Discriminative Machine Learning for Maximal Representative Subsampling

Laksan Nathan

Outline

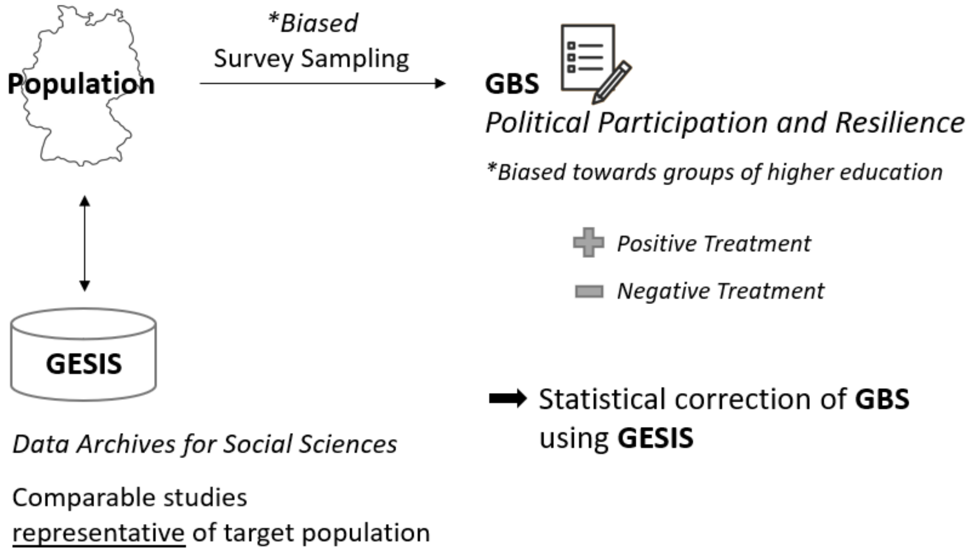
The Learning Problem

- Overfitting
- Covariate Shift

Results

- One-Class Classification
- Logistic Regression
- Positive Unlabeled Learning

Future Work



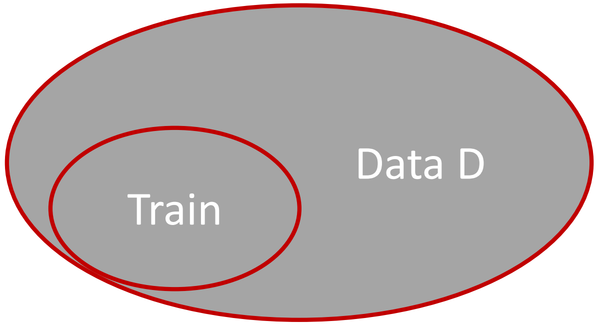
Overfitting

- Error of hypothesis h over training data $\text{errorTrain}(h)$ and error over entire distribution D of data $\text{errorD}(h)$
- Hypothesis h in H **overfits** training data if there exists an alternative hypothesis $\bar{h}$ in H such that

$$\text{errorTrain}(h) < \text{errorTrain}(\bar{h})$$

and

$$\text{errorD}(h) > \text{errorD}(\bar{h})$$



Underfitting:

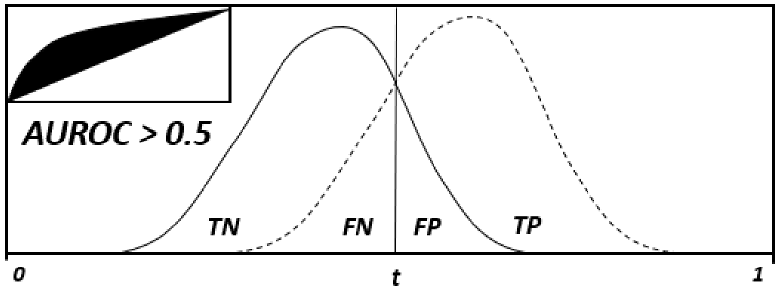
- Predictor too simplistic/rigid
- Not powerful enough to capture pattern in data
- There exists an alternative hypothesis $\bar{h}$ in H with smaller errorTrain and smaller errorD.

