SynGen: Synthetic Data Generator for Feature Selection

Firuz Kamalov (Canadian University Dubai, Dubai, UAE, firuz@cud.ac.ae)
Said Elnaffar (Canadian University Dubai, Dubai, UAE, said.elnaffar@cud.ac.ae)
Hana Sulieman (American University of Sharjah, Sharjah, UAE, hsulieman@aus.edu)
Aswani Cherukuri (Vellore Institute of Technology, Vellore, India, cherukuri@acm.org)

Abstract

Given the large number of existing and new feature selection algorithms, it has become imperative to have a uniform procedure for evaluating the performance of the algorithms. To this end, we propose a library of synthetic datasets designed specifically to test the effectiveness of feature selection algorithms. The datasets are inspired by applications in electronics and have a range of characteristics to provide a variety of test scenarios. The software comes in the form of a Python library with standard interface for loading and generating datasets. Each dataset is implemented as a function that allows to control various parameters of the data.

Keywords

feature selection, synthetic data, machine learning, data mining

Code metadata

Please replace the italicized text in the right column with the correct information about your code/software and leave the left column untouched.

Nr.	Code metadata description	Please fill in this column
C1	Current code version	*For example: v42*
C2	Permanent link to code/repository used for this code version	*For example: https : //github.com/mozart/mozart2*
C3	Permanent link to Reproducible Capsule	
C4	Legal Code License	*All software and code must be released under one of the pre-approved licenses listed in the Guide for Authors, such as Apache License, GNU General Public License (GPL) or MIT License. Write the name of the license you've chosen here.*
C5	Code versioning system used	*For example svn, git, mercurial, etc. (put 'none' if none used)*
C6	Software code languages, tools, and services used	*For example C++, python, r, MPI, OpenCL, etc.*
C7	Compilation requirements, operating environments & dependencies	
C8	If available Link to developer documentation/manual	*For example: http : //mozart.github.io/documentation/*
C9	Support email for questions	

1. Synthetic data generator for feature selection

Feature selection has been an active area of research with dozens of new algorithms being proposed every year. In this software package, we provide a Python library for generating synthetic datasets that are designed specifically to test the effectiveness of feature selection algorithms. The library consists of functions that allow to load and generate 5 different datasets. Each dataset consists of a number of relevant, redundant, correlated, and irrelevant variables. The target variable is calculated based on a predetermined rule/formula as described in [7]. The library functions allow to specify dataset parameters such as the number of irrelevant variables, the number of instances, and the random seed. Since the relevant variables are known a priori, synthetic data enables direct evaluation of feature selection algorithms. To mimic real life scenarios, the datasets are inspired by applications in electronics.

Feature selection has become an important part of the process in many data science and machine learning applications. As a result, a large number of feature selection algorithms have been proposed in the literature [1, 4, 6, 8, 12]. However, there does not exist a universal benchmark for evaluating these algorithms. To fill this gap, we propose a software package called SynGen that allows to generate synthetic data tailored explicitly to assess feature selection algorithms. We aim that the proposed software and the corresponding datasets would be used to evaluate the existing and future feature selection algorithms and provide a standard approach to measure and analyze the effectiveness of algorithms.

A summary of the datasets generated via SynGen is presented in Table 1. The table shows the default parameter values of the datasets. The datasets include different types of the target variable including binary, multi-class, and continuous values. While the number of relevant, redundant, and correlated features is fixed, the number of irrelevant features and the sample size can be specified through the corresponding data generating functions. In addition, the random seed can be specified to generate different irrelevant features for algorithm stability analysis. The details of the datasets used in the SynGen library can be found in [7].

Name	Relevant	Redundant	Correlated	Irrelevant	Samples	Target
ORAND	3	3	2	92	50	binary
ANDOR	4	4	2	90	50	binary
ADDER	3	3	2	92	50	4-class
LED-16	16	16	2	66	180	36-class
PRC	5	5	2	88	500	continuous

Table 1: Summary the SynGen datasets.

