Discriminative Machine Learning for Maximal Representative Subsampling

Bachelor's Thesis submitted

to

Prof. Dr. Stefan Kramer

and

Prof. Dr. Andreas Hildebrandt

Johannes-Gutenberg University of Mainz

Institute for Computer Science

Chair of Data Mining

by

Laksan Nathan

(2715043)

in partial fulfillment of the requirements

for the degree of

Bachelor of Science

Mainz, October 15, 2018

ABSTRACT

To allow statistical inference in social sciences, survey participants must be selected at random from the target population. When samples are drawn from parts of the population that are close to hand, subgroups might be over-represented. This leads to statistical analyses under sampling bias, which in turn may produce similarly biased outcomes. The present thesis uses machine learning to reduce this selection bias in a psychological survey using auxiliary information from comparable studies that are known to be representative. Discriminative algorithms are trained to directly characterize the divergence between representative and non-representative samples.

CONTENTS

1 INTRODUCTION

Psychological resilience is generally regarded as positive adaptation to past and ongoing exposure to potential negative effects of stressors. Accordingly, adaptation to stressful or adverse situation is a dynamic process with predictors that can differ between population groups. Within the discipline of developmental psychology, Tuescher and colleagues have provided prospective studies investigating the concept of resilience and its complex underlying mechanisms. As part of a doctoral dissertation, their studies aimed to validate the following research questions:

- Does resilience have a positive effect on the willingness to participate in politics, specifically in election?
- Does the confrontation with positive or negative statements on politics for people with lower resilience have stronger effects on the willingness to participate in politics?

The research group did its poll by selecting people from Mainz, while trying to generalize to the entire German population. The survey data (GBS, n=587) tends to over-represent groups of higher income and higher education, since participants are primarily selected from an academic environment.

Therefore, the validity of assertions about the population beyond the original observation range is affected, even if statements are made conditional upon the available data. The basic premise for standard statistical conclusions, that the training and test set are drawn independently and identically (i.i.d.) from the same probability distribution, does not hold any more. This setting is also known as covariate-shift (Shimodaira 2000). Data sets are rarely generated under ideal conditions with bias pervasive in almost all empirical studies. The oftentimes underestimated analytic errors produce misleading descriptions of populations and ultimately yield false inferences (Aurelien, West, Sakshaug, 2016).

To get a complete picture of the subject, the research group consulted the department Data Archives for the Social Sciences. Their data archive service (GESIS) holds representative data of comparable studies in politics and psychology. The acquired sample (GESIS, n=4000) encompasses the German speaking population with permanent residence in Germany.

This thesis is a practical application to reduce the sampling bias by selecting a maximal representative subsample (MRS) of GBS survey respondents with reference probability distri-

butions from GESIS. The effects of positive and negative treatments on political participation are then analysed in the resulting MRS and compared to the initial GBS data [Fig. 1.1].

To evaluate the research questions to a certain required level of significance, it is inevitable to keep the exclusion of instances at a minimum. Pruning the GBS data in any way, narrows the data variance and thus the reach of subsequent studies. This is especially harmful since the initial GBS survey data is already small.

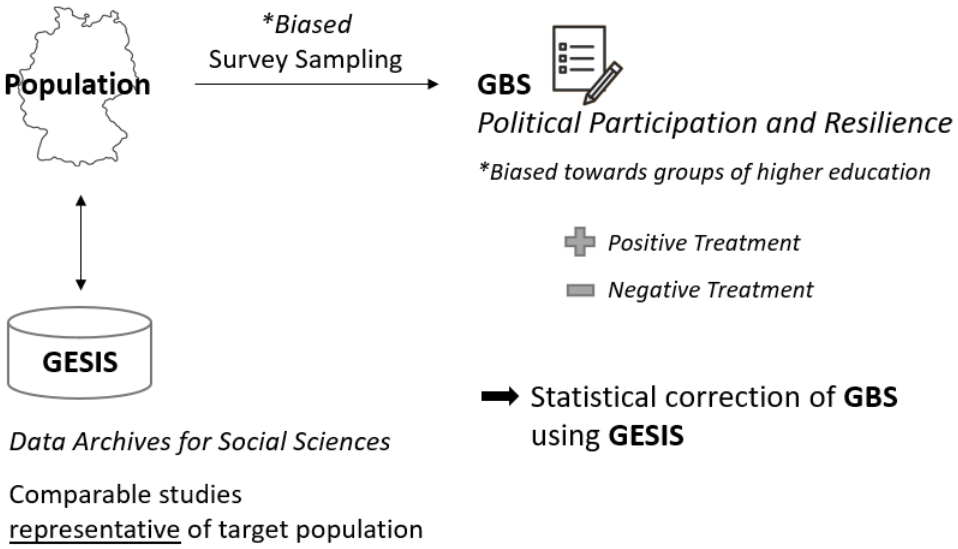


Figure 1.1: Multivariate auxiliary information GESIS linked to GBS so that expected bias can be detected and corrected for. In addition, GBS contains an attribute for positive or negative treatment of survey participants for further analysis.

In the MRS procedure, discriminative learners will look for decision boundaries to distinguish the negative class GBS from the positive class GESIS. First, both data sets are combined by adding an attribute indicating the source of origin. This label is predicted so that instances can be ranked according to their predicted probability. GBS participants that can be distinguished are removed from the result set. The remaining instances in GBS define the MRS and are expected to be more closely aligned with the target probability distribution. The area under the receiver operating characteristic curve (AUROC) is used as single number evaluation metric to measure the degree of sampling bias. The MRS is characterized by an AUROC of roughly 0.5.