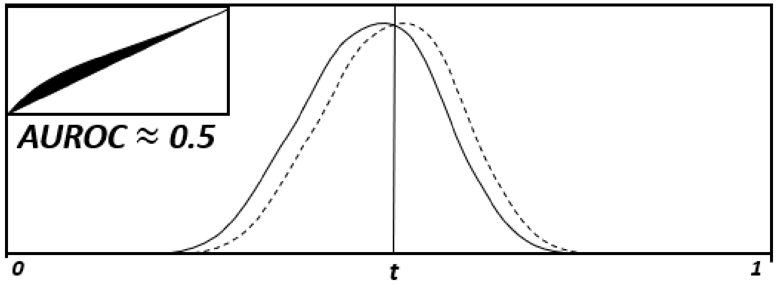
Covariate-Shift

Basic premise for (traditional) machine learning:
Training and test set are drawn i.i.d. from the same probability distribution.

Discriminative learners to predict the **probability** of the source of survey participant origin (GESIS or GBS).

- Instance classified as GBS: regions of feature space **overrepresented** by GBS
  → Remove GBS instances from result set (Undersampling).
- **(AU)ROC** as proxy measure for degree of bias.



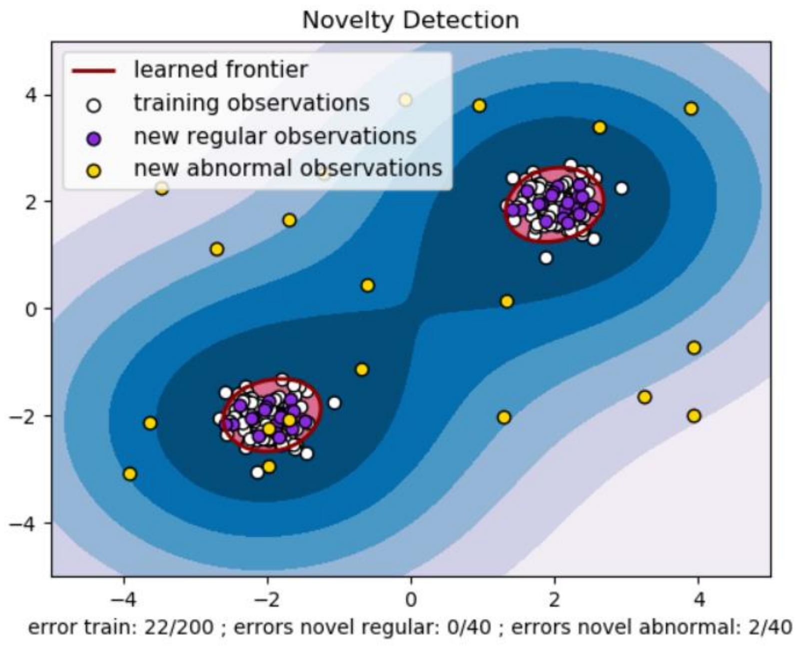


One-Class Support Vector Machines

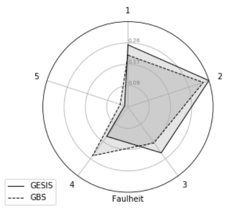
One-class SVMs to classify GBS as similar or different to GESIS.

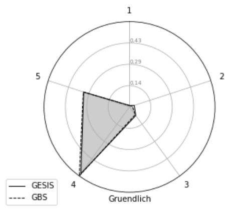
- There is no proper probabilistic interpretation of SVM. Platt scaling as small „cheat“ simply fits another, **probabilistic model**, typically Logistic Regression, on top of SVM projection).

