Evaluation of Data Augmentation for Tabular Data

Fabiana Fazio, Kate Huisinga, Ace Mejía-Sánchez, and Ellee Mortensen

Advisor: Dr. Marilyn Vazquez Bryan Summer Research Program

Problem Statement

- The effective use of machine learning is challenging due to the substantial amount of data required for accuracy.
- To address data limitations, an existing solution is to turn to data augmentation techniques to artificially increase the size of data sets and improve model performance [1].
- Most research has focused on data augmentation methods alone; instead, we focused on evaluating if these methods were effective.

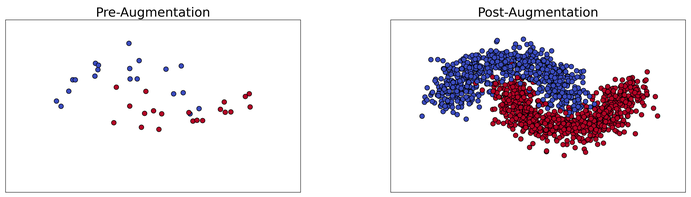
Research Question: Do current data augmentation methods effectively augment tabular data while maintaining original data patterns?

Our project goals are:

- Create evaluation metrics for augmenting tabular data
- Using evaluation metrics, compare: Feature Distributions, Features to Features and Features to Label

What is Data Augmentation?

The process of **generating new data** from existing data while **maintaining patterns** without collecting new data.

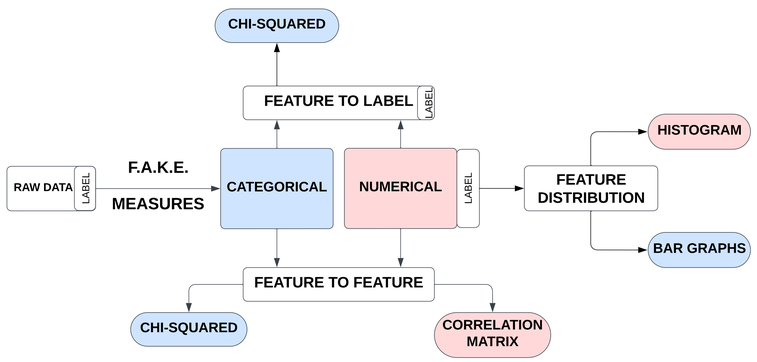


Augmentation Methods Used:

- **modPMOne:** chooses an observation (row) and adds or subtracts one unit from random variables [2]
- **randSwap:** randomly swaps a value from the augmented row with values of the same column from the original data set [2]
- **HAT:** generates a data set whose distribution is similar to that of the original data set [3]

Methods

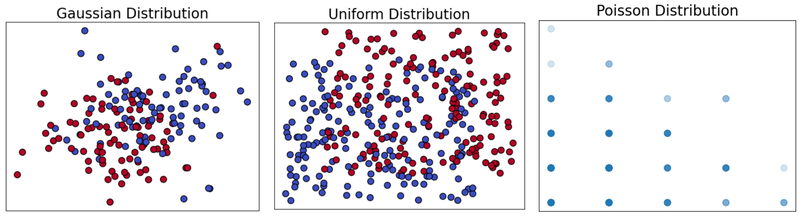
Our Process for Evaluating Patterns



Measuring Tools

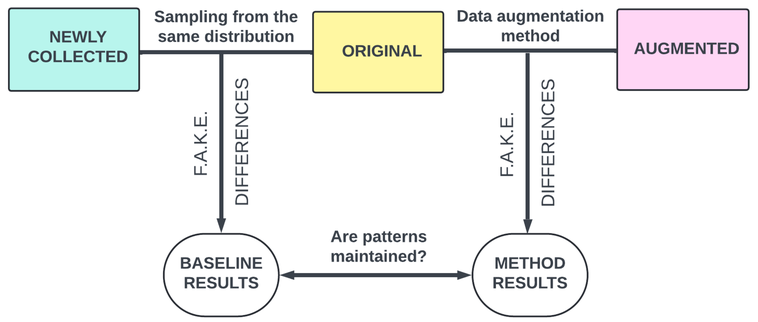
- **Correlation Matrix:** table that shows how strongly pairs of numerical columns are related
- **Frobenius Norm:** measure of the size or distance between matrices
- **Histogram:** displays the bin edges and the proportions of numerical columns in a data set
- **Mann-Whitney U Test:** measures differences between two numerical distributions [4]
- **Bar Graph:** displays the proportions of subcategories within categorical columns, illustrating the frequency or count of each subcategory
- **Euclidean Norm:** measure of the size or distance between vectors
- **Chi-Squared Test:** finds dependence between two categorical columns or between categorical columns and the label column

