Imbalanced Domain Learning

Rita P. Ribeiro

Machine Learning - 2022/2023

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
[dcc] DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

From Previous Classes

So Far...

- Data mining methodologies
- Classification tasks and scoring
- Data understanding
- Data Preparation

Now...

- Advanced issues in learning tasks
  - Imbalance Domain Learning

References

- Branco, P., et al. 2016.
  A survey of predictive modeling on imbalanced domains. ACM Computing Surveys, 49(2).
- Moreira, João, et al. 2018.
  Data Analytics: A General Introduction. Ch 9, 9.1, 9.2, 11.4.1
- Gama, João, et al. 2015.
  Data Mining 3rd Ed. Ch 9.3.
- Aggarwal, Charu C. 2015.
  Data Mining, the Textbook. Ch 11.3

Today

Imbalanced Domain Learning

- Context
- Applications
- Challenges
- Strategies
  - Data pre-processing
  - Algorithm-level
  - Post-processing predictions

Context

- Most data mining tasks focus on creating a model of the “normal” patterns in the data, extracting knowledge from what is common (e.g. frequent patterns).
- Rare patterns can also give us some crucial insights about data.
- Depending on the goal, those insights can be even more interesting/critical than the “normal” patterns.

Context

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980)

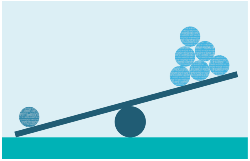
- Initially, outliers were considered errors, and their identification had data cleaning purposes.

- However, they can represent truthful deviation of data.
- For some applications, they represent critical information, which can trigger preventive or corrective actions.

Context

Imbalanced Domain Learning is based on the following assumptions:

1. the cases on the training data are not uniformly represented;
2. the underrepresented cases are the most relevant ones.



- The focus is on the identification of these scarce/outlier cases.
- The definition of these cases depends on the application domain.

Applications with Imbalanced Domains

- Medical Applications
  - Medical Sensor or Imaging for Rare Disease Diagnostics
- Earth Science Applications
  - Sea Surface Temperature Anomalies, Environmental Disasters
- Fault Detection Applications
  - Quality Control, Systems Diagnosis, Structure Defect Detection

Applications with Imbalanced Domains

- Financial Applications
  - Credit Card Fraud, Insurance Claim Fraud, Stock Market Anomalies
- (Cyber) Security Applications
  - Host-based, Network Intrusion Detection
- Text and Social Media Applications
  - Anomalous Activity in Social Networks, Fake News Detection

Challenges

In standard predictive learning tasks:

- The preference is constant over all target variable values.
- The cost of all similar errors and the benefit of all similar accurate predictions is the same.
- To achieve an overall good performance, the learning algorithm focuses on the most frequent cases.

In imbalanced predictive learning tasks:

- Non-uniform importance of values on the domain of the target variable.
- The more relevant cases are poorly represented in the training set.
- The model should be specially accurate at those cases.

Challenges

How to ...

- specify the most important subset(s) of values of the target variable?
  - In some cases can be easy.

    "I'm interested in accurate predictions of fraudulent credit card transactions"

- properly evaluate the performance of models regarding these cases?

- bias the learning algorithms to these rare cases?

Inadequacy of Standard Performance Metrics

- Standard performance metrics (e.g. *accuracy*, *error rate*) assume that all instances are equally relevant for the model performance.
- A good performance is obtained by a model that performs well on normal (frequent) cases and bad on outlier (rare) cases.

  Credit Card Fraud Detection:

  - data set D with only 1% of fraudulent transactions;
  - model M predicts all transactions as non-fraudulent;
  - M has an estimated accuracy of 99%;
  - yet, all the fraudulent transactions were missed!

Challenges

- It is of key importance that the obtained models are particularly accurate at the sub-range of the domain of the target variable for which training examples are rare.

- To prevent the models from being biased to the most frequent cases, it is necessary to consider:
  - performance metrics
    biased towards the performance of these rare cases;
  - learning strategies
    that focus on these rare cases.

