Annals of Mathematics and Computer Science
ISSN: 2789-7206
Vol 7 (2022) 45-53

IMBALANCED DATA: A CONCISE SURVEY

DILSHOD ISKHAKOV[1*], SAIDJON KAMOLOV[2]

Abstract. Imbalanced data affects a wide range of applications in machine learning. As a results, it has garnered a substantial amount of interest from the research community. Given the large volume of research related to imbalanced data, there is a need to organize and analyze the literature to facilitate the progress of the field. In this paper, we provide a state-of-the-art survey of approaches to dealing with imbalanced data. We discuss the existing methods along with their advantages and disadvantages. The presented survey will help researchers to better orient in the vast landscape of imbalanced data research.

1. Introduction

Imbalanced data refers to skewed distribution of labels in data. Concretely, it is the case when one class of labels significantly outnumbers another class of labels. The issue of imbalanced distribution occurs in many applications including medical diagnostics [15, 20], network intrusion [13, 16], finance [10, 14], fraud detection [12], manufacturing [4], and others [1, 11]. The skewed distribution of classes leads to classification bias and other issues in machine learning models [17, 18, 33]. Since the objective of most classification algorithms is to maximize the overall accuracy, they focus on the majority class. In extreme cases, when the majority class significantly dominates the minority class, classification algorithms would label all the samples as the majority. While this strategy may yield high accuracy results, it leads to poor performance on the minority class. On the other hand, the minority class samples are often of greater importance than the majority class. For example, in fraud detection, the instances of fraudulent transactions are much more rare than regular transactions. However, it is far more important to correctly identify the instance of fraud than regular transactions. Similarly, in medical data, the number of negative cases is much higher than positive cases. On the other hand, correctly identifying the positive cases can be life changing.

The main objective of a classification algorithm is to minimize its loss function. The loss function is traditionally defined based on categorical entropy or another similar metric. It is calculated based on all the sample points in the dataset. To minimize the loss function, the classification algorithm learns to separate different

Date: Received: Jan 10, 2022; Accepted: May 5, 2022.
* Corresponding author.
2010 *Mathematics Subject Classification.* Primary 46L55; Secondary 44B20.
Key words and phrases. imbalanced data; sampling; cost-sensitive; survey.

classes within different regions of feature space. Regularization is often employed to avoid overfitting. As a result, the contours of classification boundaries are smooth and gradual. To train a classifier efficiently in such setting requires a balanced dataset. In the absence of a balanced data the values of the loss function become dominated by the majority class. Since the loss function is calculated based on all the points in the data, the payoff from correctly classifying the majority class will be greater than the minority class. As a result, unless the classes are clearly removed from one another, the separation regions will be dominated by the majority class labels. The classification bias is directly correlated with the imbalance ratio. As the ratio of majority to minority classes increases, so does the classification bias. Beyond a certain threshold, the classification algorithm labels all the points as majority class. While classifying all the samples as the majority class may minimize the loss function, it has adverse effects on classifying the minority points. Since classifying the minority points is often of greater importance than the majority class, it is necessary create solutions to deal with imbalanced data.

Imbalanced data is an issue that affects a number of applications. Consequently, it has attracted a considerable amount of attention from the research community. There exist three main approaches to dealing with imbalanced data: sampling, cost-sensitive classification, and 1-class classification. Sampling is a preprocessing step, where the number of samples in each class is artificially equalized. In cost-sensitive classification, the cost of misclassifying minority points is made higher than the majority class. In 1-class classification, each class boundary is learned individually. While cost-sensitive classification and 1-class classification have been used in the literature, sampling is by far the most popular approach to dealing with imbalanced data. Given the large body of literature devoted to imbalanced data, a systematic review and analysis of the current research is required.

Our goal in this paper is to review the existing approaches to dealing with imbalanced data. We group the literature into three categories based on sampling, cost-sensitive classification, 1-class classification approaches, and ensemble methods. We identify the latest trends in research. In addition, we highlight the advantages and disadvantages of each approach. Our analysis yields several avenues for improvement in the current research.

