cluster-over-sampling: A package for clustering based oversampling

Georgios Douzas, Fernando Bacao*

NOVA Information Management School, Universidade Nova de Lisboa

Abstract

Learning from class-imbalanced data is a common and challenging problem in supervised learning. Standard classification algorithms are designed to handle balanced class distributions. While different strategies exist to tackle this problem, methods that generate artificial data to achieve a balanced class distribution, called oversampling algorithms, are more versatile than modifications to the classification algorithms. SMOTE algorithm, the most popular oversampler, as well as any other oversampling method based on it, generates synthetic samples along line segments that join minority class instances. SMOTE addresses only the issue of between-classes imbalance. On the other hand, by clustering the input space and applying any oversampling algorithm for each resulting cluster with appropriate resampling ratio, the within-classes imbalanced issue can be addressed. This approach, implemented in the cluster-over-sampling Python open source project, has been shown in various publications to outperform other standard oversamplers in a large number of datasets. In this paper we describe cluster-over-sampling in detail and make it available to the machine learning community. An important point is that the implementation integrates effortlessly with the Scikit-Learn ecosystem. Therefore, machine learning researchers and practitioners can integrate it directly to any pre-existing work.

Keywords: Machine learning, Classification, Imbalanced learning, Oversampling

*Postal Address: NOVA Information Management School, Campus de Campolide, 1070-312 Lisboa, Portugal, Telephone: +351 21 382 8610

Email addresses: `gdouzas@novaims.unl.pt` (Georgios Douzas), `bacao@novaims.unl.pt` (Fernando Bacao)

Code metadata	
Current code version	v0.1.1
Permanent link to code/repository used for this code version	`https://github.com/AlgoWit/cluster-over-sampling`
Legal Code License	MIT
Code versioning system used	git
Software code languages, tools, and services used	Python, Travis CI, AppVeyor, Read the Docs, Codecov, CircleCI, zenodo, Anaconda Cloud
Compilation requirements, operating environments & dependencies	Linux, Mac OS, Windows
If available Link to developer documentation/manual	`https://cluster-over-sampling.readthedocs.io/`
Support email for questions	georgios.douzas@gmail.com

Table 1: Code metadata

1. Motivation and significance

1.1. Introduction

The imbalanced learning problem is defined as a machine learning classification task using datasets with binary or multi-class targets where one of the classes, called the majority class, outnumbers significantly the remaining classes, called the minority class(es) [1]. Learning from imbalanced data is a frequent and non-trivial problem for academic researchers and industry practitioners alike. The imbalance learning problem can be found in multiple domains such as chemical and biochemical engineering, financial management, information technology, security, business, agriculture or emergency management [2].

The imbalanced learning problem describes the case where in a machine learning classification task using datasets with binary or multi-class targets, one of the classes, called the majority class, has a significantly higher number of samples compared to the remaining classes, called the minority class(es) [1]. Learning from imbalanced data is a non-trivial problem for both academic researchers and industry practitioners. Additionaly, imbalanced data can be frequently found in multiple domains such as chemical and biochemical engineering, financial management, information technology, security, business, agriculture or emergency management [2].

A bias towards the majority class is induced when imbalanced data are used to train standard machine learning algorithms. This results in low classification accuracy, especially for the minority class(es), when the classifier is evaluated on unseen data. An important measure for the degree of data imbalance is the Imbalance Ratio (IR), defined as the ratio between the number of samples of the majority class and each of the minority classes. Using a rare disease detection task as an example, with 1% of positive cases corresponding to an $IR = \frac{0.99}{0.01} = 99$, a trivial classifier that always labels a person as healthy will score a classification accuracy of 99%. However in this case, all positive cases remain undetected. The observed values of IR are often between 100 and 100.000 [3], [4]. Figure 1 presents an example of imbalanced data in two dimensions as well as the decision boundary identified by a typical classifier when they are used as a training set.