The fraction of misclassified GBS instances is kept as proxy measure for the subsequent method positive-unlabeled learning (Denis et al. 2005). Positive-unlabeled learning is a semi-

supervised technique that does not make the simplifying assumption of GBS being positive or negative. GBS is in actual fact unlabeled, i.e. subgroups of the target population might be over-represented or actually representative. Apparently, subsampling does not help in the case of under-represented subgroups.

These procedures can only work when most of the observations come from a feature space which is not specific to one of the surveys. Thus, attributes must be matched correctly.

1.1 Related Work

The influence of sampling bias could be alleviated by weighting the instances according to their importance (Shimodaira 2000). Statistical adjustment might also be reached by developing survey weights for calibration estimators [5]. However, these techniques require density estimation which is known to be a hard problem especially in high-dimensional cases (Haerdle et al. 2004). Directly estimating the importance without estimating the density ratios would be more promising: Discriminative machine learning for maximal representative subsampling.

1.2 Outline

The remainder of this thesis is organized as follows. Sec. 2 starts with an initial data analysis step and focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, i.e. handling missing values and making transformations of variables. Sec. 3 defines key terminology and definitions and discusses discriminative learning and positive-unlabeled learning. The resulting maximal representative subset of GBS is presented in Sec. 4 and compared to GBS regarding political participation and resilience. Sec. 5 concludes.

2 INITIAL DATA ANALYSIS

In order to diagnose to what extent an algorithm suffers from sampling bias, it will be useful to have another dataset. Initial data analysis is conducted independently of the problem statements to understand what properties of the data differ between GBS and GESIS for matching attributes. A brief characterization of the data currently employed in the studies is given in this chapter. The GitHub repository further specifies the list of transformations that are sequentially applied to each group of features in order to prepare the inputs for survey comparisons. Preprocessing steps and methods used to evaluate outcomes are documented as well. Scaling methods that apply to both data sets, e.g: centering and scaling of skewed continuous features for SVMs, are not mentioned but can be deduced from code easily. If an attribute is removed at some point, it will only be mentioned in the relevant section.

CODE	GBS+00027	GESIS 288506501
Gruppe	NEGATIV	
Geschlecht	männlich	Männlich
Geburtsjahr	1949-01-15	1946
Geburtsland	Deutschland	Deutschland
Nationalitaet	1	Deutschland
Familienstand	Verheiratet und lebe mit meinem/r Ehepartner/-...	Verheiratet/ Eing. LP zus. lebend
Hoechster Bildungsabschluss	Realschulabschluss (Mittlere Reife)	Fachhochschulreife, Fachoberschule
Berufliche Ausbildung	Ausbildung an einer Fach-, Meister-, Techniker...	Fachhochschulabschluss
Erwerbstaetigkeit	1	Nicht erwerbstätig
Berufsgruppe	Beamter/Beamtin, Richter/-in, Berufssoldat/-in	Missing by filter
Personen im Haushalt	2	3
Nettoeinkommen Selbst	2 250\t bis unter\t2 500 Euro	2600 bis unter 3200
Nettoeinkommen Haushalt	4 500\t bis unter\t5 000 Euro	4000 bis unter 5000
Schlechter Schlaf	3	Manchmal
Leben genießen	3	Meistens
Zu Nichts aufraffen	3	Nie
Alles anstrengend	2	Fast nie
Wahlteilnahme	NaN	Ja
Wahlabsicht	5	Ja, ich würde wählen.
Desinteresse Politiker	4	3
Zufriedenheit Leben	3	9
Aktiv	3	Erheblich
Verärgert	4	Ein bisschen
Wach	4	Erheblich
Nervös	3	Gar nicht
Ängstlich	4	Gar nicht
Zurueckhaltend	eher zutreffend (4)	3 Weder noch
leicht Vertrauen	eher zutreffend (4)	3 Weder noch
Faulheit	trifft eher nicht zu (2)	1 Trifft überhaupt nicht zu
Entspannt	eher zutreffend (4)	3 Weder noch
wenig kuenstlerisches Intere	eher zutreffend (4)	4 Eher zutreffend
Gesellig	weder noch (3)	4 Eher zutreffend
Andere kritisieren	weder noch (3)	3 Weder noch
Gruendlich	eher zutreffend (4)	4 Eher zutreffend
Nervoes	trifft eher nicht zu (2)	3 Weder noch
Phantasievoll	weder noch (3)	4 Eher zutreffend

Figure 2.1: GBS - GESIS attribute and value comparison. Not all attributes are used in every learning task. See GitHub documentation for more information.

2.1 Feature Selection and Data Imputation

If not stated differently, deletion of rows is applied to every instance with missing values in GESIS to reduce the class imbalance in the later described classification problem. Missing values are sparse in GBS and can be imputed, e.g: with median substitution, with negligible effects. Mean substitution can not be used as this might lead to previously unseen values. Discriminative algorithms will then use these new values to distinguish GBS and GESIS that have been created by ill-considered data imputation. Note that the following figures represent the data after attribute and value matching described in Sec. 2.2.

Fig. A.2 lends itself to first thoughts about whether missing data elements depend on observable attributes or occur entirely at random and is also able to detect functional dependencies in GESIS. Fig. A.3 summarizes the main observations from GBS. The correlation matrix with ratio -1 to +1 in Fig. A.4 is used as another way to visualize differences in GESIS and GBS and to further simplify preprocessing decisions. Potential bugs and issues can also be detected with this graph.