2. Impact and use cases

The majority of the existing synthetic datasets used to evaluate feature selection algorithms were originally designed for classification tasks [2, 3, 5, 11]. On the other hand, the SynGen data is designed specifically for use in feature selection. The SynGen data includes redundant as well as correlated features to provide a rich setting to test the algorithms. There are two primary advantages of using synthetic data over real life data: i) the knowledge of the relevant features, and ii) control of the data parameters. In the traditional approach using real life data, feature selection algorithms are evaluated based on the accuracy of the classifier trained on the selected features. On the other hand, the nature of all the variables in synthetic data is known so the selected features can be evaluated directly.

The ANDOR dataset generated via SynGen was used to compare the performance of several feature selection algorithms in [7]. The study showed that while most of the algorithms are able to distinguish between the relevant and irrelevant variables, they fail to separate the relevant variables from the redundant and correlated variables. Several SynGen generated datasets were used in [9] to evaluate the performance of a new feature selection algorithm called Nested Ensemble Selection.

Synthetic data enables an in-depth analysis of feature selection algorithms by controlling the parameters of the dataset. In particular, SynGen allows to specify the number of irrelevant features and the size of the dataset. By varying the number of irrelevant variables the corresponding performance of the selection algorithm can be observed and analyzed [10]. Similarly, the sensitivity of an algorithm to the size of the dataset can be investigated by varying the number of instances in SynGen.

As an illustration of the use of SynGen, consider the results of applying the χ^2 univariate feature selection algorithm on the ADDER dataset. The results are shown in Figure 1, where the top and bottom subfigures are based on sample size 20 and 50, respectively. It shows that increase in the sample size increases the difference between the relevant and irrelevant features. On the other hand, the algorithm fails to distinguish between the relevant and redundant variables. It also shows that the algorithm assigns high scores to the (randomly) correlated variables. Thus, we obtain a better understanding of the performance of the algorithm and its characteristics.

3. Conclusion and future development

In this report, we presented a Python package called SynGen which allows to generate synthetic data designed for feature selection. While SynGen is aimed at evaluating feature selection algorithms, it can also be used for classification and regression tasks. For example, researchers can analyze the performance of a classification model with different sample sizes or number of irrelevant variables.

As part of future development, we aim to expand the collection of the datasets in SynGen. In addition, we hope to establish a forum where researchers can share the results of feature selection algorithms based on SynGen datasets.

References

[1] Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through boosting. Expert Systems with Applications, 187, 115895.

[2] Belanche, L. A., & González, F. F. (2011). Review and evaluation of feature selection algorithms in synthetic problems. arXiv preprint arXiv:1101.2320.

[3] Bolon-Canedo, V., Sánchez-Marono, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3), 483-519.

[4] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis, 143, 106839.

[5] John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In Machine learning proceedings 1994 (pp. 121-129). Morgan Kaufmann.

[6] Kamalov, F. (2021). Orthogonal variance decomposition based feature selection. Expert Systems with Applications, 182, 115191.

[7] Kamalov, F., Sulieman, H., & Cherukuri, A. K. (2022). Synthetic Data for Feature Selection. arXiv preprint arXiv:2211.03035.

[8] Kamalov, F., Thabtah, F., & Leung, H. H. (2022). Feature Selection in Imbalanced Data. Annals of Data Science, 1-15.

[9] Kamalov, F., Reyes, J., Moussa, S., & Safaraliev, M. (2023). Nested Ensemble Selection: an effective hybrid feature selection method. Under review.

[10] Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34(4), 1060-1073.

[11] Kim, G., Kim, Y., Lim, H., & Kim, H. (2010). An MLP-based feature subset selection for HIV-1 protease cleavage site analysis. Artificial intelligence in medicine, 48(2-3), 83-89.

[12] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375. Chicago