ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/365610619

Benchmarking Data Augmentation Techniques for Tabular Data

Chapter *in* Lecture Notes in Computer Science · November 2022

DOI: 10.1007/978-3-031-21753-1_11

CITATIONS
14

READS
299

3 authors:

Pedro Machado
University of Minho
1 PUBLICATION **14** CITATIONS
SEE PROFILE

Bruno Fernandes
University of Minho
32 PUBLICATIONS **264** CITATIONS
SEE PROFILE

Paulo Jorge Novais
University of Minho
680 PUBLICATIONS **6,057** CITATIONS
SEE PROFILE

All content following this page was uploaded by Pedro Machado on 13 February 2025.

The user has requested enhancement of the downloaded file.

Benchmarking Data Augmentation Techniques for Tabular Data

Pedro Machado[0000−0002−1697−1667], Bruno Fernandes[0000−0003−1561−2897], and Paulo Novais[0000−0002−3549−0754]

ALGORITMI Center, University of Minho, Braga, Portugal
pedrofcmachado26@gmail.com, bruno.fernandes@algoritmi.uminho.pt, pjon@di.uminho.pt

Abstract. Imbalanced learning and small-sized datasets are usual in machine learning problems, even with the increased data availability provided by recent developments. The performance of learning algorithms in the presence of unbalanced data and significant class distribution skews is known as the "imbalanced learning problem". The models' performance on such problems can drastically decrease for certain classes with an uneven distribution because the models do not learn the distributive features of the data and present accuracy too favorable for a specific set of classes of data. As an example, this can have negative consequences when talking about cancer detection since the model may poorly identify unhealthy patients. Hence, data augmentation techniques are usually conceived to evaluate how models would behave in non-data-scarce environments, generating synthetic data that mimics the characteristics of real data. By applying those techniques, the amount of available data can be increased, balancing the class distributions. However, there are no standardized data augmentation processes that can be applied to every domain of tabular data. Therefore, this study aims to identify which characteristics of a dataset provide a better performance when synthesizing samples by a data augmentation technique in a tabular data environment.

Keywords: data augmentation · imbalanced data · machine learning

1 Introduction

In recent years, the imbalanced learning problem has become a highly frequent topic among academia, industry, and government funding agencies. The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly decrease the performance of machine learning algorithms [4]. These algorithms, when faced with imbalanced data, do not learn the distributive features of the data and present accuracies too favorable to a specific set of classes of data, in this case, the majority classes, compromising the performance of the other classes (the minority classes) because of that bias. In fairness, a dataset is considered imbalanced when it exhibits an unequal distribution between its classes. Nevertheless, the community usually considers that imbalanced data corresponds to a large unequal distribution and, in some cases, extremes.

In an imbalanced data problem, the real problem can't be solved with data treatment and/or model changes since the limitation is in the data itself. The same goes for a small-sized dataset, since the model cannot learn enough features to classify the problem in a real-time situation. Therefore, data augmentation appears as a way to surpass that limitation [3].

The present article, by identifying more favorable properties in a dataset when synthesizing samples, aims to conceive and benchmark several candidate models to overcome the imbalanced learning problem as well as increase the amount of available data without a loss in quality. With this in mind, we used classical techniques, such as SMOTE, a very uncommon clustering technique like Gaussian Mixture Model (GMM), and deep learning ones, such as Variational Autoencoder (VAE). Finally, this manuscript is structured as follows: the next section describes the literature review on the addressed domains; the third section presents the conducted experiments as well as the achieved results for this benchmark; the last section summarizes the obtained conclusions and outlines future work.

2 State of Art

Data Augmentation refers to methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables [2]. With these techniques, we can increase the amount of data, thus balancing the target variable in an imbalanced dataset. Data augmentation can be applied to images or tabular data, but this article will focus on the latter. When used alongside images, techniques tend to apply transformations to samples of datasets like geometric transformations, flipping, color modification, cropping, etc. One other way is to introduce new synthetic images created by machine learning algorithms, for example, VAEs. Instead, if we are dealing with tabular data, we cannot apply simple transformations to samples, but instead synthesize samples (new or duplicated) based on the class distributions and features.

