on the samples' selection strategy. The support vectors are extracted by the improved alien k-neighbors algorithm, and the oversampling strategy is only performed to the boundary decision samples of the minority class rather than all the minority samples for making the data more balance [30].

Kesinee Boonchuay , Krung Sinapiromsaran, and Chidchanok Lursinsap proposed a novel impurity measure(minority entropy) that is performed on decision tree induction for determining the best split. A decision tree algorithm which uses this novel impurity measure shows better performance over the distinct class-based splitting measure, a top-down decision tree , Hellinger distance decision tree, C4.5 and asymmetric entropy on UCI imbalanced data sets compared with F-measure and geometric mean which described in paper Decision tree induction based on minority entropy for the class imbalance problem[31].

Qi Fan, Zhe Wang, Dongdong Li, Daqi Gao, and Hongyuan Zha Proposed an Entropy-based fuzzy support vector machine algorithm which proposes a novel fuzzy membership and assigns it to the training samples for reflecting the various importances of these samples which described in paper Entropy-based Fuzzy Support Vector Machine for Imbalanced Datasets[19].

Table 2:sets of algorithms for handling with imbalanced dataset

Sr No	**Paper**	**Year**	**Contents**
1	Imbalance class problems in data mining: A review	Indonesian Journal of Electrical Engineering and Computer Science,2019	Data level solutions,Algorithmic level solution,Cost Sensitive learning.

2	A Review on Handling Imbalanced Data	International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018.	Over sampling , under sampling and hybrid methods
3	Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review	IOP conference series: earth and environmental science,2017	presents review of synthethic over sampling methods for handling imbalance data problem
4	Imbalanced data classification using complementary fuzzy support vector machine techniques and SMOTE	International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2017.	A hybrid sampling technique(CMTFSVM+SMOTE) Which achieves the best G-mean and AUC on the imbalanced real world data based on the optimised membership function .
5	Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance	Pattern Recognition Letters,2017	A hybrid sampling method that integrates Condensed Nearest Neighbor and Tomek-link undersampling techniques with machine learning classifiers(Back Propagation Neural Network , K-Nearest- Neighbor , Support Vector Machine and Naïve Bayes)
6	A New Approach for Handling Imbalanced Dataset using ANN and Genetic Algorithm	International Conference on Communication and Signal Processing (ICCSP). IEEE, 2016	Algorithmic level solution is presented using a hybrid approaches between ANN and Genetic Algorithm
7	Imbalanced classification using genetically optimized cost sensitive classifiers	IEEE Congress on Evolutionary Computation (CEC). IEEE, 2015	Propose a cost sensitive approach which automatically generating optimized cost matrices using a genetic algorithm
8	Oversampling Method for Imbalanced Classification	Computing and Informatics,2016	propose a new oversampling method SNOCC that compensates the drawbacks of SMOTE.
9	Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification	Information Sciences,2017	Presents a cost sensitive strategy by proposing A new weight that is assigned to a weighted support vector machine(SVM) as a weak learner of the AdaBoost Algorithm
10	Entropy-Based Classifier Enhancement to Handle Imbalanced Class Problem	Procedia Computer Science,2017	Presents an algorithm level approach which proposes a simple growing for the entropy-based classifiers to handle the imbalanced class problem
11	Classification with class imbalance problem: A Review	Int. J. Advance Soft Compu. Appl,2015	Presents Data level sampling , Algorithm level and Cost Sensitive strategies

7. CONCLUSION

Data imbalance is the most common snag. Standard classification algorithms fail to classify unbalancing data in a perfect way, so we need to prepare and balance the data. Cost-sensitive and sampling are strategies for dealing with the imbalanced data snag. At the data level, the most diffused technique of handling imbalanced data is sampling. Over-sampling is obviously more efficient for locally-based classifiers than under-sampling, while some under-sampled strategies outperform over-sampling when using global learning classifiers.

8. REFERENCES

[1] Aurelio, Y.S., et al., Learning from imbalanced data sets with weighted cross-entropy function. Neural Processing Letters, 2019: p. 1-13.

[2] Ali, Z., et al. Empirical Study of Associative Classifiers on Imbalanced Datasets in KEEL. in 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). 2018. IEEE.

[3] Arafat, M., A. Qusef, and G. Sammour. Detection of Wangiri Telecommunication Fraud Using Ensemble Learning. in 2019 IEEE Jordan International Joint

Conference on Electrical Engineering and Information Technology (JEEIT). 2019. IEEE.