- Instead, OCC-SVM voting ensemble trained on **resamples** of GESIS to get probabilities.



One-Class Support Vector Machines

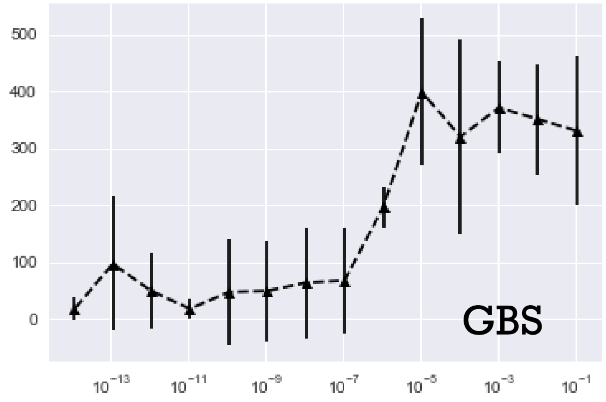


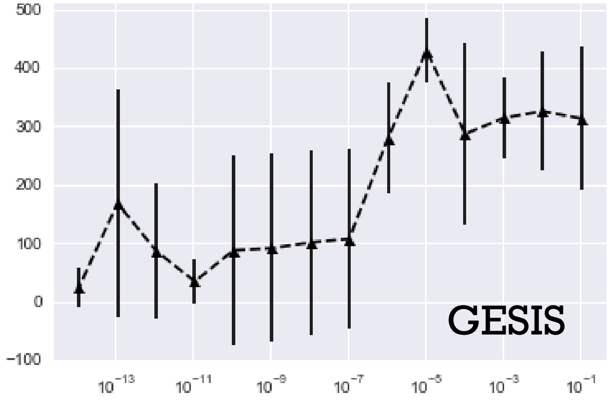


Ideally, OCC-SVMs correctly classify most GESIS instances.
Experiments:

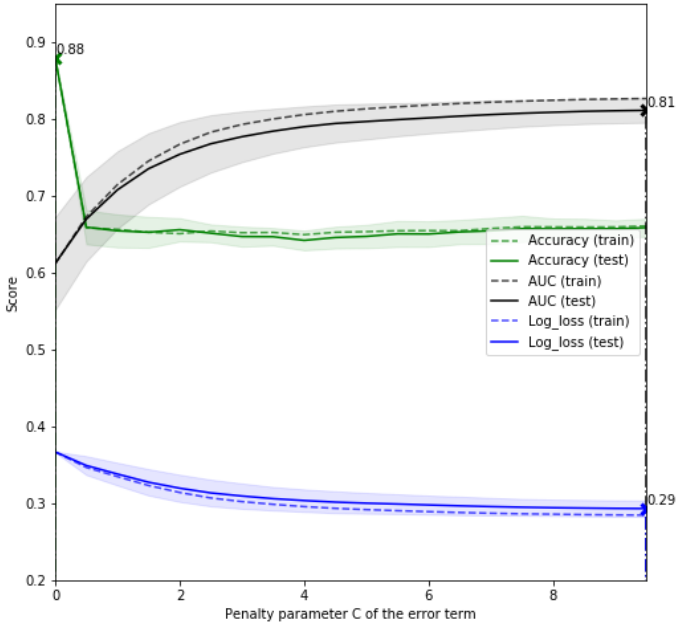
- Ensemble suffers from **high variance.**
- Not enough training data (?)