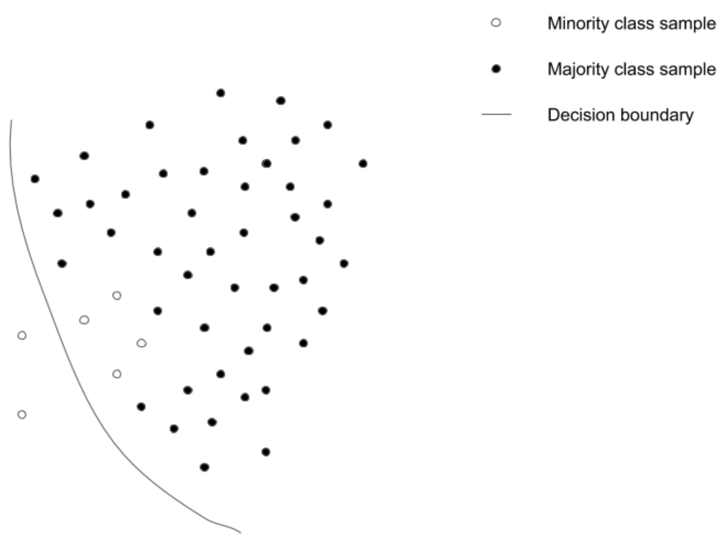

Figure 1: Imbalanced data in two dimensions. The decision boundary of a typical classifier shows a bias towards the majority class.

1.2. Oversampling algorithms

Various approaches have been proposed to improve classification results when the training data are imbalanced, a case also known as between-class imbalance. The most general approach, called oversampling, is the generation of artificial data for the minority class(es) [5]. Synthetic Minority Oversampling Technique (SMOTE) [3] was the first non-trivial oversampler proposed and remains the most popular one. Although SMOTE has been shown to be effective for generating artificial data, it also has some drawbacks [6]. In order to improve the quality of the artificial data many variants of SMOTE have been proposed. Nevertheless, they utilize the SMOTE data generation mechanism, which consists of a linear interpolation between minority class samples to generate synthetic instances as shown in figure 2.

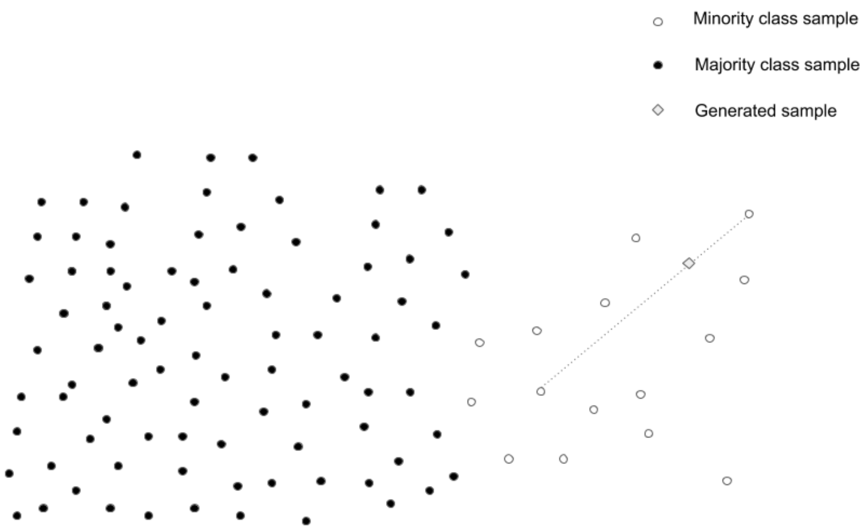

Figure 2: Visual representation of the SMOTE data generation mechanism.

A Python implementation of SMOTE and several of its variants is available in the Imbalanced-Learn [7] library, which is fully compatible with the popular machine learning toolbox Scikit-Learn [8].

1.3. Clustering based oversampling

In addition to between-class imbalance, within-class imbalance refers to the case where areas of sparse and dense minority class instances exist. As a first step of generating synthetic samples, the SMOTE data generation mechanism selects randomly, with uniform probability, minority class instances. Consequently, dense minority class areas have a high probability of being inflated further, while the sparsely populated are likely to remain sparse. This allows to combat between-class imbalance, while the issue of within-class imbalance is ignored [9].