Suitable Performance Metrics

Imbalanced Classification Task

- In a classification setting, this type of problem is usually represented by a 2-class problem where outliers are the minority (positive) class.

2-class Confusion Matrix

		True		
		Negative	Positive	Total
Predicted	Negative	TN	FP	PNEG
	Positive	FN	TP	PPOS
	Total	NEG	POS	

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

- Standard performance metrics (e.g. *accuracy*) are unsuitable.

Suitable Performance Metrics

Example: Diagnose of a rare disease

Model B Confusion Matrix			
		Disease	
		absent	present
Diagnose	negative	TN = 63	FN = 2
	positive	FP = 27	TP = 8

Model C Confusion Matrix			
		Disease	
		absent	present
Diagnose	negative	TN = 68	FN = 7
	positive	FP = 22	TP = 3

- The accuracy for both models is 71%.
- Model B correctly diagnosed 80% of the sick individuals
- Model C diagnosed only 30%
- The goal is to achieve a good performance on the rare cases.

Suitable Performance Metrics

- Precision: proportion of positive predictions of the model that are correct.

$$precision = \frac{TP}{TP + FP}$$

- Recall: proportion of the positive cases that are captured by the model.

$$recall = \frac{TP}{TP + FN}$$

- But maximizing one of them comes at the cost of the other.
- It is easy to achieve 100% recall: always predict positive events.
- What is difficult is to achieve high values for both precision and recall.

Suitable Performance Metrics

- F-measure: trade-off measure between precision and recall.

$$F_\beta = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

where β controls the relative importance of *precision* and *recall*

  - when $\beta = 1$, F_1 is the harmonic mean between *precision* and *recall*
  - when $\beta \to 0$, the weight of *recall* decreases
  - when $\beta \to \infty$, the weight of *precision* decreases

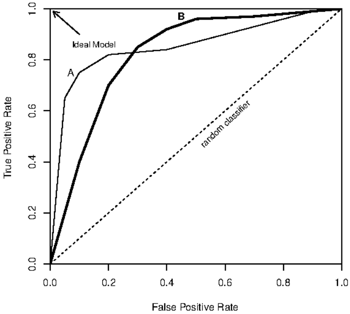
Suitable Performance Metrics

- Receiver Operating Characteristic (ROC) Curve:
  trade-off between *TPR* (*recall*) and *FPR* as the discrimination threshold for the two classes varies.
- False Positive Rate (FPR): proportion of negative cases wrongly predicted as positive.

True Class	Predicted Probability	FPR	TPR	Thr.
1	0.95			
0	0.92	1/4	1/2	> 0.9
0	0.85			
0	0.81	3/4	1/2	> 0.8
1	0.78			
0	0.73	4/4	2/2	> 0.7

Suitable Performance Metrics

- Area Under Curve (AUC) of ROC: performance measure that tells how good the model is in distinguishing the two classes.



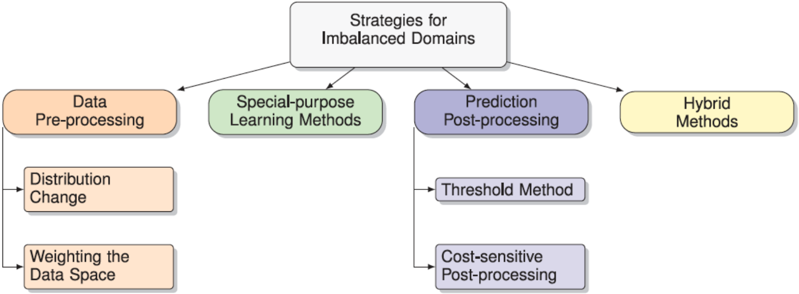
- The higher the AUC, the better.