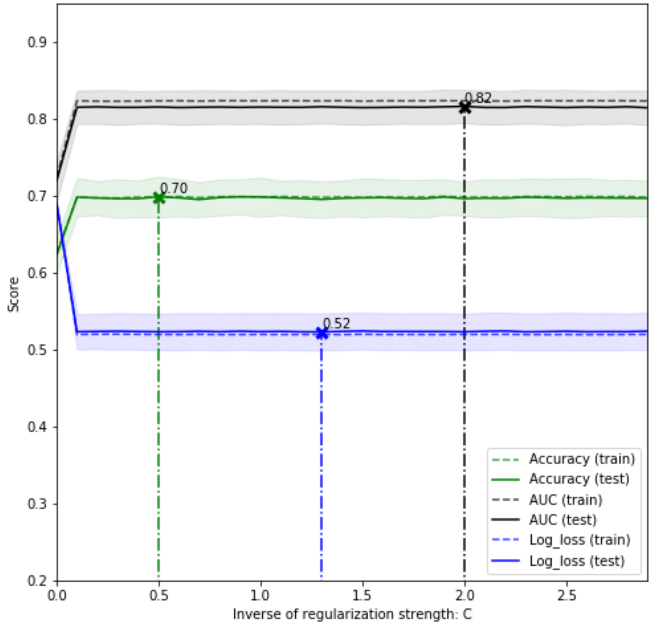
→ GESIS and GBS **indistinguishable** (with respect to BFI-10 attributes).

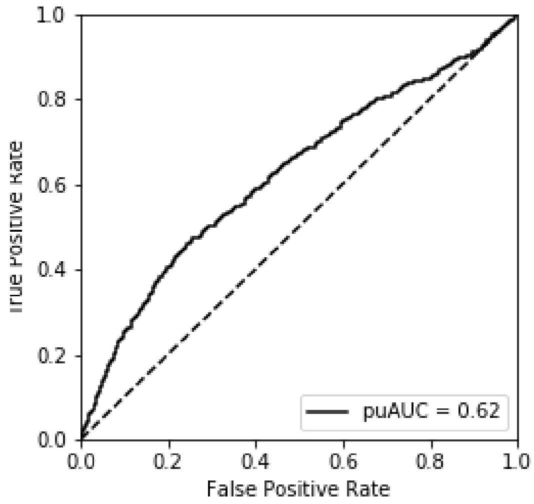




Binary Classification







Linear SVM (left) outperformed by Logistic Regression (right).

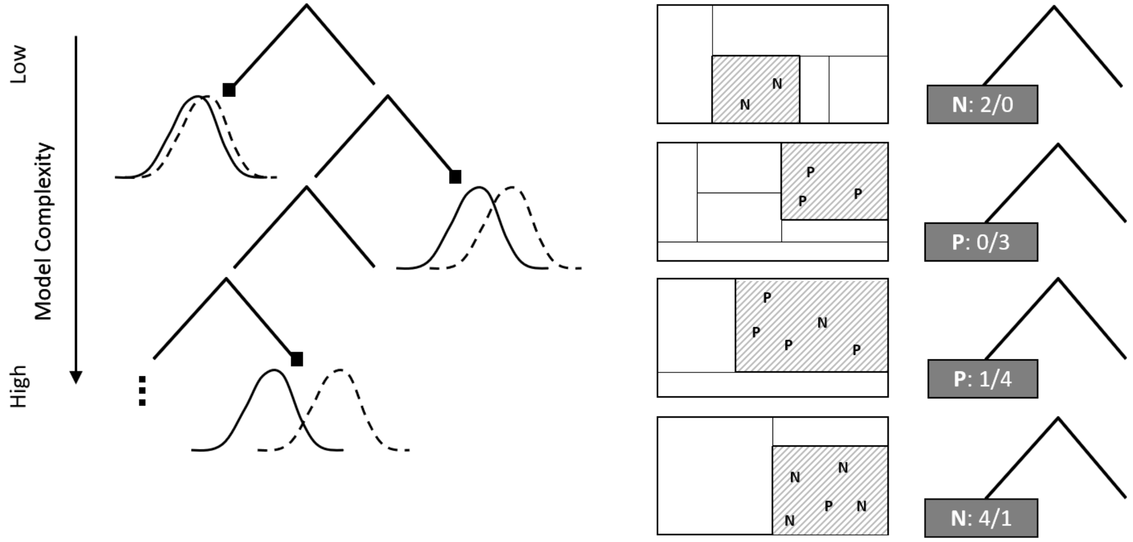
- LR with **puROC = 0.62** on hold-out set.
- Nested strat. 10-fold CV for param tuning and training.