2.2 Perspectives on Response Styles

In survey analysis, scales measuring attributes need to be reliable and valid. Therefore, GBS and GESIS almost entirely use already tested scales from the literature. Although the same characteristics are asked in both surveys, they may have been covered differently, for example by different scales. In addition, GBS surveys are generally more detailed. Especially characteristics which are not exactly predefined are often recorded differently. Features must be engineered carefully, whereby potential loss of information must be minimized. Shortcomings in attribute mappings may result in inappropriate representation and therefore incorrect conclusions. The Big Five are a consistent set of attributes across the surveys. In contrast, there also exist vivid examples for data mismatch on Likert-type scales.

2.2.1 Likert-Type Scale

The Big Five is an empirically-derived model of human personality and psyche. When factor analysis is applied to personality survey data, five clusters of traits consistently emerge. The BFI-10 is a 10-item scale measuring the Big Five personality traits. Figures 2.5 and 2.6 visualizes the response distribution of two BFI items for each dimension, representing both

the high and low pole of each factor. Likert scales are the most frequently used instruments in GBS and GESIS. They consist of statements which measure the intensity of one's estimation towards the preceding statement. Respondents are asked to rate the BFI-10 items on a level of agreement on a consistent rating scale ranging from "Strongly Agree (5)" to Strongly Disagree (1)" for all items in both survey.

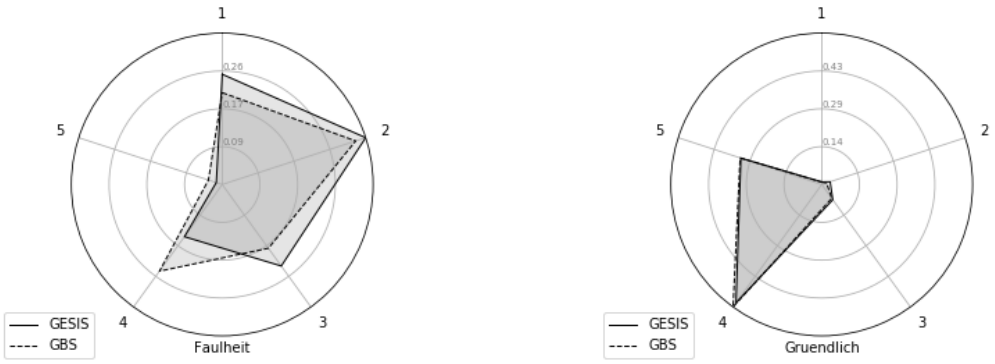


Figure 2.5: Conscientiousness is the degree of organization, self-regulation, and responsibility one exhibits. *"I see myself as someone who tends to be lazy."*(left). *"I see myself as someone who does a thorough job."*(right). The graphs are almost identical for the Likert item "Gruendlich". Respondents specify their level of disagreement for "Faulheit" stronger in GBS.

There are still discussions about whether to use a Likert scale element as a categorical or numerical characteristic. The intervals between positions on the scale are monotonous, but never so well defined that they are numerically uniform steps. A "Strongly Agree (5)" answer indicates more agreement than "Agree", but it shows no agreement that is five times stronger than "Strongly Disagree (1)". In this work, the BFI-10 items are considered categorical.

2.2.2 Data Mismatch

It can be assumed that different numbers of rating bars in a subjective rating scale can have significant effects on the subjective measurement. Table 2.1 gives an example for two different scales of the same attribute while trying to find a proper representation in Table 2.2. Figure 2.7 shows what happens when attributes are partially transformed using cut-off mappings.

attribute	GBS values	GBS values count	GESIS values	GESIS values count
Wach	4	311	Einigermassen	1697
	3	183	Erheblich	1389
	2	66	Ein bisschen	467
	1	14	Aeusserst	367
	-1	1	Gar nicht	184
	-9	1		

Table 2.1: "Wach" as an example of a Likert item discrepancy. GESIS uses an odd number of responses with a "neutral" option, such as "no opinion", "neither agree nor disagree" or some phrase to that effect. In contrast, there is an even number of responses for this item in GBS encouraging participants to voice a positive or negative opinion. In some cases, an additional "opt-out" option is provided for those respondents who truly cannot respond in GBS only (here: value "-9"). Missing values are indicated by "-1".GESIS often contains textual values, which often cannot be clearly assigned, so that the value counts have to be used to support the decision making.

Raw Data	GESIS	1	2	3	4	5	6
	GBS	1	2	3	4	5	
Max Scaler	GESIS	0.83	1.7	2.5	3.3	4.2	5.0
	GBS	1	2	3	4	5	
Min-Max Scaler	GESIS	1	1.8	2.6	3.4	4.2	5
	GBS	1.0	2.25	3.5	4.75	6.0	
Cut-Off Mapping	GESIS	1	2	3	4	5	5
	GBS	1	2	3	4	5	

Table 2.2: Demonstration of potential scalings for differing attribute values. The cut-off mapping comes with a loss of information. Max scaler and min-max scaler introduce new unseen values. Depending on the discriminative algorithm, values that only appear in one of the two surveys are enough to perfectly classify instances: "*If value equals 1.7* $\rightarrow$ *Instance of class GESIS.*" Decision-tree based learning algorithms will likely suffer from erroneous mappings, while logistic regression with a proper scoring rule might be the most suitable algorithm to limit the effects of measurement discrepancy.