On the other hand, clustering based oversampling, as presented in [10] and [11], aims to deal with both between-class and within-class imbalance problems. Initially a clustering algorithm is applied to the input space. The resulting clusters allow to identify sparse and dense minority class(es) areas. A small IR, compared to a threshold, of a particular cluster is used as an indicator that it can be safely used as a safe data generation area, i.e. noise generation is avoided. Furthermore, sparse minority clusters are assigned more synthetic samples, which alleviates within-class imbalance.

Specific realizations of the above approach are SOMO [10] and KMeans-SMOTE [11] algorithms. Empirical studies have shown that both algorithms outperform SMOTE and its variants across multiple imbalanced datasets, classifiers and evaluation metrics. In this paper, we present a generic Python implementation of clustering based oversampling, in the sense that any combination of a Scikit-Learn compatible clusterer and Imbalanced-Learn combatible oversampler can be selected to produce an algorithm that identifies clusters on the input space and apply oversampling on each one of them. In section 2, the software description is given while section 3 provides a demonstrative example of its functionalities.

2. Software description

The `cluster-over-sampling` software project is written in Python 3.7. It contains an object-oriented implementation of the cluster based oversampling procedure as well as detailed online documentation. The implementation provides an API that is compatible with Imbalanced-Learn and Scikit-Learn libraries. Therefore, standard machine learning functionalities are supported while the generated clustering based oversampling algorithm includes any selected oversampler as a special case.

2.1. Software Architecture

The `cluster-over-sampling` project contains the Python package `clover`. The main modules of `clover` are called `distribution` and `over_sampling`.

The module `distribution` implements the functionality related to the distribution of the generated samples to the identified clusters. It contains the files `base.py` and `density.py`. The former provides the implementation of the `BaseDistributor` class, the base class for distributors, while the later includes the `DensityDistributor` class, a generalization of the density based distributor presented in [10] and [11], that inherits from `BaseDistributor`. Following the Scikit-Learn API, `BaseDistributor` includes the public methods `fit` and `fit_distribute`. The `fit_distribute` method calls the `fit` method and returns two Python dictionaries that describe the distribution of generated samples inside each cluster and between clusters, respectively. Particularly, the `fit` method calculates various statistics related to the distribution process, while it calls `_fit` method to calculate the actual intra-cluster and inter-cluster distributions. This is achieved by invoking the `_intra_distribute` and `_inter_distribute` methods. The `BaseDistributor` class provides a trivial implementation of them, that should be overwritten when a realization of a distributor class is considered. Therefore, DensityDistributor overwrites both methods as well as the `_fit` method. The later calls the methods `_identify_filtered_clusters` and `_calculate_clusters_density` that identify the clusters used for data generation and calculate their density, respectively. Subsection 2.2 provides a detailed description of the initialization and functionality of the `DensityDistributor` class. Figure ?? shows a visual representation of the above classes and functions hierarchy.

The initialization of a `GeometricSMOTE` instance includes G-SMOTE's hyperparameters that control the generation of synthetic data. Additionally, `GeometricSMOTE` inherits from the `BaseOverSampler` class of Imbalanced-Learn library. Therefore, an instance of `GeometricSMOTE` class provides the `fit` and `fit_resample` methods, the two main methods for resampling as explained in subsection 2.2. This is achieved by implementing the `_fit_resample` abstract method of the parent class `BaseOverSampler`. More specifically, the function `_make_geometric_sample` implements the data generation mechanism of G-SMOTE as shortly described in section 1.3. This function is called in the `_make_geometric_samples` method of the `GeometricSMOTE` class in order to generate the appropriate number of synthetic data for a particular minority class. Finally, the method `_make_geometric_samples` is called in `_fit_resample` method to generate synthetic data for all minority classes. Figure 3 provides a visual representation of the above classes and functions hierarchy.