2. Sampling

Sampling is the most popular approach to dealing with imbalanced data. It involves equalizing the number of samples in each class by either increasing the number of minority samples or decreasing the number of majority samples. The latter approach is called undersampling, while the former is referred to as oversampling. Hybrid approaches combine undersampling and oversampling in the attempt to obtain better performance.

In oversampling, new minority points are generated to balance the data. There exist a plethora of oversampling techniques. The simplest approach is random oversampling (ROS). In ROS, the new minority points are randomly chosen from

the existing points with replacement. As a result, there are instances of multiple points in the same location. It is a simple yet effective approach. The disadvantage of ROS is overfitting. Since there are multiple points in a single location, classifier may overfit that location. Several algorithms are based on generating new minority points in the neighborhoods of the original minority points. In [3], the authors propose the SMOTE algorithm which is based on generating new points along the straight line connecting neighboring points. Concretely, suppose x_0 is a minority point. Then another neighboring minority point x_1 is chosen and a straight line between x_0 and x_1 is created. Finally, using the uniform probability distribution, a random point is chosen along the line segment. SMOTE is a popular sampling method that is widely used in different applications. The SMOTE algorithm has inspired a number of extensions and variations including polynom-fit-SMOTE [5], ProWSyn [2], and SMOTE-IPF [27]. The authors in [8], employ a similar approach to SMOTE to generate the new minority points along the straight line connecting neighboring minority points albeit using the Gamma distribution. The authors show that the use of Gamma function yields a more natural distribution of the points. Classifiers with built in sampling mechanisms have also been proposed by researchers. In [19], the authors propose an ensemble model where each base classifier is trained on independent artificially balanced subsets of the data.

Several methods have been proposed that utilize a targeted approach to point generation. In particular, the new points are created in the regions with the greatest classification ambiguity. The goal is to address the regions with the greatest likelihood of misclassification. Traditionally, the most problematic regions for class separation occur along the border between the different classes. Therefore, sampling algorithms generate the greatest number of new points in these regions. In Borderline SMOTE [6], the new points are created in near the border between the minority and majority classes. In ADASYN [7], a similar intelligent approach is utilized. The authors in [**?**], proposed a method that considers simultaneously the interclass and intracluster distributions using a distributed fuzzy-based adaptive oversampling method.

Statistical sampling approaches are based on estimating the underlying probability distribution of the minority points. The estimated distribution function is then used to generate the new minority points. In parametric statistical modeling a known distribution is fit to the data. There are a number of parametric distributions that can be employed to fit the data including the Gaussian, Gamma, χ^2, multimodal and other distributions. The choice of the appropriate distribution depends on the individual dataset. Different heuristics such as the histogram and quantile metrics can be consulted in this regard. The parameters of the distribution are usually estimated to maximize the likelihood of the data. One popular approach to fit a distribution to the data is via the robust covariance algorithm [26]. It fits a robust covariance estimate to the data, and thus fits an ellipse to the central data points, ignoring points outside the central mode.

In nonparametric estimation, the distribution of the points does not follow a fixed function. Histogram and kernel-based approaches can be employed to model the distribution of the points. In [9], the authors employed kernel density

estimation (KDE) to model the underlying distribution of minority points. In particular, the distribution is modeled as the sum of Gaussian distributions each centered at individual minority points. By controlling the standard deviation of the kernel function, the fit of the estimate can be affected. In [30], the authors proposed a solution for regression on imbalanced data based on KDE.

Recently ensemble methods have gained popularity. Following the trend of ensemble classification models, a number of ensemble sampling methods have been proposed. Ensemble sampling methods combine multiple tools to generate the new points. In [41], the authors proposed a novel approach based on the mix of SMOTE and Tomek links to address the issue of imbalanced data in dimensional imbalanced settings.