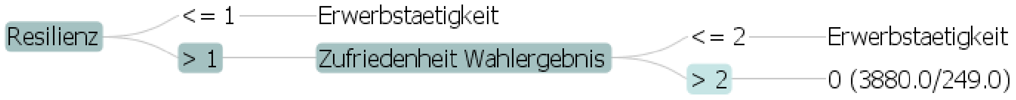
Positive Unlabeled Learning
(or: Iterative Removal of Most Accurate TNs)

*Whats the meaning of a **negative** instance in the context of „representative or not“?*

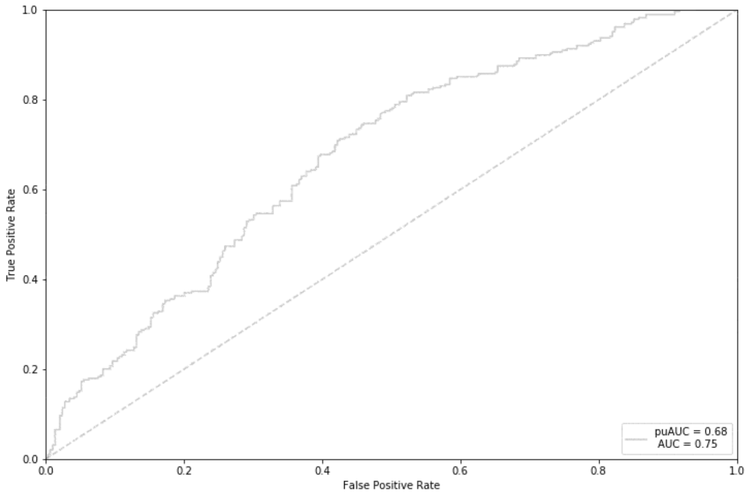
- So far: Remove every correctly classified GBS instance (TN).
- Now: **Iteratively** remove the GBS instance with the highest predicted probability (TN). **Stop: AUROC ≈ 0.5.**

Additionally:
Interpret learning problem as Positive-Unlabeled setting. AUROC can then be corrected with puAUROC estimation.



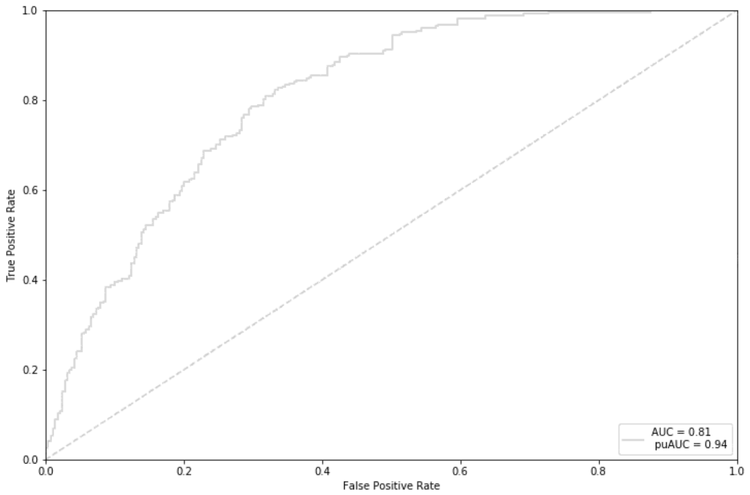


Positive Unlabeled Learning



Zurückhaltend, leicht Vertrauen, Faulheit, Entspannt, wenig künstlerisches Interesse, Gesellig, Andere kritisieren, Gründlich, Nervös, Phantasievoll, Geschlecht, Netto-Haushalt, Netto-Selbst, Geburtsjahr, Geburtsland