[4] Wang, H., Utilizing Imbalanced Data and Classification Cost Matrix to Predict Movie Preferences. arXiv preprint arXiv:1812.02529, 2018.

[5] Maheshwari, S., R. Jain, and R. Jadon, A Review on Class Imbalance Problem: Analysis and Potential Solutions. International Journal of Computer Science Issues (IJCSI), 2017. 14(6): p. 43-51.

[6] Bermejo, P., J.A. Gámez, and J.M. Puerta, Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. Expert Systems with Applications, 2011. 38(3): p. 2072-2080.

[7] Huang, C., et al., Deep imbalanced learning for face recognition and attribute prediction. IEEE transactions on pattern analysis and machine intelligence, 2019.

[8] Chan, R., et al., Application of decision rules for handling class imbalance in semantic segmentation. arXiv preprint arXiv:1901.08394, 2019.

[9] Potharlanka, J.L. and M.P. Turumella, Weighted SVMBoost based Hybrid Rule Extraction Methods for Software Defect Prediction. International Journal of Rough Sets and Data Analysis (IJRSDA), 2019. 6(2): p. 51-60.

[10] Raskutti, B. and A. Kowalczyk, Extreme re-balancing for SVMs: a case study. ACM Sigkdd Explorations Newsletter, 2004. 6(1): p. 60-69.

[11] Folorunso, S. and A. Adeyemo, Empirical Study of Enhanced Sampling Schemes with Ensembles to Alleviate the Class Imbalance Problem.

[12] Schubach, M., et al., Variant relevance prediction in extremely imbalanced training sets. F1000Research, 2017. 6: p. 1392.

[13] Abdullah, Z., et al. 2M-SELAR: A Model for Mining Sequential Least Association Rules. in Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015). 2019. Springer.

[14] Wu, G.P. and K.C. Chan. Clustering driving trip trajectory data based on pattern discovery techniques. in 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA). 2018. IEEE.

[15] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

[16] Freund, Y. and R. Schapire, A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 1999. 14(771-780): p. 1612.

[17] Liu, H. and M. Cocea, Granular computing-based approach of rule learning for binary classification. Granular Computing, 2019. 4(2): p. 275-283.

[18] Xia, Y., et al., A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 2017. 78: p. 225-241.

[19] Fan, Q., et al., Entropy-based fuzzy support vector machine for imbalanced datasets. Knowledge-Based Systems, 2017. 115: p. 87-99.

[20] Chawla, N.V., et al., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002. 16: p. 321-357.

[21] Garcı, S., et al., Evolutionary-based selection of generalized instances for imbalanced classification. Knowledge-Based Systems, 2012. 25(1): p. 3-12.

[22] Yang, P., et al., Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. IEEE transactions on cybernetics, 2013. 44(3): p. 445-455.

[23] Ramyachitra, D. and P. Manikandan, Imbalanced dataset classification and solutions: a review. International Journal of Computing and Business Research (IJCBR), 2014. 5(4).

[24] Tomek, I., Two modifications of CNN. IEEE Trans. Systems, Man and Cybernetics, 1976. 6: p. 769-772.

[25] Hart, P., The condensed nearest neighbor rule (Corresp.). IEEE transactions on information theory, 1968. 14(3): p. 515-516.

[26] Zhao, P., et al. Cost-sensitive online classification with adaptive regularization and its applications. in 2015 IEEE International Conference on Data Mining. 2015. IEEE.

[27] Chen, C., A. Liaw, and L. Breiman, Using random forest to learn imbalanced data. University of California, Berkeley, 2004. 110(1-12): p. 24.

[28] Yao, D., J. Yang, and X. Zhan, An improved random forest algorithm for class-imbalanced data classification and its application in PAD risk factors analysis. The Open Electrical & Electronic Engineering Journal, 2013. 7(1).

[29] Gong, J. and H. Kim, RHSBoost: Improving classification performance in imbalance data. Computational Statistics & Data Analysis, 2017. 111: p. 1-13.

[30] Xie, W., et al., An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data. Mathematical Problems in Engineering, 2019. 2019.

[31] Boonchuay, K., K. Sinapiromsaran, and C. Lursinsap, Decision tree induction based on minority entropy for the class imbalance problem. Pattern Analysis and Applications, 2017. 20(3): p. 769-782