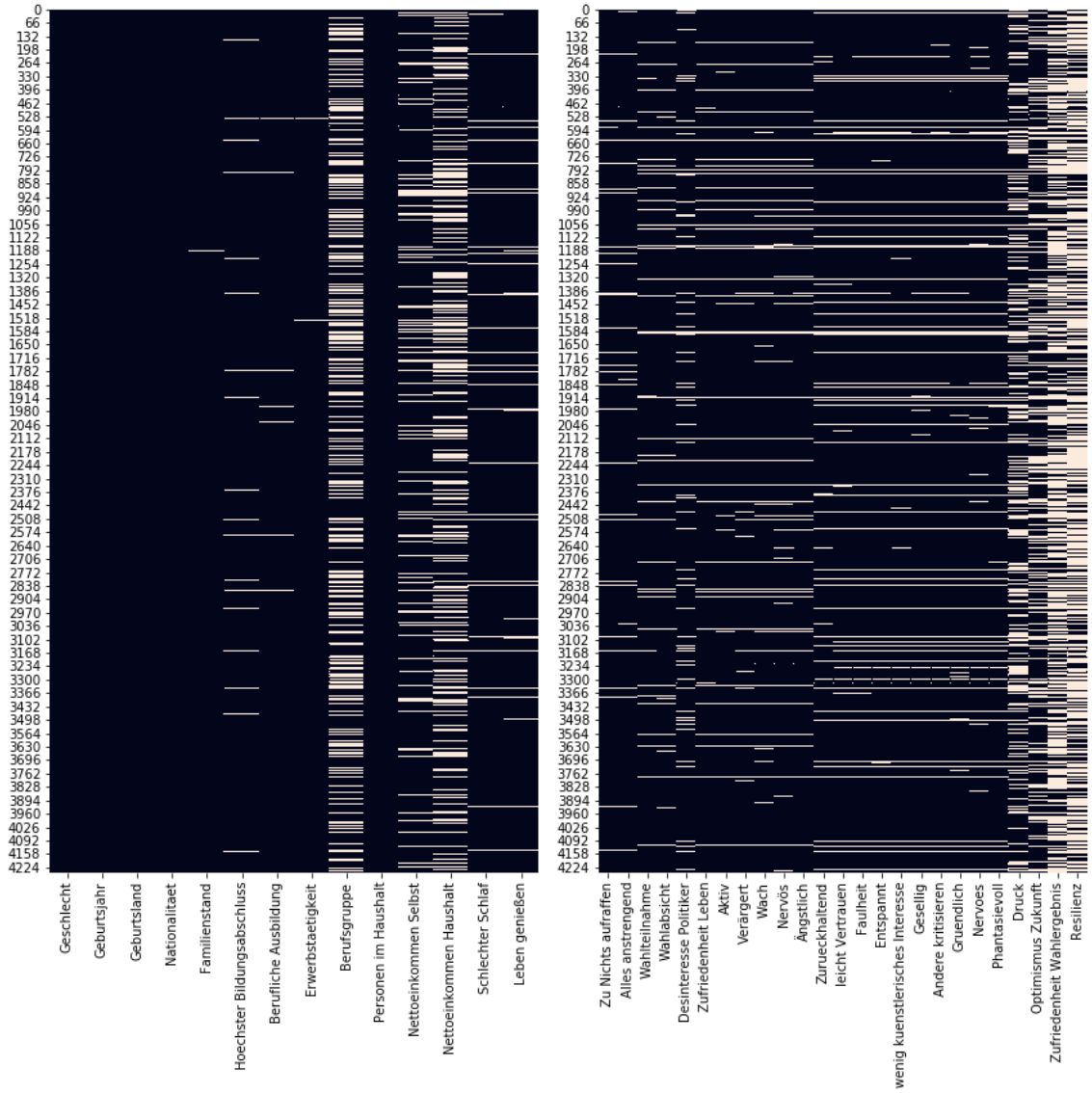


Figure 2.2: Missing values in GESIS. The attribute values of a participant are always known for "Geschlecht", "Geburtsland", "Geburtsjahr", "Nationalitaet", "Familienstand", "Personen im Haushalt". In contrast, the last three columns "Druck", "Optimismus Zukunft", "Zufriedenheit Wahlergebnisse" and "Resilienz" are almost always missing and are therefore removed from the analysis. Participants with missing BFI-10 elements are removed. Sample size will only be reduced slightly as missing values often occur for the same instance. These dependencies form a line pattern in the graph. "Berufsgruppe" was surveyed as a text field so that the column clearly suffers from ambiguous value mismatch. To include "Berufsgruppe" mappings need to be redefined first. For now, "Berufsgruppe" is removed.