Suitable Performance Metrics

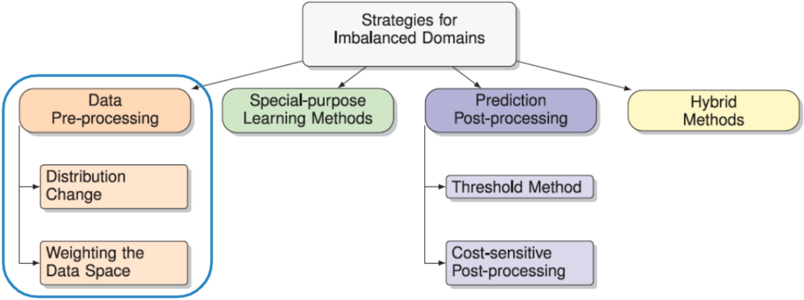
Other metrics that account for the performance in both classes differently:

- AUC-PR
- G-Mean
- Index of Balanced Accuracy
- etc.

Learning Strategies for Imbalanced Domains



Learning Strategies for Imbalanced Domains



Data Pre-Processing Strategies

Proposal

Change the data distribution to make the standard algorithm focus on rare and relevant cases.

Advantages

- They allow the application of any learning algorithm
- The obtained model will be biased toward the domain goals
- Models will be interpretable

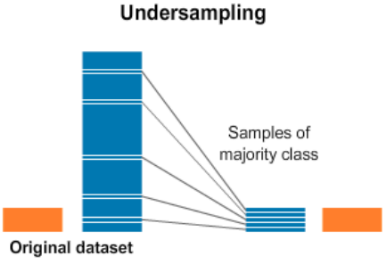
Data Pre-Processing Strategies

Techniques

- **Distribution change**
  - change the data distribution to address the issue of poor representativeness of the more relevant cases
- **Weighting the data space**
  - some algorithms allow different weights to be assigned to different data instances.

Random undersampling

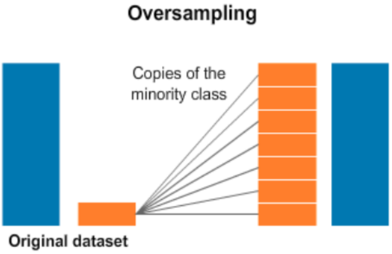
- removes examples from the majority class or with common values from the original dataset, reducing the size of the dataset.
- Problem: useful examples for the learning task may be discarded



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Random oversampling

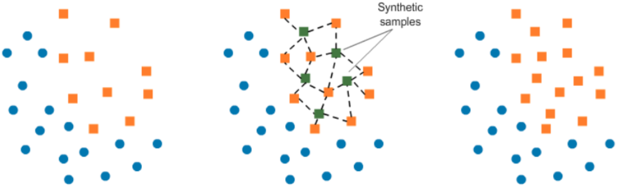
- a random set of copies of minority class or rare values examples are added to the dataset.
- Problem: possible overfitting, i.e. poor generalization ability of the model



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002)

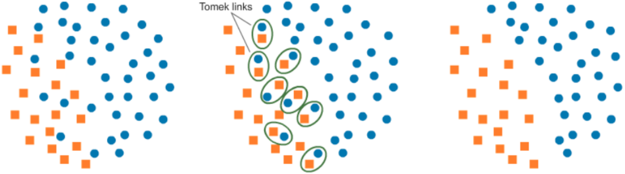
- over-samples the minority class examples by generating new synthetic data;
- reduces the risks of under-sampling and over-sampling;
- creates new examples by interpolating a seed minority example and one of its k minority class nearest neighbours



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

SMOTE can be combined with under-sampling of the majority class

- random under-sampling
- informed under-sampling (e.g. by identifying Tomek links)



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Distribution Change Techniques

SMOTE can be problematic depending on the distribution of minority examples (e.g. too far apart)

- Several SMOTE variants have been proposed that generate synthetic in harder-to-learn regions of the input space
  - put effort into the borders between classes (e.g. Borderline-SMOTE)
  - put effort into minority examples found in spaces dominated by the majority class (e.g. Adaysn)
  - there are many more variants

Distribution Change Techniques

Regardless of the distribution change you choose, one question remains.

- Where should these distribution change strategies be applied?
  - All data set?
  - Just training data?
  - Just test data?