puAUC reduction: 0.15 - 0.18



In addition: Berufsgruppe, Aktiv, Familienstand, Berufliche Ausbildung

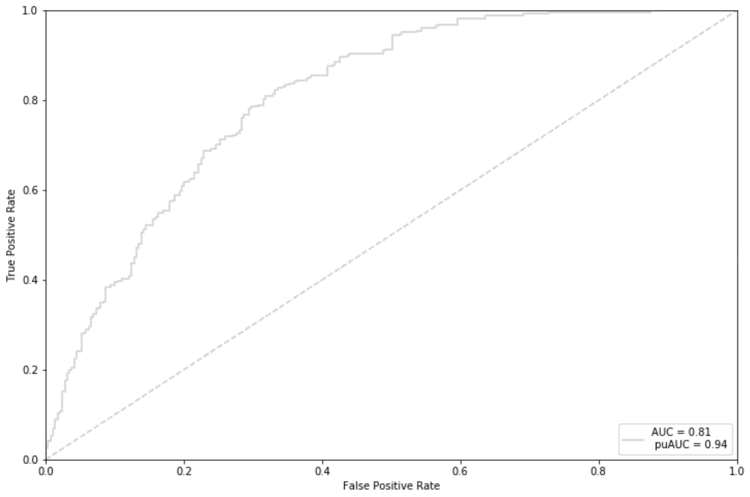
puAUC reduction: 0.27 - 0.35

Positive Unlabeled Learning

Running (PU)-algorithm on these (additional) attributes pinpoints potential weaknesses:

If not properly stopped, then AUC keeps decreasing without **converging** to „random guessing".

- Stopping criterion cannot be (pu)AUC = 0.5.
- Stopping criterion cannot be Highest Predicted Probability (of GBS) = 0.5.



In addition: Berufsgruppe, Aktiv, Familienstand, Berufliche Ausbildung

puAUC reduction: 0.27 - 0.35

Future Work

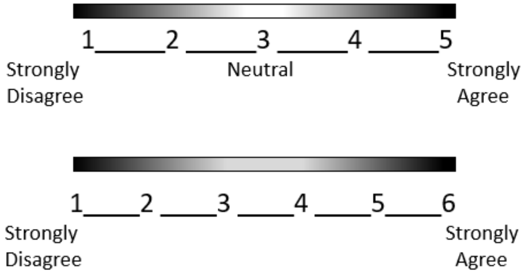
- **SMOTE** as another way to validate results („*oversampling before undersampling*“).
- Test methodology on academic datasets (less prone to mistakes in preprocessings).
- Introduce stopping criterion (based on proper scoring rules). Only if number of instances to exclude is not known in advance.
- For GESIS vs. GBS use F-Measure instead of AUC. This is due to high class imbalance.
- Analyse Prediction of **Political Participation and Resilience** in GBS before and after Maximal Representative Subsampling

and…

…wait for new dataset „Brief Resilience Scale (BRS)“.

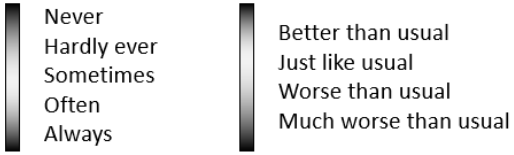
Appendix

(a)