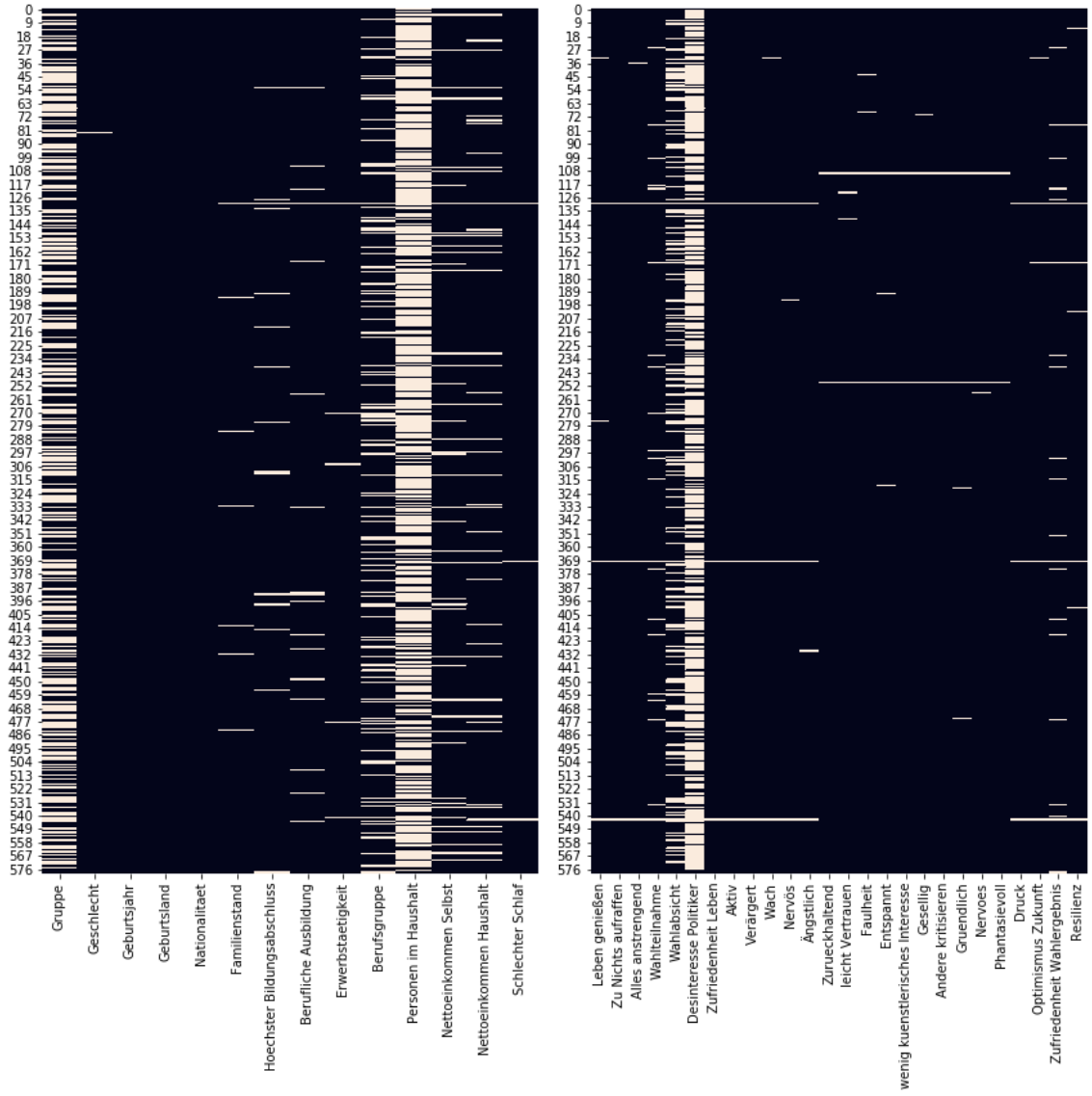


Figure 2.3: Missing values in GBS. There is one more attribute in "Gruppe". Not every participant received a positive a negative psychological treatment. Therefore, "Gruppe" is more likely to be missing than not. However, the absence of a value indicates no treatment rather than a missing positive or negative. "Gruppe" is not properly represented yet. "Desinteresse Politiker" is given by multiple data sources from different excel files. Some of them being the inverse of the attribute itself. The surey design regarding this issue is unclear to me. To incorporate "Desinteresse Politiker" the attribute(s) need to be imported correctly, if possible. Another import issue is given by "Personen im Haushalt". If the actual value is greater than one, the cell will be empty. To correct this, the corresponding csv-file needs to be fixed. The text field "Berufsgruppe" suffers on both ends, GBS and GESIS, due to current oversimplification of value and potential data mismatch.