Wrap-up

Advantages

- They allow the application of any learning algorithm
- The obtained model will be biased toward the goals of the user
- Models will be interpretable

Disadvantages

- difficulty of relating the modifications in the data distribution and the user preferences
- mapping the given data distribution into an optimal new distribution according to the user goals is not easy

Weighting the Data Space

- Typically, the goal is to minimize the errors and thus FP+FN
- But, FP and FN can incur different costs
- Ex: Diagnose of a rare disease

Model B Confusion Matrix

		Disease	
		absent	present
Diagnose	negative	TN = 63	FN = 2
	positive	FP = 27	TP = 8

- FP: unnecessary exams and anxiety;
- FN: unnecessary suffering, more expensive procedures, and eventually death;
- imagine cost(FP)=100 and cost(FN)=1000
- the relative cost of the absent (Neg) class is 1, and of the present (Pos) class is 10

Weighting the Data Space

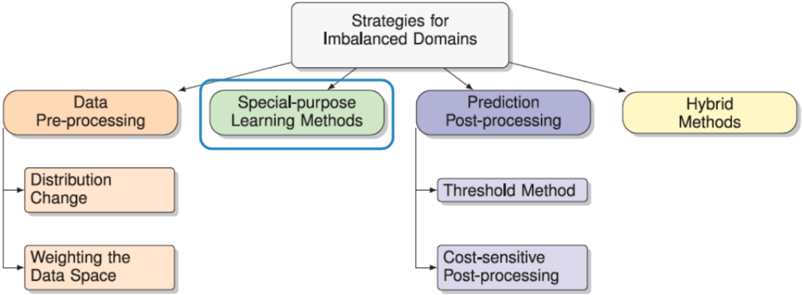
Possible solution:

- Use misclassification costs to obtain a better training distribution
- Each instance is assigned a weight proportional to its importance (relative cost)
- Resampling instances based on their weight

Disadvantages:

- Risk of model overfitting
- Real cost values are often unavailable

Learning Strategies for Imbalanced Domains



Special-purpose Learning Strategies

Proposal

Change the learning algorithm so it can learn from imbalanced data.

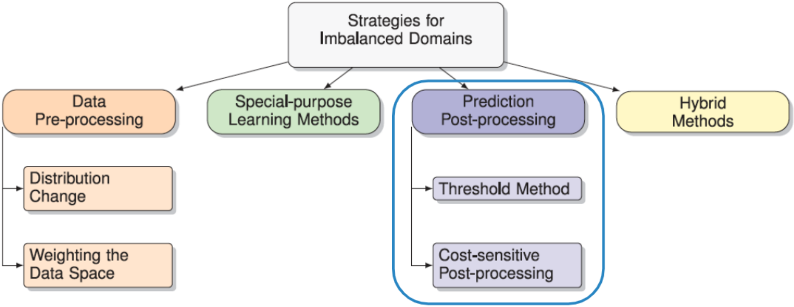
Advantages

- domain preferences incorporated as a preference criterion.
- models will be interpretable

Disadvantages

- restricted set of modified learning algorithms
- if the preference criterion changes, models have to be relearned and, possibly the algorithm has to be re-adapted
- mapping domain preferences with a suitable preference criterion is not straightforward

Learning Strategies for Imbalanced Domains



Prediction Post-processing Strategies

Proposal

Manipulate the predictions of the models according to the domain preferences (e.g. thresholding, cost-sensitive)

Advantages

- original dataset and a standard algorithm
- same model can be applied to different deployment scenarios without having to be relearned

Disadvantages

- the models do not reflect domain preferences
- models interpretability is jeopardized; they are obtained by optimizing a function that is not following the domain preference

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Branco, P., L. Torgo, and R. P. Ribeiro. 2016. “A Survey of Predictive Modeling on Imbalanced Domains.” *ACM Computing Surveys* 49. https://doi.org/10.1145/2907070.

Gama, João, André Carlos Ponce de Leon Ferreira de Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. 2015. *Extração de Conhecimento de Dados: Data Mining -3rd Edition*. Edições Sílabo.

Moreira, João, Andre Carvalho, and Tomás Horvath. 2018. *Data Analytics: A General Introduction*. Wiley.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2018. *Introduction to Data Mining*. 2nd ed. Pearson.