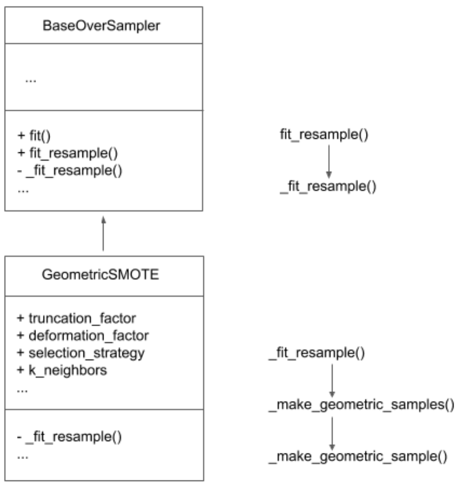

Figure 3: UML class diagrams and callgraphs of main classes and methods.

2.2. Software Functionalities

As it was mentioned in subsection 2.1, the class GeometricSMOTE represents the G-SMOTE oversampler. The intializer of GeometricSMOTE includes the following G-SMOTE's hyperparameters: truncation_factor, deformation_factor, selection_strategy and k_neighbors as explained in subsection ??. Once the GeometricSMOTE object is initialized with a specific parametrization, it can be used to resample the imbalanced data represented by the input matrix X and the target labels y. Following the Scikit-Learn API, both X, y are array-like objects of appropriate shape.

Resampling is achieved by using the two main methods of fit and fit_resample of the GeometricSMOTE object. More specifically, both of them take as input parameters the X and y. The first method computes various statistics which are used to resample X while the second method does the same but additionally returns a resampled version of X and y.

The geometric-smote project has been designed to integrate with the Imbalanced-Learn toolbox and Scikit-Learn ecosystem. Therefore the GeometricSMOTE object can be used in a machine learning pipeline, through Imbalanced-Learn's class Pipeline, that automatically combines samplers, transformers and estimators. The next section provides examples of the above functionalities.

3. Illustrative Examples

3.1. Basic example

An example of resampling multi-class imbalanced data using the fit_resample method is presented in Listing 1. Initially, a 3-class imbalanced dataset is generated. Next, GeometricSMOTE object is initialized with default values for the hyperparameters, i.e. truncation_factor = 1.0, deformation_factor = 0.0, selection_strategy = combined. Finally, the object's fit_resample method is used to resample the data. Printing the class distribution before and after resampling confirms that the resampled data X_res, y_res are perfectly balanced. X_res, y_res can be used as training data for any classifier in the place of X, y.

Listing 1: Resampling of imbalanced data using the fit_resample method.

```python
# Import classes and functions.
from collections import Counter
from gsmote import GeometricSMOTE
from sklearn.datasets import make_classification

# Generate an imbalanced 3-class dataset.
```

```python
X, y = make_classification(
    random_state=23,
    n_classes=3,
    n_informative=5,
    n_samples=500,
    weights=[0.8, 0.15, 0.05]
)

# Create a GeometricSMOTE object with default hyperparameters.
gsmote = GeometricSMOTE(random_state=10)

# Resample the imbalanced dataset.
X_res, y_res = gsmote.fit_resample(X, y)

# Print number of samples per class for initial and resampled data.
init_count = list(Counter(y).values())
resampled_count = list(Counter(y_res).values())

print(f'Initial class distribution: {init_count}.')
# Initial class distribution: [400, 75, 25].

print(f'Resampled class distribution: {resampled_count}.')
# Resampled class distribution: [400, 400, 400].
```

3.2. Machine learning pipeline

As mentioned before, the `GeometricSMOTE` object can be used as a part of a machine learning pipeline. Listing 2 presents a pipeline composed by a G-SMOTE oversampler, a PCA tranformation and a decision tree classifier. The pipeline is trained on imbalanced binary-class data and evaluated on a hold-out set. The user applies the process in a simple way while the internal details of the calculations are hidden.

Listing 2: Training and evaluation of a machine learning pipeline that contains the `GeometricSMOTE` object.

```python
# Import classes and functions.
from gsmote import GeometricSMOTE
from sklearn.datasets import make_classification
from sklearn.decomposition import PCA
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score
```

```python
from imblearn.pipeline import make_pipeline

# Generate an imbalanced binary-class dataset.
X, y = make_classification(
        random_state=23,
        n_classes=2,
        n_samples=500,
        weights=[0.8, 0.2]
)

# Split the data to training and hold-out sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state

# Create the pipeline's objects with default hyperparameters.
gsmote = GeometricSMOTE(random_state=11)
pca = PCA()
clf = DecisionTreeClassifier(random_state=3)

# Create the pipeline.
pip = make_pipeline(gsmote, pca, clf)

# Fit the pipeline to the training set.
pip.fit(X_train, y_train)

# Evaluate the pipeline on the hold-out set using the F-score.
test_score = f1_score(y_test, pip.predict(X_test))

print(f'F-score on hold-out set: {test_score}.')
# F-score on hold-out set: 0.7.
```