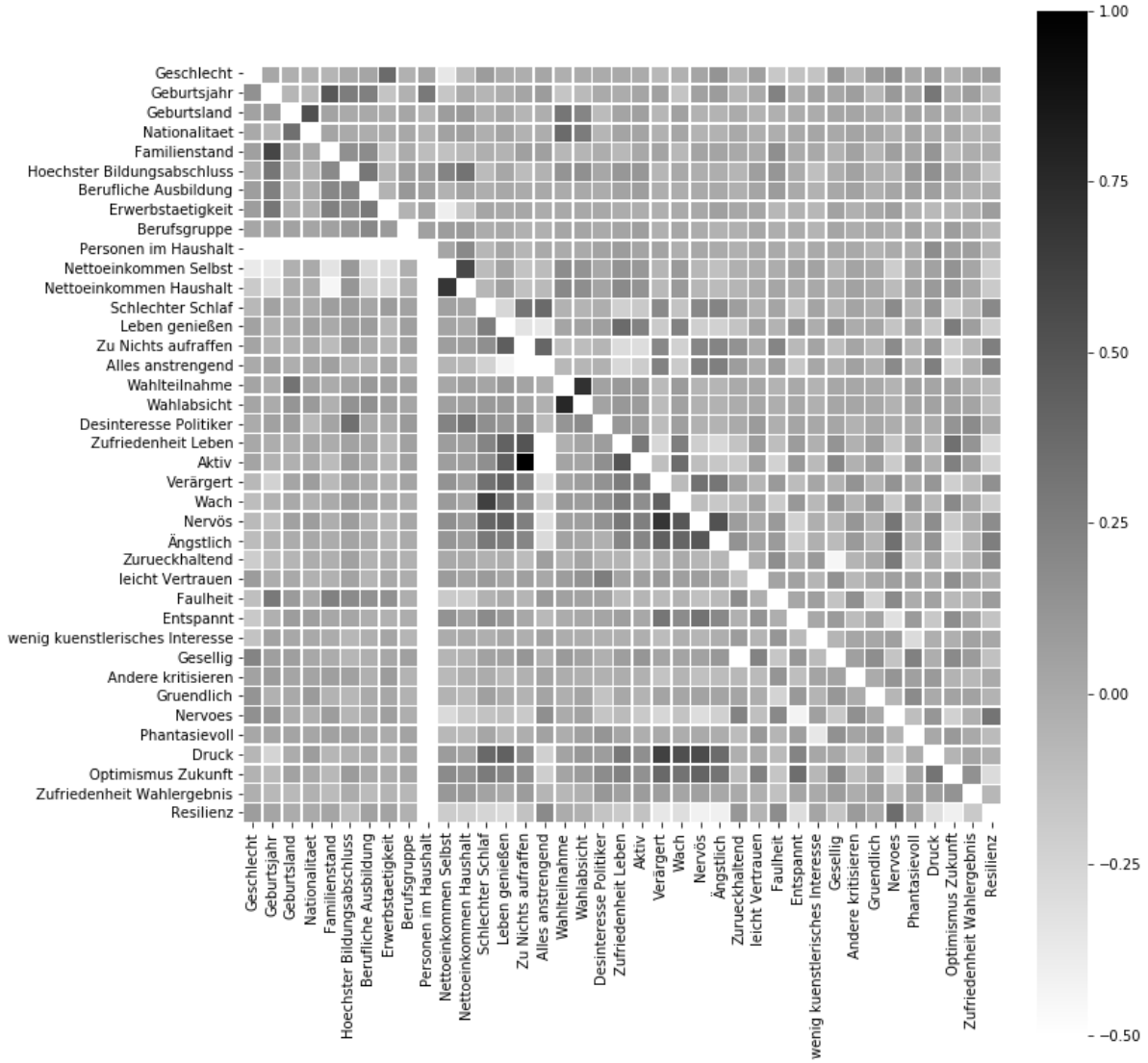


Figure 2.4: The upper right triangular matrix shows GESIS correlations while GBS correlations are shown in the lower left. The main diagonal should not be confused with white squares. These trivial combinations are simply excluded and not colored black. As can be seen "Personen im Haushalt" in GBS can not be calculated, since there is only one possible value. "Nettoeinkommen Selbst" and "Nettoeinkommen Haushalt" are highly correlated but not removed or handled at all. I will keep this in mind, when facing the naive bayes assumption in the learning process. Entropy-based mutual information in "Wahlteilnahme" and "Wahlabsicht" have led to almost perfect classification performances in predicting political participation. "Wahlabsicht" is therefore removed.

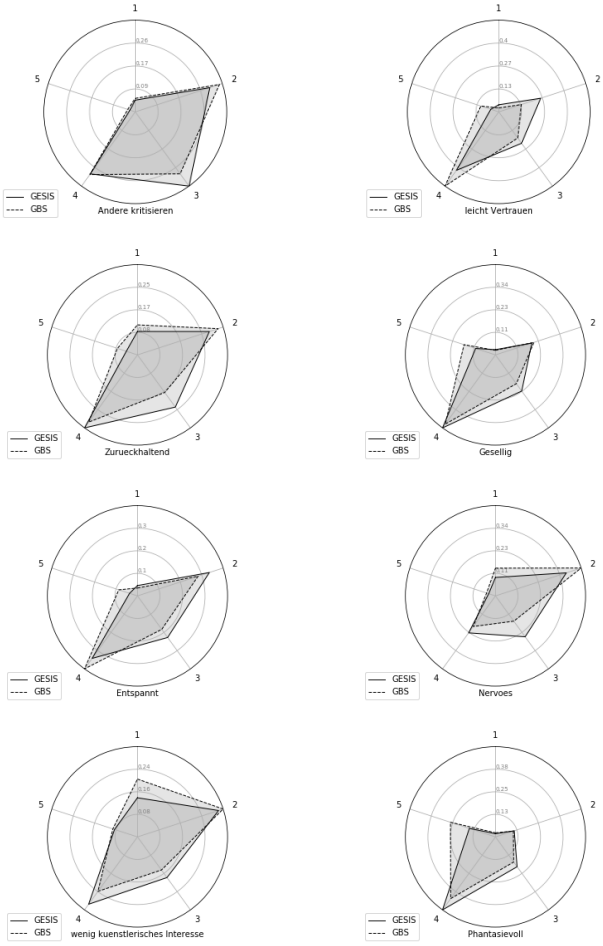


Figure 2.6: Agreeableness is the measure of one's cooperation, empathy, and willingness to trust and help others. Openness estimates whether one is hesitant or eager about new objects or situations. Extraversion refers to level of sociability, seeking and enjoyment of social contact, and energy and assertiveness in social situations. Neuroticism is characterized by easily experiencing negative emotions, and a poor coping response to those emotions. The cyclic structure of the chart, i.e. "Strongly Disagree" next to Strongly Agree", provides a vivid example for the central-tendency bias across all items.