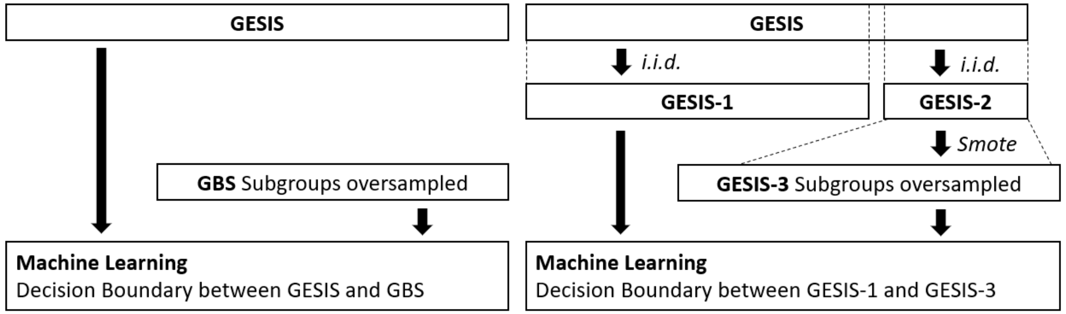


1 2 3 4 5
Strongly Disagree
Neutral
Strongly Agree
1 2 3 4 5 6
Strongly Disagree
Strongly Agree

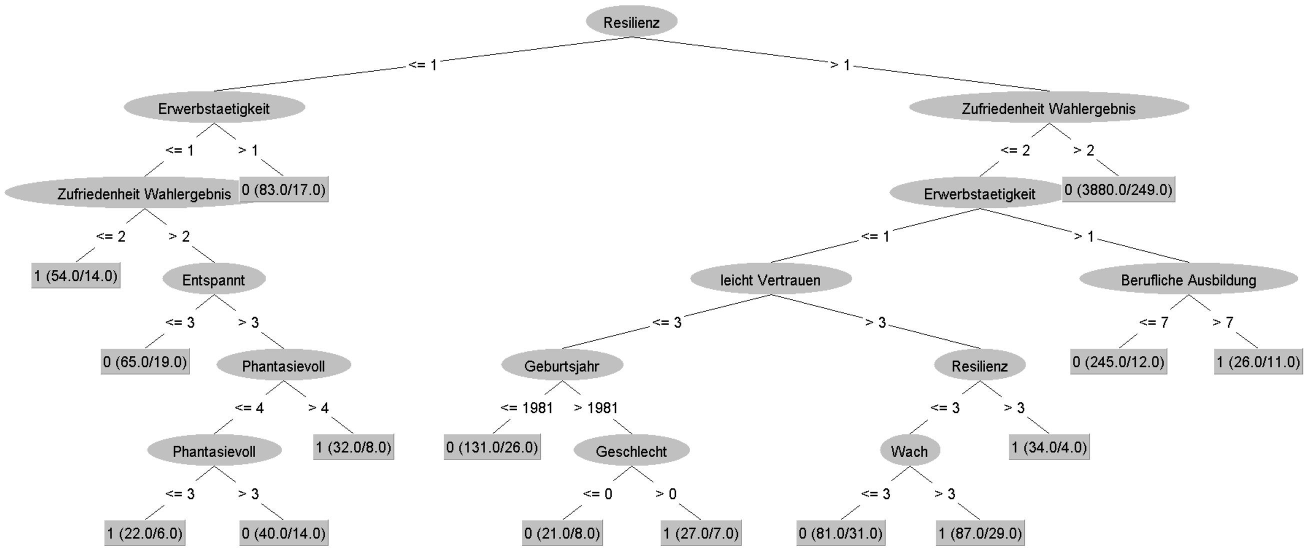
(b)



