МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ ПОЛІТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІНСТИТУТ КОМП'ЮТЕРНИХ СИСТЕМ

МАТЕРІАЛИ ДЕВ'ЯТОЇ МІЖНАРОДНОЇ НАУКОВОЇ КОНФЕРЕНЦІЇ СТУДЕНТІВ ТА МОЛОДИХ ВЧЕНИХ

ICS
ODESSA
ONPU

ПРИСВЯЧЕНА 55-РІЧЧЮ
ІНСТИТУТУ КОМП'ЮТЕРНИХ СИСТЕМ

" Сучасні інформаційні технології 2019 "

" Modern Information Technology 2019 "

GAMEHUB

NetCracker®

банк
ПІВДЕННИЙ

23-24 травня

Одеса
«Екологія»
2019

УДК 004.8

HANDLING IMBALANCED CLASSES IN MULTICLASS CLASSIFICATION PROBLEM

Slonskii O.V.
PhD, prof. Arsirii O.O.
Odessa national polytechnic national polytechnic university, UKRAINE

ABSTRACT. In this work, imbalanced classes balancing methods for multiclass classification problem are suggested. Conducted research presents data engineering approach for enhancing machine learning models in multiclass classification problems.

Introduction. There are a lot of multiclass classification problems in data science. We can observe, that most of the data for multiclass classification problems contains objects of classes with non-equal quantitative ratio, which causes dramatically decrease of quality of machine learning models. In this paper, we would like to present class balancing methods and show its ability to enhance machine learning quality.

Goal. The main goal of this work was overview and combining of class balancing methods with purpose to gain machine learning models quality which are being fitted on imbalanced classes data.

Main part.

Statistical distribution of real world data is rarely corresponds to uniform distribution law. Data for most of classification problems contains objects of classes with non-equal quantitative ratio. Models, trained on such data, overfit on majority classes and underfit on minority classes. Such models make a lot of type-I errors on test data.

Measuring accuracy of trained models gives us confusing results. Suppose, dataset consists 80% of class A objects and 20% of class B objects. If model classifies all test objects as class A, its accuracy reach 80%, although model does not analyze passed data at all. Better classifier quality metric on imbalanced classes dataset would be precision, recall, or f1-score (Formula 1), which describe model with ability to overcome as well I type errors as II type errors. For our experiments we will use f1-measure (mean harmonic value of precision and recall) as model quality metric.

$$F_1 = 2\frac{precision \cdot recall}{precision + recall};\ precision = \frac{TP}{TP+FP};\ recall = \frac{TP}{TP+FN}$$

Formula 1 – f1-score, precision and recall, where: TP – true positive, FP – false positive, FN – false negative

Most common methods for classes balancing are:

- up-sampling of minority classes – make copies of the minority classes objects, until we reach nearly equal class ratio;
- down-sampling of majority classes – drop objects of majority classes, until we reach balanced classes ratio;
- generate object of minority classes – using special algorithms or machine learning models and generate so many minority classes sample, as to make class ratio equal [2].

Another method, suggested in our paper, is majority class splitting method. It is to split majority classes objects, using random choice, to groups with size of the smallest minor class. For example, we have 600 objects of class A, 300 instances of class B, and 100 instances of class C. Acting in this method, we will split class A objects to six “sub-classes” with each size of 100 objects and class B object to three “sub-classes” with same size of 100 objects. So we transform imbalanced classes dataset with three classes to balanced dataset with ten classes.

Datasets used in research are:

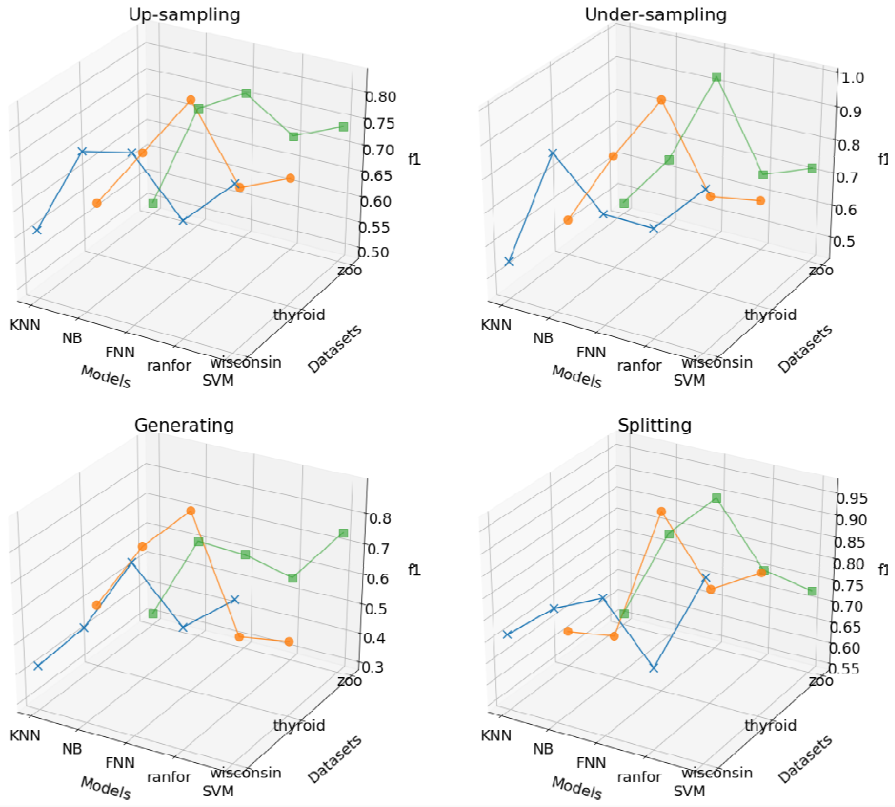
- Breast Cancer Wisconsin dataset (2 classes, 65% objects of majority class, marked as “wisconsin”) [3];
- Thyroid Disease Data Set (3 classes, 83% objects of majority classes, marked as “thyroid”) [4];

- Zoo Data Set (7 classes, 95% objects of majority class, marked as “zoo”) [5].

Models, used in research are:

- K-nearest neighbors (marked as “KNN”);
- naive Bayes (marked as “NB”);
- feed-forward neural network (marked as “FNN”);
- random forest (marked as “ranfor”);
- support vector machine (marker as “SVM”);

Each dataset was balanced with listed before four methods. All classifiers were trained on each dataset with default parameters, set in sklearn library implementation. Results of f1-measure for methods are shown on picture 1.



Picture 1 – Comparison of different classes balancing methods

We can observe, that the beset result are being achieved with majority class splitting method (0.68 average f1-score), but the best common result among all models we get with up-sampling method (0.72 average f1-score). By duplicating minority classes objects, we provide additional information to model, so it might generalize train data as well.

REFERENSES

1. How to Handle Imbalanced Classes in Machine Learning — https://elitedatascience.com/imbalanced-classes.

2. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset — https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset.

3. Breast Cancer Wisconsin (Diagnostic) Data Set — https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

4. Thyroid Disease Data Set — http://archive.ics.uci.edu/ml/datasets/thyroid+disease.

5. Zoo Data Set — https://archive.ics.uci.edu/ml/datasets/zoo.