Data Sets



- Gaussian Data Set: 172 x 13
- Uniform Data Set: 343 x 25
- Mixed Distribution Data Set (Gaussian, uniform, and Poisson distributions horizontally mixed, respectively): 240 x 25

Experimental Design



Preliminary Results

Table 1. Relative Error of Correlation Matrices using Frobenius Norm (Feature-to-Feature)

	Gaussian Original	Uniform Original	Mixed Original
Original + New	**0.099**	**0.105**	**0.114**
Augmented with modPMOne	0.151	0.134	0.145
Augmented with randSwap	0.225	0.274	0.341
Augmented with HAT	0.082	0.124	0.116

Table 2. Count of Numerical Columns with Significant p-values using Mann-Whitney U Test

	Gaussian Original	Uniform Original	Mixed Original
Original + New	**0**	**0**	**0**
Augmented with modPMOne	0	0	0
Augmented with randSwap	0	0	0
Augmented with HAT	0	0	0

Table 3. Relative Error of Categorical Proportions using Euclidean Norm

	Gaussian Original	Uniform Original	Mixed Original
Original + New	**0.138**	**0.029**	**0.100**
Augmented with modPMOne	0.387	0.046	0.283
Augmented with randSwap	0.107	0.058	0.098
Augmented with HAT	0.106	0.060	0.102

Table 4. Count of Changes in Chi-Squared Significance (Feature-to-Label)

	Gaussian Original	Uniform Original	Mixed Original
Original + New	**0**	**0**	**0**
Augmented with modPMOne	0	5	9
Augmented with randSwap	0	0	2
Augmented with HAT	0	0	9

Table 5. Count of Changes in Chi-Squared Significance (Feature-to-Feature)

	Gaussian Original	Uniform Original	Mixed Original
Original + New	**0**	**0**	**2**
Augmented with modPMOne	8	6	28
Augmented with randSwap	0	6	8
Augmented with HAT	10	10	60

Conclusions

- Previously, no comprehensive methodology existed to test whether the patterns of a data set are maintained during data augmentation.
- We implemented our methodology through two functions, F.A.K.E. Measures and F.A.K.E. Differences, to ensure the original patterns are preserved.
- We designed experiments with synthetic data to test pattern retention after augmentation.

Future Work

- It would be beneficial for a variety of real-life data sets to be tested with our methods in order to determine their efficacy.
- A measuring tool could be added to our framework that compares numerical and categorical variables.
- Further work could focus more on the details of augmentation and how to optimize the process.

Acknowledgments

This work is part of the Bryan Summer Research Program, which is generously funded by Dr. Albert H. and Greta A. Bryan, as well as the SIAM-Simons Undergraduate Summer Research Program, which is funded by the Society for Industrial and Applied Mathematics (SIAM) through Simons Foundation award 1036702.

References

[1] Pedro Filipe Costa Machado. *Conception and evaluation of data augmentation techniques for tabular data*. PhD thesis, Universidade do Minho, 2022.

[2] Christina Dietrich, Jeffrey Roberts, Jason White, and Marilyn Vazquez. Data augmentation for tabular data sets. 2022.

[3] Balachander Sathianarayanan, Yogesh Chandra Singh Samant, Prahalad S Conjeepuram Guruprasad, Varshin B Hariharan, and Nirmala Devi Manickam. Feature-based augmentation and classification for tabular data. *CAAI Transactions on Intelligence Technology*, 7(3):481–491, 2022.

[4] K. Perry. Determine if two distributions are significantly different using the mann-whitney u test. 2019.

GitHub Link for Our Code

faziofa@kean.edu kate.huisinga@my.simpson.edu acemejiasanchez@college.harvard.edu ellee.mortensen@my.simpson.edu