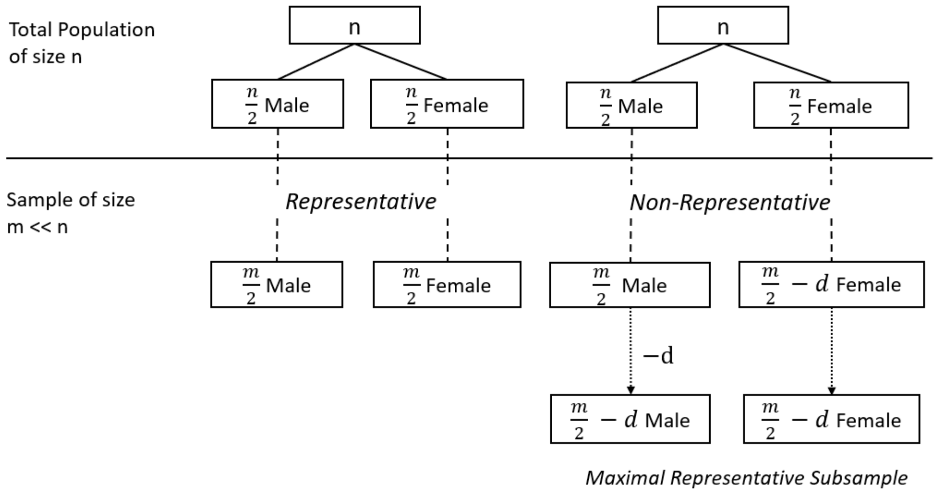
Never
Hardly ever
Sometimes
Often
Always
Better than usual
Just like usual
Worse than usual
Much worse than usual



GESIS
GBS Subgroups oversampled
Machine Learning
Decision Boundary between GESIS and GBS
GESIS
i.i.d.
i.i.d.
GESIS-1
GESIS-2
Smote
GESIS-3 Subgroups oversampled
Machine Learning
Decision Boundary between GESIS-1 and GESIS-3



Resilienz
<= 1
> 1
Erwerbstaetigkeit
<= 1
> 1
Zufriedenheit Wahlergebnis
0 (83.0/17.0)
<= 2
> 2
1 (54.0/14.0)
Entspannt
<= 3
> 3
0 (65.0/19.0)
Phantasievoll
<= 4
> 4
Phantasievoll
1 (32.0/8.0)
<= 3
> 3
1 (22.0/6.0)
0 (40.0/14.0)
Zufriedenheit Wahlergebnis
<= 2
> 2
Erwerbstaetigkeit
0 (3880.0/249.0)
<= 1
> 1
leicht Vertrauen
Berufliche Ausbildung
<= 3
> 3
<= 7
> 7
Geburtsjahr
Resilienz
0 (245.0/12.0)
1 (26.0/11.0)
<= 1981
> 1981
0 (131.0/26.0)
Geschlecht
<= 0
> 0
0 (21.0/8.0)
1 (27.0/7.0)
<= 3
> 3
Wach
1 (34.0/4.0)
<= 3
> 3
0 (81.0/31.0)
1 (87.0/29.0)



Total Population of size n
n
n
$\frac{n}{2}$ Male
$\frac{n}{2}$ Female
$\frac{n}{2}$ Male
$\frac{n}{2}$ Female
Sample of size m << n
Representative
Non-Representative
$\frac{m}{2}$ Male
$\frac{m}{2}$ Female
$\frac{m}{2}$ Male
$\frac{m}{2} - d$ Female
$-d$
$\frac{m}{2} - d$ Male
$\frac{m}{2} - d$ Female
Maximal Representative Subsample

Appendix

Algorithm 1: PU training procedure

Input: P: set of positive instances (GESIS)

U: set of unlabeled instances (GBS)

n_{models}: number of base models in ensemble

n_P: size of bootstrap sample of P

n_U: size of bootstrap sample of U

Result: Scoring function $f : U \to \mathbb{R}$

Initialize: $f(x) \leftarrow 0$ and $c(x) \leftarrow 0$

for *t = 1 to n_{models}* **do**

Draw a bootstrap sample P_t of size n_P.

Draw a bootstrap sample U_t of size n_U.

Train classifier f_t to discriminate P_t against U_t.

For any $x \in U \backslash U_t$, update:

$f(x) \leftarrow f(x) + f_t(x)$,

$c(x) \leftarrow c(x) + 1$

end for

Return: $s(x) = f(x)/c(x)$

Instead: Remove correctly classified GBS instance with highest predicted probability.