In [11], Data Augmentation is utilized to surpass the data limitations of the minority class, in this case, fraudulent transactions. The classification performance improved considerably and overfitting was alleviated, demonstrating the benefits of using a these techniques. These techniques can also be applied to automated skin lesion analysis by applying traditional color and geometric transformations, and more unusual augmentations such as elastic transformations, random erasing, and a novel augmentation that mixes different lesions, as stated in [8]. They prove the importance of data augmentation techniques in both training and testing, leading to more performance gains than simply obtaining new images.

In this study, we focused on some of the most popular data augmentation techniques, namely, SMOTE, GMM, and VAE. These techniques are present in multiple data augmentation studies and are going to be developed and evaluated in order to augment the used datasets.

SMOTE. One of the classic data augmentation techniques is SMOTE. It oversamples the minority class by creating synthetic samples [1]. One way to solve the imbalance problem is to duplicate minority samples. However, this does not provide any new information to the machine learning algorithm training on the data. Therefore, instead of duplicating minority samples, SMOTE synthesizes new examples from that class.

This technique synthesizes the minority class by operating in the feature space. It selects examples that are close in the feature space and introduces synthetic samples along the line drawn from these examples. SMOTE is effective because the new synthetic samples from the minority class are somewhat close in feature space to real samples from that same class. This makes the created samples plausible.

Gaussian Mixture Model The GMM's generative nature provides an opportunity to explore its performance as a data augmentation technique, contrarily to other clustering algorithms, such as K-means. The use of a simple radial distance metric by k-means to assign cluster membership results in poor performance and a typical circular form for the clusters. This algorithm has no built-in way of accounting for non-circular clusters (oblong or elliptical), which do not represent the true shape of the data points sometimes. Moreover, this algorithm does not have a probabilistic nature when forming clusters.

Therefore, GMMs are an extension of the ideas behind k-means. This algorithm aims to model the data as a combination of multiple multi-dimensional Gaussian probability distributions and it works on the basis of the Expectation-Maximization algorithm. Because of this, the EM algorithm finds the maximum likelihood, i.e., finds a set of parameters that results in the best fit for the joint probability of the data sample [7]. Due to the generative nature of GMM, it can generate synthetic data close to the distribution of the fitted data [10]. After the algorithm fits the data and learns its distribution, it can generate an arbitrary number of samples from the learned distribution.

Variational Autoencoder. Nowadays, deep learning has gained a lot of interest and has made some amazing improvements regarding its performance. From the deep learning models, the family of generative models has also increased in popularity, showing a magnificent ability to produce highly realistic samples of various kinds, such as images, text, and sounds. These families of models, like all deep learning models, rely on huge amounts of data, well-structured architectures, and smart training techniques. One of these popular deep learning generative models is the Variational Autoencoder. In short, a VAE is an autoencoder whose encoding distribution is regularized during the training in order to ensure that its latent space[1] has good properties, allowing us to generate some new data [9].

[1] Latent space is a representation of compressed data in which similar data points are closer together in space. It is useful to learn the features.

A VAE consists of an encoder and a decoder, just like an autoencoder, but the loss term and the encoded layers of the autoencoder are altered in order for the model to be used as a generative model [5]. Its training is adjusted to avoid overfitting, making sure that the latent space has good properties that enable the generative process. On the other hand, an autoencoder is trained to encode and decode with as few losses as possible, making no difference how the latent space is organized. The main distinction between the two encoding layer algorithms is that they encode an input as a distribution throughout the latent space rather than a single point [9]. With this in mind, the VAE avoids having some points in the latent space that would provide meaningless information once decoded.

3 Experiments

The data generated was used in two different ways in order to evaluate the data augmentation techniques. First, it was added to the original training data that trained the classifiers and then evaluated based on the test data. The second approach is to train the classifiers only with synthetic data and evaluate them with the test data. Note that all the test data is real.

