AI Strategy and Digital transformation
6. Data rebalancing methods

Piotr Wójcik
University of Warsaw (Poland)
pwojcik@wne.uw.edu.pl

January 2025

Unbalanced sample

- the problem of data imbalance refers to the **clasiffication** problem
- **unbalanced sample** is when for a binary dependent variable one of the values occurs **much more often** than the other (in 80-90% and more cases)
- the classification model estimated on such a sample will usually **predict the value occurring more often** much better, especially for the default probability value of 0.5 in many models
- prediction of the **less frequent** value will be subject to **a much larger error**

Unbalanced sample - cont.

- measures of model quality depending on a selected cut-off value (eg. accuracy, specificity, sensitivity, etc.) are **less useful in the case of an unbalanced sample**
- **a much better measure** of the model assessment is in this situation, for example **area under the ROC curve**, independent of the cut-off value or **balanced accuracy**
- it is also possible to transform the training sample by correcting the imbalance

Unbalanced sample – methods of correction

- **there is no single right answer** how to best correct the unbalanced sample
- one can do nothing and be lucky enough to get a good model
- alternatively one can **weight observations** – give larger weights to observations from a less frequent group and smaller weights to observations from the dominant group
- in modeling this will result in a **higher cost of incorrect classification** for observations from a small group
- one can also balance the sample:
  - by **eliminating observations** from a larger class (**down-sampling**, **undersampling**)
  - by **replicating** observations from a smaller class (**up-sampling**, **oversampling**)

More complex methods

- besides random over- and under-sampling, there are **more complex methods**
- instead of creating **copies of existing instances** of minority class, we can generate *synthetic* observations through interpolation
- one may also **combine under-sampling with the generation of additional data**
- two of the most popular complex methods are **ROSE** and **SMOTE**
- another include undersampling with **Tomek link** or **Near Miss**

ROSE

- **ROSE** (*Random OverSampling Examples*) – applies smoothed bootstrapping to draw **artificial observations** from the feature space in the **neighbourhood of the minority class**
- in simple ROSE tries to estimate the probability distribution $P(x|y = k)$ for each class k and then draws the needed N_k observations from it
- one way to estimate such density is through **kernel density estimation** which you can derive from more crude versions such as histogram analysis
- in contrast to random oversampling it **generates a new point** instead of repeating existing observations

SMOTE

- **SMOTE** (*Synthetic Minority Oversampling TEchnique*) – draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space:
  - operating **only on observations from a smaller group** for each observation of i find the k of its nearest neighbors
  - then **create a new** (synthetic) observation assigned to this group **in the middle of the distance** between the observation i and the average of its neighbors
  - and occurs in different variants – it can be combined with the decrease in the size of the dominant group

ROSE and SMOTE – comments

- it is always useful to **check both**, but SMOTE often gives better results than ROSE
- artificial observations created by ROSE tend to be **less realistic** (out of the sensible feature space – e.g. negative age)
- one can **limit this problem** by playing with the parameters defining the neighbourhood in ROSE
- both SMOTE and ROSE usually give **better results** than simple under- or oversampling

Tomek Links Undersampling

- **Tomek Links** method selects pairs of observations (say, a and b) that fulfill the properties:
  - b is the nearest neighbor of a
  - a is the nearest neighbor of b
  - a and b belong to a different classes – minority and majority class (or vice versa), respectively
- observations from the majority class that have **the lowest Euclidean distance** with the minority class and then **removed**

Near Miss Undersampling

- **Near Miss** refers to a collection of undersampling methods that select observations to omit based on the distance of majority class examples to minority class examples
- There are three versions of the technique:
  - **NearMiss-1** selects observations from the majority class that have the **smallest average distance** to the three **closest** observations from the minority class
  - **NearMiss-2** selects observations from the majority class that have the **smallest average distance** to the three **furthest** observations from the minority class
  - **NearMiss-3** involves selecting a given number of majority class observations for each observation in the minority class that are closest

How to do it correctly?

- the main disadvantage of **downsampling** is that we loose potentially relevant information from the left-out samples.
- in **upsampling** there is a risk of overfitting our model as we are more likely to get the same samples in the **training and in the validation** datasets
- resampling should be applied in a **correct way** – we should **not** simply apply over- or under-sampling on our training data and then use cross-validation to find the best model
- during cross-validation we need to perform resampling **on each fold independently** to get a reliable estimate of model performance!

Some Rules of Thumb

- try **down-sampling** if you **have an a lot data** (tens- or hundreds of thousands of instances or more)
- try **up-sampling** if you **don't have a lot of data** (tens of thousands of records or less)
- consider testing random and non-random (e.g. stratified) sampling schemes
- consider testing **different resampled ratios** (e.g. you don't have to target a 1:1 ratio in a binary classification problem, try other ratios)

Result of balancing of the sample

- it is worth to be aware that weighing or balancing the sample will have a **significant effect on measures that depend on the cut-off point**
- this is due to the fact that the artificial balancing of the sample will ensure that the standard usual cut-off point 0.5 will be **close to the incidence of "positives" (1s)** in the corrected training sample
- assessment measures independent of the cut-off point also **may** improve, but **it will not be so significant**
- it is always a good idea to **compare several methods** to see which works best on the analyzed data set
- in general, the **more extreme** initial output imbalances, the **more often** balancing the sample will bring **improvement in the quality of predictions**

Data rebalancing methods – practical exercises in python

Thank you for your attention