4. Impact and conclusions

Classification of imbalanced datasets is a challenging task for standard machine learning algorithms. G-SMOTE, as a enhancement of the SMOTE data generation mechanism, provides a flexible and effective way for resampling the imbalanced data. G-SMOTE's emprical results prove that it outperforms SMOTE and its variants. Machine learning researchers and industry practitioners can benefit from using G-SMOTE in their work since the imbalanced learning problem is a common characteristic of many real-world applications.

The `geometric-smote` project provides the only Python implementation, to the best of our knowledge, of the state-of-the-art oversampling algorithm G-SMOTE. A significant advantage of this implementation is that it is built on top of the Scikit-Learn's ecosystem. Therefore, using the G-SMOTE oversampler in typical machine learning workflows is an effortless task for the user. Also, the public API of the main class `GeometricSMOTE` is identical to the one implemented in Imbalanced-Learn for all oversamplers. This means that users of Imbalanced-Learn and Scikit-Learn, that apply oversampling on imbalanced data, can integrate the `gsmote` package in their existing work in a straightforward manner or even replace directly any Imbalanced-Learn's oversampler with `GeometricSMOTE`.

5. Conflict of Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

[1] N. V. Chawla, A. Lazarevic, L. Hall, K. Boyer, SMOTEBoost: improving prediction of the minority class in boosting, Principles of Knowledge Discovery in Databases, PKDD-2003 (2003) 107–119doi:10.1007/b13634.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.80.1499

[2] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Systems with Applications 73 (2017) 220–239. doi:10.1016/j.eswa.2016.12.035.
URL https://doi.org/10.1016/j.eswa.2016.12.035

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. arXiv:1106.1813, doi:10.1613/jair.953.

[4] S. Barua, M. M. Islam, X. Yao, K. Murase, MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning, IEEE Transactions on Knowledge and Data Engineering 26 (2) (2014) 405–425. doi:10.1109/TKDE.2012.232.

[5] A. Fernández, V. López, M. Galar, M. J. del Jesus, F. Herrera, Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches, Knowledge-Based Systems 42 (2013) 97–110. doi:http://dx.doi.org/10.1016/j.knosys.2013.01.018.
URL http://www.sciencedirect.com/science/article/pii/S0950705113000300

[6] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284. arXiv:arXiv:1011.1669v3, doi:10.1109/TKDE.2008.239.

[7] G. Lemaitre, F. Nogueira, C. K. Aridas, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research 18 (2016) 1–5. arXiv:1609.06570, doi:http://www.jmlr.org/papers/volume18/16-365/16-365.pdf.
URL http://arxiv.org/abs/1609.06570

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Vol. 12, 2011. arXiv:arXiv:1201.0490v2, doi:10.1007/s13398-014-0173-7.2.
URL http://dl.acm.org/citation.cfm?id=2078195

[9] R. C. Prati, G. Batista, M. C. Monard, Learning with class skews and small disjuncts, in: SBIA, 2004, pp. 296–306. doi:10.1007/978-3-540-28645-5_30.

[10] G. Douzas, F. Bacao, Self-organizing map oversampling (somo) for imbalanced data set learning, Expert Systems with Applications 82 (2017) 40 – 52. doi:https://doi.org/10.1016/j.eswa.2017.03.073.
URL http://www.sciencedirect.com/science/article/pii/S0957417417302324

[11] G. Douzas, F. Bacao, F. Last, Improving imbalanced learning through a heuristic oversampling method based on k-means and smote, Information Sciences 465 (2018) 1 – 20. doi:https://doi.org/10.1016/j.ins.2018.06.056.
URL http://www.sciencedirect.com/science/article/pii/S0020025518304997