3.1 Data

In this experiment, multiple datasets were chosen to perform a good comparison of these data augmentation techniques in generating new data. Moreover, these datasets are inserted into different domains, such as health and fraud detection, and are imbalanced. Furthermore, these datasets were also chosen due to being mainly composed by continuous or categorical features. The chosen datasets were the following:

- **Adult**. The adult dataset was extracted from the census bureau and has information about multiple adults. This dataset serves as a binary classification, predicting if a certain adult has an income superior to fifty thousand in a year. The target class is clearly imbalanced, as the majority class (income superior to fifty thousand) is three times more frequent than the minority class. The dataset has over $30K$ instances.
- **Breast Cancer**. Another health domain analyzed in this experiment was breast cancer prediction. This dataset was obtained from the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg and contains samples of clinical cases gathered periodically. The dataset contains a target class imbalanced with 66% of the instances belonging to benign cases and the rest being malignant. This dataset, beyond being imbalanced, is also small in size, since it only has 570 instances.
- **Credit Card Fraud**. Fraud detection is also a recurrent domain where imbalanced data is present. Therefore, in this dataset, transactions made by credit cards in September 2013 by European cardholders are analyzed. This dataset originally had more than $280K$ instances but was reduced (while maintaining the target class ratio) to $85K$ due to computational reasons.

3.2 Assessment Metrics

In order to compare the performances of all the data augmentation techniques, it is required to define how their performances can be compared. Therefore, we need to define which metrics fit better into an imbalanced data problem. Traditionally, the most often used metrics are *accuracy* and *error rate*. Although accuracy provides an easy way to describe the model's performance, it can mislead in certain situations. Therefore, accuracy and error rate do not provide enough information about a classifier's functionality in terms of the sort of classification required.

As a means to provide comprehensive assessments of imbalanced learning problems, the research community adopted other evaluation metrics, such as *precision*, *recall*, *F-measure*[2], and *G-mean*.

First, precision is a metric that measures how many correct positive predictions the model makes (a measure of exactness)[3]. Therefore, precision calculates the accuracy of the positive class and is sensitive to data distribution. Second, recall is a metric that measures how many correct positive predictions were produced out of all possible positive predictions. Unlike precision, which only gives information on the correct positive predictions of all positive predictions, recall indicates the missed positive predictions and it is not sensitive to data distributions. Moreover, recall is also known as sensitivity. When used correctly, recall and precision can evaluate an imbalanced learning problem adequately. Nevertheless, the F-measure metric combines the two previous metrics as a weighted focus on either recall or precision. Finally, the G-mean (Geometric mean) metric evaluates the balance of classification between the majority and minority classes. Even if the negative cases are accurately identified, a low G-Mean suggests poor performance in the classification of positive cases.

3.3 Experimental Results

Regarding the synthetic data generated by all data augmentation techniques, the experiments have produced a variety of findings. These analyses consist, mainly, of:

1. Comparing each feature's distribution throughout statistical methods;
2. Training the classifiers only with synthetic data;
3. Training the classifiers with real and synthetic data.

In all cases, all the techniques implemented generated the same amount of synthetic data. In this case, this amount is the size of the training dataset (i.e., seventy percent of the entire dataset). As a result, all the synthetic data generated can be compared to one another and to the original data.

In order to compare the synthetic data of each data augmentation technique, we first compared how each feature of the real and synthetic datasets behaves,

[2] It is also known as *F-score*.
[3] In this case, the positive class is considered the minority.

i.e., if they possess the same distribution. Ideally, a synthetic dataset should have properties very similar to the original one. Therefore, we implemented some statistical methods to compare each feature distribution on the real and synthetic datasets. However, due to the very different behavior of continuous and categorical features, it was necessary to apply different statistical tests. The categorical feature distributions were analyzed by the chi-square test and the continuous features by the Kolmogorov-Smirnov test.

In the adult dataset, most of the features are categorical, with only one continuous feature. The statistical tests showed that the data augmentation techniques had almost no difficulty representing the original categorical features in the synthetic data. However, the techniques couldn't represent the continuous feature distributions since all of the techniques failed the test. In regards to the Breast Cancer dataset, all features are continuous since they are medical measures. SMOTE showed as the best technique to represent the feature distributions as the other techniques couldn't. Finally, the credit card fraud dataset had similar properties to the previous dataset, containing only continuous features. However, all the data augmentation techniques had difficulties representing similar continuous feature distributions.