The latest sampling methods have been based on generative models. Generative models have the capacity to generate new samples based on the trained model. In particular, variational autoencoders (VAE) have been employed to construct sampling algorithms [38, 35]. A VAE consists of encode and decoder. After training the VAE, the decoder can be used to generate new minority samples. Given the success of generative models in image and sound applications, it has a potential to be an effective sampling method. Similarly, generative adversarial networks (GAN) have been put forth by several researchers [28, 40].

Several studies have a tried to compare the performances of different sampling methods. The in [4], considered different sampling approaches in the context of manufacturing. They concluded that there is no single-best approach and that the optimal approach depends on individual datasets. Similarly, the authors in [36] compared different sampling methods in the context of credit card fraud. The results showed that the performance of the sampling algorithms varies according to the dataset and classifier. Another comparative study, considered the performance of KDE-based sampling approach to other techniques showing its robust performance [25].

While the newly proposed algorithm have achieved improved accuracy, it comes at the cost of increased computational load. The modern sampling algorithm require multistage calculations. The design of the algorithms is often highly involved leading to issues with implementations for third party users. In addition, the increased number of calculations adds to the computational load. Thus, the benefits of improved accuracy of new sampling algorithms must be carefully weighted against the cost of increased complexity.

2.1. **Undersampling.** Undersampling methods reduce the number of the majority points to the size of the minority class. The simplest approach is random undersampling (RUS). In RUS, the points are randomly selected from the existing majority set without replacement. It is a simple yet efficient sampling approach. More sophisticated undersampling methods have also been proposed. In particular, targeted undersampling aims to select the majority points near the border between different classes. The advantage of undersampling is in that it utilizes actual data points. While oversampling may introduce various biases in the data, undersampling uses only the true points. On the other hand, the size of the undersampled data is significantly smaller than the original imbalanced set. In case

when the ratio of the majority to minority points is very high, undersampling may lead to extremely small datasets. As a results, there may not be enough points to train the classifier effectively.

Several cluster-based undersampling methods have been utilized in the literature. In particular, the majority class is grouped into a number of clusters according to a specified strategy after which the majority points are selected from each cluster [22, 37] Recently, the authors in [21] proposed to modify radial-based oversampling to undersampling. The authors showed that the proposed method significantly reduced the time complexity of the method.

3. Cost-sensitive classification

Cost-sensitive classification is a relatively popular approach to dealing with imbalanced data. In cost-sensitive classification, the loss function of the classifier is modified to increase the penalty for misclassifying the minority points. In a standard loss function, the penalty for misclassifying a minority point is the same as the majority point. Thus, by increasing the penalty for misclassifying the minority points the classifier is encouraged to classify correctly the minority points.

There are two types of cost-sensitive approaches: static and dynamic. In the static approach, the penalty for misclassification of the minority points is fixed during the entire training process. The penalty is determined by the user. It often depends on the imbalance ratio. A simple approach is to equal the penalty ratio to the class distribution ratio. For instance, if the number of the majority points is twice as many as the minority points, then the penalty for misclassifying the minority points is twice greater than majority points. As a result, the classifier would emphasize learning the minority points twice as much. In the end, the greater penalty for misclassification will offset the fewer number of samples.

In the dynamic cost-sensitive approach, the loss function is continuous updated during the training process. The classification performance is evaluated after each training epoch and the loss function is adjusted as necessary to maintain balanced accuracy. If the classifier is not identifying the minority points as required, the loss function is adjusted to increase the penalty for misclassifying the minority points. Conversely, if the classifier is overperforming on the minority set the loss function is adjusted to decrease the penalty.

The advantage of cost-sensitive learning is simplicity and fidelity. The implementation of fixed cost-sensitive learning requires just one line of code. Even dynamic cost-sensitive learning is very simple to implement. In addition, cost-sensitive algorithms do not require additional preprocessing. Cost-sensitive approach is implemented directly into the learning model providing a simple and efficient solution to the problem of imbalanced data.

