Imbalanced dataset

presentation

<https://www.dropbox.com/sh/n90sc69x7gbzxtg/AABvS3QdPgJWm1r_HYPtzdCda?dl=0>

google repo

[solegalli/machine-learning-imbalanced-data: Code repository for the online course Machine Learning with Imbalanced Data (github.com)](https://github.com/solegalli/machine-learning-imbalanced-data)

imbalance dataset: certain number of class too high

machine learning needs balance dataset

small classes diff to found

sp they get mis classified

we are particularly interested in minority classes in situations like(fraud and cancer)

imbalance degree: ration between number of observation from majority class to minority

typically 1:10

**example or use case:**

fraud detection

medical diagnosis

equipment manufacturing and testing

detection of oil spills from radar images of ocean

**nature of imbalance data:**

**some cases ratio is very low 1:35**

**factors that influence the ability of classifier to identify rare events**

**are:**

**small sample size**

**class separability**

**within class sub clusters**

small sample size:

small the data set bigger the error

sample size says how good is your model

having bimbalance data may not be a big problem if sample is s big

class separability:

if patterns among classes overlap- hard to separate them

how clearly different the observations differ from each other- easy to classify

linearly separable are not sensitive to any amount of data

within class imbalance: or sub clusters

each class may have separate sub classes witch are again imbalance

Solutions:

Data level approach- modify dataset

Cost sensitive approach : modify cost(apply higher cost)

Ensemble algorithm: built over several algorithm , combind weak learners , multiple classifiers and combine result or aggregate result

**Data level approach**: modify dataset

Change distribution of data

Random over/under sampling

Create new synthetic data which look like minority(smote)

Remove noise or alternatively

**Additional Reading Resources (Optional)**

Here is a list of reviews that discuss the problem of imbalanced data.

* [A Survey of Predictive Modelling under Imbalanced Distributions](https://arxiv.org/pdf/1505.01658.pdf)
* [Classiﬁcation of imbalanced data: a review](https://www.researchgate.net/publication/263913891_Classification_of_imbalanced_data_a_review)
* [Learning from Imbalanced Data](https://www.ele.uri.edu/faculty/he/PDFfiles/ImbalancedLearning.pdf)
* [Learning from imbalanced data: open challenges and future directions](https://www.researchgate.net/publication/301596547_Learning_from_imbalanced_data_Open_challenges_and_future_directions)
* [Machine Learning from Imbalanced Data Sets 101](http://pages.stern.nyu.edu/~fprovost/Papers/skew.PDF)

Matrix evaluation:

Matrix independent on probability threshold: roc-aoc curve , precision and recall curve

Matrics dependent on probability threshold: accuracy , confusion matric, precision, recall, f score, fpf,fnr, imbalance accuracy

Accuracy:

Percent of correct prediction

Correct predi/total prediction

Tp+tn/all values (TP+TN+FP+FN)

Not suitable for imbalance data because accuracy doent distinguish between the number of correctly classified examples of different class

Minority class has little impact on accuracy when compare to majority

-------------------------

**Precision, recall, f1 score:**

**Recall: sensitivity/tpr: TP/TP+FN- minority class**

**TRUE negative rate: TN/FP+TN- majority class**

**Positive predictive value/precision: TP/TP+FP- minority class**

**Nagative predictive value: TN/TN+FN- majority class**

Precision **: total positive identified by model how many are actually true**

**Recall: total positive correctly identifies as positive, increase recall, decrease misclassification**

**F1 score= 2\*prec\*recall/pre+recall,**

**Support: number of actual occurrence of class in specific dataset**

Change in the threshold will effect recall and precision and f1

1. Threshold – precision will be effect , all shample will be classified as positive as all positive will be classified as positives or minority class(if precision is 1 recall will be one)

As we increase threshold recall will be decreased

--

As we increase threshold – decrease in false positive- precision increases

F measure : balance between 2 matrics

Which is optimal threshold?

Optimal threshold is that at which f measure is highest

-----------------------------------

**Install Yellowbrick**

Make sure you have installed the open source Python package **Yellowbrick**.

If not, from a Mac terminal, Windows command line interface or Anaconda prompt, execute:

1. pip install yellowbrick

More details about **Yellowbrick**:

* [PyPI](https://pypi.org/project/yellowbrick/)
* [Documentation](https://www.scikit-yb.org/en/latest/)
* [Github](https://github.com/DistrictDataLabs/yellowbrick)

Recall gives: how many minority class predicted correctly

Confusion matrix:

Fpr=fp/fp+tn

Fnr=fn/tp+fn

Both between 0to1

Goal is to minimize them

They vary depending on threshold

Minimize FNR: min sick people that we don’t diagnose correctly

Minimize FRR:Drug discovery : minimize number of drugs that we think could be beneficial but they are not