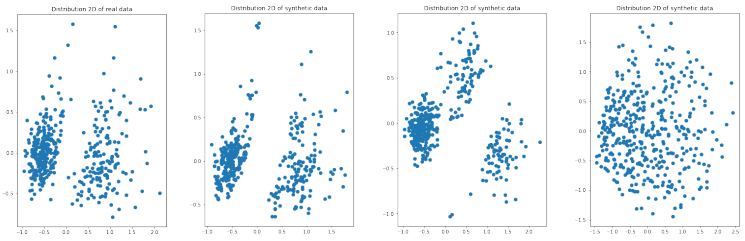
The results of the data augmentation techniques throughout all datasets indicated the increased difficulty in representing continuous feature distributions, with SMOTE being the technique with the best representations. However, the categorical feature distributions were much easier to represent, with GMM and VAE being the ones with better results.

Moreover, one important factor noticed during the training of more complex and computationally resource-demanding techniques, such as VAE, was the fact that continuous features should be normalized in order to reduce computational cost and avoid crashes during training. These kinds of crashes are detectable when the loss is NaN during training.

In regards to the data distribution of one of the datasets, as we can see in Figure 1, most data augmentation techniques can represent the entirety of the data distribution. Also, VAE seems to be the technique with the most difficulty in separating what seems to be the two target classes, but it doesn't appear to affect the classifier's performance as we will observe.

We can still perform two additional crucial analyses to determine how reliable the generated data is after the analysis of the synthetic data properties. First, we are going to train the machine learning classifiers with real data and then compare the results with training with only synthetic data. Note that there were multiple classification models, but we only represented the best model for each case.

During the experiments and implementation of the data augmentation techniques, there were some obstacles with regard to the performance of some techniques, mainly the Variational Autoencoder. The implemented VAE suffered from the phenomenon called *posterior collapse* [6]. As a result, the minority class was unable to be synthesized, and the solution was to add a weight decay on the loss function as well as change the latent space dimension.



(a) Real Data (b) SMOTE (c) GMM (d) VAE

Fig. 1: Comparison of two dimension data throughout all data augmentation techniques on the Breast Cancer Dataset.

As observed in Table 1, synthetic data can achieve similar training scores in comparison with training with real data. SMOTE and VAE demonstrated better performance in generating samples on the three datasets. In this experiment, the VAE showed better performance for datasets with fewer categorical features (in this experiment, the Breast Cancer and Credit Card fraud datasets), while GMM indicated difficulties in generating good synthetic data. These experiments demonstrated that data augmentation techniques such as SMOTE or VAE can synthesize data in order to replace the real data in an efficient way. This could be very interesting in datasets where some data is sensitive and privacy matters.

Table 1: Best classifier performance on different kinds of training data.

Dataset	DA Technique	Minority Class			Majority Class			G-Mean
		Precision	Recall	F1 Score	Precision	Recall	F1 Score	
Adult	—	0.6957	0.6327	0.6626	0.8868	0.9123	0.8993	0.7597
	SMOTE	0.6088	0.6173	**0.6130**	0.8781	0.8742	**0.8762**	**0.7643**
	GMM	0.2352	0.6033	0.3385	0.7504	0.3780	0.5028	0.4776
	VAE	0.4565	0.2742	0.3427	0.7958	0.8965	0.8431	0.4958
Breast Cancer	—	1.0	0.9524	0.9756	0.9737	1.0	0.9867	0.9759
	SMOTE	1.0	0.8095	0.8947	0.9024	1.0	0.9487	0.8997
	GMM	0.6111	0.5238	0.5641	0.7500	0.8108	0.7792	0.6517
	VAE	0.9090	0.9524	**0.9302**	0.9722	0.9459	**0.9589**	**0.9492**
Credit Card Fraud	—	1.0	0.8667	0.9286	0.9998	1.0	0.9999	0.9309
	SMOTE	0.9286	0.8667	0.8966	0.9998	0.9998	0.9998	0.9309
	GMM	0.0027	0.8667	0.0054	0.9995	0.4411	0.6121	0.6005
	VAE	0.9286	0.8667	**0.8967**	0.9998	0.9999	**0.9999**	**0.9309**

With those results in mind, the following analyses focus on training the classifiers with an increased amount of data (in this case, twice the original data

size). At Table 2, we get to see that the data augmentation techniques that achieved better results in Table 1 got the best results with the addition of real data into the classifier's training. We can also observe that these techniques increase or maintain the classifier's performance in both major and minor classes. One example of that is the dataset Breast Cancer, where the application of a Variational Autoencoder made the classifier's performance go up in both classes' f1-score and g-mean metrics. This implies that the data augmentation can increase not only the minority class's performance but all classes' performances as well. Another interesting finding was the performance gained by the GMM technique when joining its synthetic and real data. This may be explained by the variety of generated samples that, in this case, benefited the classifier training.