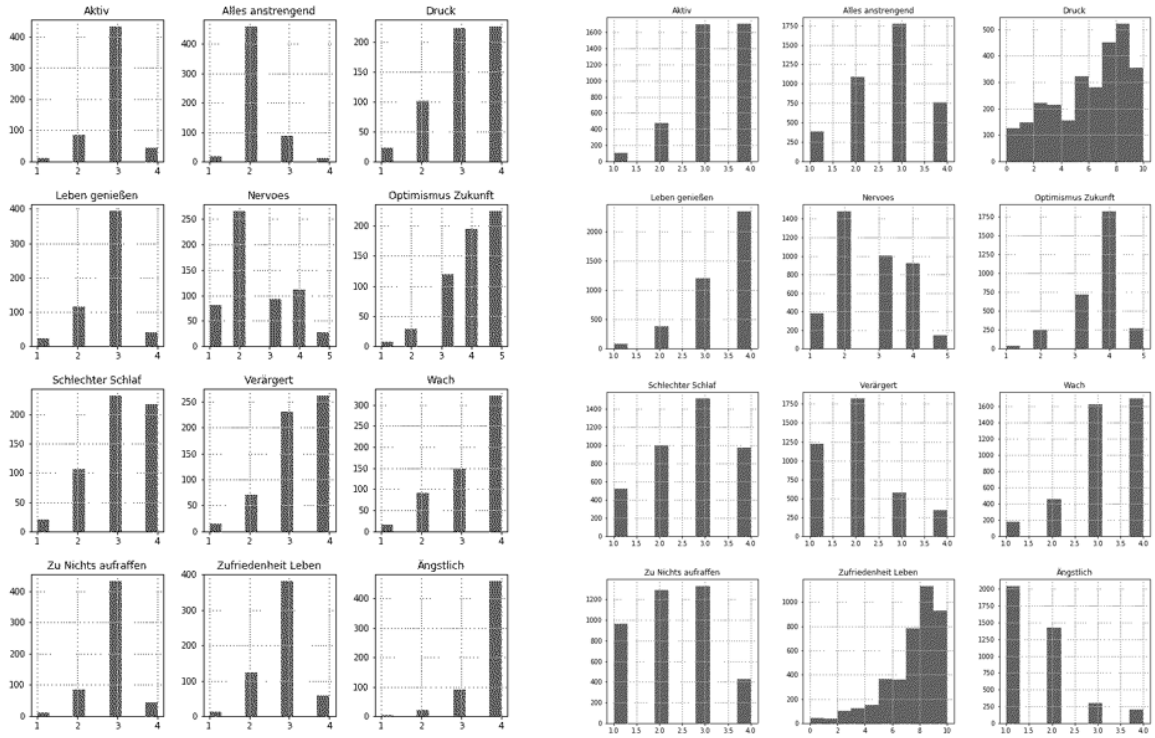


Figure 2.7: Overview of attributes that can not be used for further analysis due to data mismatch. GESIS histograms (right) are show differently scaled values, some of which have been transformed already. Taking the inverse of "Aengstlich", "Alles anstrengend" and perhaps "Verargert" for either the left or right side is likely to match the values properly. The remaining attributes do not seem to be reliable and will not be included for the time being.

3 FEASIBILITY OF LEARNING

No practical amount of data can distinguish between two distributions, thus instances of GBS can not be proven to come from GESIS. However, machine learning allows to infer the conditional probability of *'GBS participant is representative'* given the survey data within a probabilistic framework. A well-defined learning problem involves a number of design choices, including selecting the target function to be learned, a representation for this target function, and an algorithm to learn from the source of training experience [2]. This chapter covers the theoretical aspects of the MRS procedure introducing positive-unlabeled learning in addition to traditional machine learning.

3.1 Terminology and Definitions

Key terminology and destinctions made in the field of survey analytics are defined in this section. When learning from biased data, basic design issues and approaches to discriminative learning must be adjusted. The role of noise in the bias-variance decomposition will be analyzed and further broken down. Discriminative learning then uses techniques to address high variance.

3.1.1 Sampling Bias

Sampling bias is often referred to as selection bias or sample selection bias. I will stick to the more descriptive term sampling bias. It underlines the fact that the bias arises in how the data was sampled. Also, the use of the term becomes less ambiguous, because there exists another notion of selection bias in the context of model selection. This type of bias is usually referred to as bad generalization, where the performance of the selected hypothesis is overly optimistic. [input: convenience sampling]

Although one could employ a census to measure the entire population, it is more common to take a sample of the population. A properly designed probability sample (see probability sampling) can be used to make estimates for not only the sample itself, but also for the underlying population from which it was selected. A probability sample is one in which each element of the (underlying) population has a known and non-zero chance of being

selected. That is, every person has a chance to be included in the study and have his or her characteristics, opinions, etc., become part of the data. It should be noted that everyone does not have to have an equal chance of being selected just a known non-zero chance of being selected. Probability samples have several desirable characteristics. They enable us to put a margin of error or confidence interval on our estimates essentially a measure of how accurate the estimate is compared to the same estimate calculated on the full population. Probability samples make it possible to not only compare the sample to the population, but also to compare a sample from one population to a sample from another population,

3.1.2 Representative Sample

Variables considered in the study must accurately reflect the populations characteristics. Some examples include sex, age, education level, socioeconomic status or marital status. Consider a randomly chosen survey participant, i.e. an instance of GBS or GESIS. It is usually difficult to draw a simple random sample from the population, due to cost and practical considerations such as no comprehensive sampling frame available.