In one of the earliest attempts to apply cost-sensitive learning to imbalanced data [31], the authors modified the loss function of the AdaBoost algorithm to construct three different cost-sensitive boosting strategies. The algorithm was applied to medical data to improve classification performance. In a more recent attempt [32] the authors proposed a cost-sensitive ensemble for imbalanced data

classification based on Support Vector Machine (SVM). In the proposed method, the authors not only apply cost-sensitive SVMs as basic weak leaner but also concurrently modify the standard boosting scheme to cost-sensitive ones in order to to guarantee consistent optimization objectives between weak learners and the boosting scheme. The authors present a self-adaptive sequential misclassification cost weights determination method to ensure more training minority instances for successive classifiers, especially borderline minority instances. Recently, researchers proposed an adapted ResNet classifier with a cost-sensitive framework for PCB cosmetic defect detection [39]. Experiments show that the cost-sensitive framework improves the performance of the ResNet classifier.

4. One-class learning

Traditional classifiers consider all the available classes during the model training. As a result, in case of imbalanced data, the majority class can overwhelm the minority class. Therefore, traditional classifiers often exhibit biased tendencies in imbalanced data. To avoid the influence of the majority class, one-class learning can be employed. In one-class learning, only the data from a single class is considered. In particular, only the minority class labels are used in training. As a result, the majority class does not influence the performance of the classifier.

A popular one-class learning algorithm based on the Support Vector Machine SVM classifier was introduced in [29]. In the proposed algorithm, the SVM classifier creates a maximally tight fit to the data. The algorithm requires the choice of a kernel and a scalar parameter to define a frontier. The radial basis function (RBF) kernel is usually chosen although there exists no exact formula or algorithm to set its bandwidth parameter. An SVM-based one-class learning model was explored in [23]. A rule induction algorithm named Rule Extraction for MEdical Diagnosis (REMED), as a symbolic one-class learning approach was proposed in [24]. Experiment on medical data showed improved performance. More recently, a large scale study of one-class methods was studied in [34]. Using 55 datasets with different ranges of class imbalance ratios and one-class support vector machine, isolation forest, and local outlier factor as the representative one-class classification (OCC) classifiers, the authors found that the OCC classifiers perform well on high imbalance ratio data, doing better thatn the C4.5 baseline.

5. Conclusion

Given the rich literature related to imbalanced data, there is a need to analyze and summarize the research. In this paper, we presented a concise review of the current research in the field of imbalanced data.

References

[1] Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data, 8(1), 1-41.

[2] Barua, S., Islam, M. M., & Murase, K. (2013, April). ProWSyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 317-328). Springer, Berlin, Heidelberg.

[3] 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.
[4] Fathy, Y., Jaber, M., & Brintrup, A. (2020). Learning with imbalanced data in smart manufacturing: A comparative analysis. IEEE Access, 9, 2734-2757.
[5] Gazzah, S., & Amara, N. E. B. (2008, September). New oversampling approaches based on polynomial fitting for imbalanced data sets. In 2008 the eighth iapr international workshop on document analysis systems (pp. 677-684). IEEE.
[6] Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.
[7] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.
[8] Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368.
[9] Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. Information Sciences, 512, 1192-1201.
[10] Kamalov, F. (2020). Forecasting significant stock price changes using neural networks. Neural Computing and Applications, 32(23), 17655-17667.
[11] Kamalov, F., Cherukuri, A., Sulieman, H., Thabtah, F., & Hossain, A. (2021). Machine learning applications for COVID-19: a state-of-the-art review. arXiv preprint arXiv:2101.07824.
[12] Kamalov, F., Sulieman, H., & Santandreu Calonge, D. (2021). Machine learning based approach to exam cheating detection. Plos one, 16(8), e0254340.
[13] Kamalov, F., Moussa, S., Zgheib, R., & Mashaal, O. (2020, December). Feature selection for intrusion detection systems. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 265-269). IEEE.
[14] Kamalov, F., Smail, L., & Gurrib, I. (2020, November). Stock price forecast with deep learning. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 1098-1102). IEEE.
[15] Rajab, K., Kamalov, F., & Cherukuri, A. K. (2022). Forecasting COVID-19: Vector autoregression-based model. Arabian journal for science and engineering, 1-10.
[16] Kamalov, F., Zgheib, R., Leung, H. H., Al-Gindy, A., & Moussa, S. (2021, October). Autoencoder-based Intrusion Detection System. In 2021 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-5). IEEE.
[17] Kamalov, F., & Leung, H. H. (2020, November). Deep learning regularization in imbalanced data. In 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI) (pp. 1-5). IEEE.
[18] Kamalov, F., Thabtah, F., & Leung, H. H. (2022). Feature Selection in Imbalanced Data. Annals of Data Science, 1-15.
[19] Kamalov, F., Elnagar, A., & Leung, H. H. (2021, August). Ensemble Learning with Resampling for Imbalanced Data. In International Conference on Intelligent Computing (pp. 564-578). Springer, Cham.
[20] Kamalov, F., & Thabtah, F. (2021). Forecasting Covid-19: SARMA-ARCH approach. Health and Technology, 11(5), 1139-1148.
[21] Koziarski, M. (2020). Radial-based undersampling for imbalanced data classification. Pattern Recognition, 102, 107262.
[22] Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. Information Sciences, 409, 17-26.