Table 2: Best classifier performance while training with real and synthetic data.

Dataset	DA Technique	Minority Class			Majority Class			G-Mean
		Precision	Recall	F1 Score	Precision	Recall	F1 Score	
Adult	SMOTE	0.6935	0.6263	0.6583	0.8850	0.9123	0.8884	0.7559
	GMM	0.7149	0.6301	**0.6698**	0.8870	0.9203	**0.9034**	**0.7615**
	VAE	0.7032	0.6199	0.6589	0.8839	0.9171	0.9002	0.7540
Breast Cancer	SMOTE	1.0	0.9048	0.9500	0.9487	1.0	0.9737	0.9512
	GMM	1.0	0.9048	0.9500	0.9487	1.0	0.9734	0.9512
	VAE	0.9545	1.0	**0.9767**	1.0	0.9730	**0.9863**	**0.9864**
Credit Card Fraud	SMOTE	1.0	0.8000	0.8889	0.9996	1.0	0.9998	0.8944
	GMM	1.0	0.7333	0.8462	0.9996	1.0	0.9998	0.8563
	VAE	1.0	0.8667	**0.9286**	0.9998	1.0	**0.9999**	**0.9308**

4 Conclusion

In this study, we went through a benchmark of different data augmentation techniques in multiple datasets of various domains. We observed that classical techniques such as SMOTE are competitive with more recent and powerful techniques like VAE. Also, the introduction of a not so frequent technique like GMM gave a new look to cluster models as a possibility to generate samples. Even though the Variational Autoencoder is more complex and susceptible to training problems such as the *posterior collapse*, it is a very powerful technique.

Furthermore, VAE was shown to be a better solution for a dataset with more continuous features. On the contrary, SMOTE had a better performance in a dataset with more categorical features. One other important factor to take into account is the normalization of continuous features during the preprocessing of the data. This avoids higher losses that may stall the training process.

Regarding the obtained results, the data augmentation techniques showed a great capability to create almost identical datasets to the real ones and have very similar scores. Moreover, these techniques can combat the imbalanced data problem by increasing the performance of the minority class, and they can also increase the size of a dataset without a classifier's performance loss. Future work will focus on further benchmarking techniques and new analyses of the classifier's performances.

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, *FCT – Fundação para a Ciência e Tecnologia*, within the project DSAIPA/AI/0099/2019.

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
2. van Dyk, D.A., Meng, X.L.: The art of data augmentation. Journal of Computational and Graphical Statistics **10**, 1–50 (2001)
3. Fernandes, B., Silva, F., Alaiz-Moretón, H., Novais, P., Analide, C., Neves, J.: Traffic flow forecasting on data-scarce environments using arima and lstm networks. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) New Knowledge in Information Systems and Technologies. pp. 273–282. Springer International Publishing, Cham (2019)
4. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009). https://doi.org/10.1109/TKDE.2008.239
5. Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J.: Crash data augmentation using variational autoencoder. Accident Analysis & Prevention **151**, 105950 (2021). https://doi.org/https://doi.org/10.1016/j.aap.2020.105950
6. Lucas, J., Tucker, G., Grosse, R.B., Norouzi, M.: Understanding posterior collapse in generative latent variable models. In: DGS@ICLR (2019)
7. McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions, vol. 382. John Wiley & Sons (2007)
8. Perez, F., Vasconcelos, C., Avila, S., Valle, E.: Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. pp. 303–311. Springer International Publishing, Cham (2018)
9. Rocca, J.: Understanding variational autoencoders (vaes) (03 2021), `https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73`, last Visited July 11, 2022
10. Sarkar, T.: How to use a clustering technique for synthetic data generation (9 2019), `https://towardsdatascience.com/how-to-use-a-clustering-technique-for-synthetic-data-generation-7c84b6b678ea`, last Visited July 10, 2022
11. Shao, M., Gu, N., Zhang, X.: Credit card transactions data adversarial augmentation in the frequency domain. In: 2020 5th IEEE International Conference on Big Data Analytics (ICBDA). pp. 238–245 (2020). https://doi.org/10.1109/ICBDA49040.2020.9101344

View publication stats