[23] Maldonado, S., & Montecinos, C. (2014). Robust classification of imbalanced data using one-class and two-class SVM-based multiclassifiers. Intelligent Data Analysis, 18(1), 95-112.
[24] Mena, L., & Gonzalez, J. A. (2009). Symbolic one-class learning from imbalanced datasets: application in medical diagnosis. International Journal on Artificial Intelligence Tools, 18(02), 273-309.
[25] Plesovskaya, E., & Ivanov, S. (2021). An Empirical Analysis of KDE-based Generative Models on Small Datasets. Procedia Computer Science, 193, 442-452.
[26] Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), 212-223.
[27] 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.
[28] Salazar, A., Vergara, L., & Safont, G. (2021). Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets. Expert Systems with Applications, 163, 113819.
[29] 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(7), 1443-1471.
[30] Steininger, M., Kobs, K., Davidson, P., Krause, A., & Hotho, A. (2021). Density-based weighting for imbalanced regression. Machine Learning, 110(8), 2187-2211.
[31] Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. Pattern recognition, 40(12), 3358-3378.
[32] Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J. (2019). Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. Information Sciences, 487, 31-56.
[33] Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441.
[34] Tsai, C. F., & Lin, W. C. (2021). Feature Selection and Ensemble Learning Techniques in One-Class Classifiers: An Empirical Study of Two-Class Imbalanced Datasets. IEEE Access, 9, 13717-13726.
[35] Wan, Z., Zhang, Y., & He, H. (2017, November). Variational autoencoder based synthetic data generation for imbalanced learning. In 2017 IEEE symposium series on computational intelligence (SSCI) (pp. 1-7). IEEE.
[36] Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. Information Sciences, 569, 508-526.
[37] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.
[38] Zhang, C., Zhou, Y., Chen, Y., Deng, Y., Wang, X., Dong, L., & Wei, H. (2018, June). Over-sampling algorithm based on vae in imbalanced classification. In International Conference on Cloud Computing (pp. 334-344). Springer, Cham.
[39] Zhang, H., Jiang, L., & Li, C. (2021). Cs-resnet: cost-sensitive residual convolutional neural network for PCB cosmetic defect detection. Expert Systems with Applications, 185, 115673.
[40] Zheng, M., Li, T., Zhu, R., Tang, Y., Tang, M., Lin, L., & Ma, Z. (2020). Conditional Wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification. Information Sciences, 512, 1009-1023.
[41] Zhou, P., Hu, X., Li, P., & Wu, X. (2017). Online feature selection for high-dimensional class-imbalanced data. Knowledge-Based Systems, 136, 187-199.

[1]Bank Card Department, Spitamen Bank, Dushanbe, Tajikistan.
Email address: d.iskhakov@spitamen.com

[2]Faculty of Engineering, Tajik Technical University, Dushanbe, Tajikistan.
Email address: said.kamolov@ttu.tj