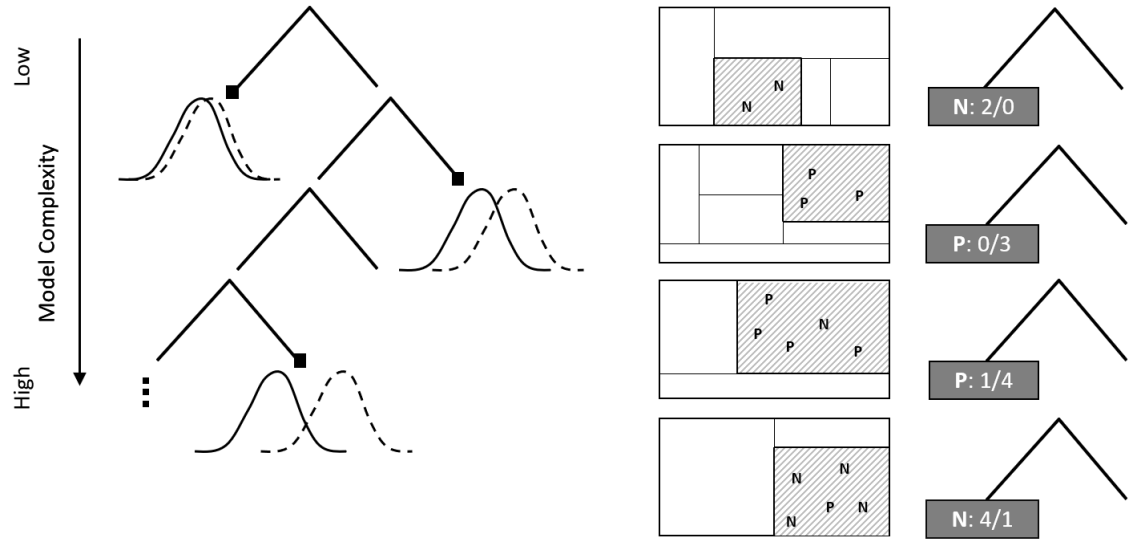


Figure 3.1: .

3.1.3 The Problem of Overfitting

Overtting stands out as one of the biggest challenges for machine learning. It is not exclusive to machine learning but rather a fundamental problem across science and is at the

very heart of the dangers of statistical inference. Hypothesis h in H overts training data if there exists an alternative hypothesis $\bar{h}$ in H such that $errorTrain(h) < errorTrain(\bar{h})$ and $errorD(h) > errorD(\bar{h})$, where $errorTrain(h) :=$ error of hypothesis h over training data and $errorD(h) :=$ error over entire distribution D of data. A hypothesis overfits the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances [1].

When overfitting occurs, the learned hypothesis is very good at calculating the answers for the given data, but much less so for new instances it encounters. In a sense, the machine learning algorithm has become xated on unimportant features of the training data overlooking the big picture. It often occurs when the algorithm has too many options to play with in designing its mappings, an approximation of the target function. The freedom to tune parameters and add complexity until it exactly matches the training data, rather than looking for large, systematic patterns leads to high variance [6].

The expected prediction error at any given data point x_0, the generalization error, can be decomposed as follows [5]:

$$\begin{aligned} Err(x_o) &= \mathbb{E}[(Y - \bar{f}(x_0))^2 \mid X = x_0] \\ &= \sigma_\epsilon^2 + [\mathbb{E}\,\bar{f}(x_0) - f(x_0)]^2 + \mathbb{E}[\bar{f}(x_0) - \mathbb{E}\,\bar{f}(x_0)]^2 \\ &= \sigma_\epsilon^2 + Bias^2(\bar{f}(x_0)) + Var(\bar{f}(x_0)) \\ &= IrreducibleError + Bias^2 + Variance \end{aligned}$$

The bias and variance terms make up the error of $\bar{f}(x_0)$ in estimating $f(x_0)$. The bias component is the squared difference between the true mean $f(x_0)$ and the expected value of the estimate $[\mathbb{E}\,\bar{f}(x_0) - f(x_0)]^2$, where the expectation averages the randomness in the training data. The variance term refers to the amount by which the estimate of the target function would change if it was estimated using a different training data set. Ideally the estimate for the underlying pattern should not vary too much between training sets. More generally, as the model complexity is increased, the variance tends to increase and the squared bias tends to decrease. The opposite behavior occurs as the model complexity is decreased. The first term in this expression is the irreducible error, a combination of stochastic and deterministic noise. More precisely, stochastic noise are fluctuations or measurement errors that can not be modelled. Re-measuring y_n changes this component. Deterministic noise is the part of the target function that can not be represented. Changing H changes this component. With a single dataset D and fixed H it is impossible to distinguish [6].

3.2 Discriminative Learning

The intuition for these results comesfrom the fact that in many practical situations, the posteriordistributions in traditional and non-traditional setting pro-vide the same optimal ranking of data points on a given testsample (Jain et al. 2016; Jain, White, and Radivojac 2016).

The holdout estimate can be made more reliable by repeating the process with different subsamples. The error rates on the different iterations are averaged to yield an overall error rate. To further reduce the variance of the error estimate, each class is sampled with approximately equal proportions in both datasets, a technique called stratication.

3.3 Learning from Positive and Unlabeled Data

We have shown the eect of resampling contaminated sets and provided some basic insight into the mechanics of bagging. We will now link these two elements to justify bagging approaches in the context of contaminated training sets. Its usefulness can be considered by both the variance reduction argument of Bauer and Kohavi [4] and equalizing the inuence of training points as described by Grandvalet [24]. Variance reduction. Resampling a contaminated set yields dierent levels of contamination in the resamples as explained in Section 3.1. Varying the contamination between base model training sets induces variability between base models without increasing bias. This observation enables us to create a diverse set of base models by resampling both P and U. The variance reduction of bagging is an excellent mechanism to exploit the variability of base models based on resampling [4, 10]. In the context of RESVM, a tradeo takes place between increased variability (by training on smaller resamples, see Figure 1) and base models with increased stability (larger training sets for the SVM models). (https://arxiv.org/pdf/1402.3144.pdf)

PU learning is a semi-supervised technique that does not make the simplifying assumption of GBS instances being negative. Instead, a one-class classifier is trained on GESIS only. [...] This can result in even better assessment. [Read Literature] - Imporance weighted cross validation and pu learning with proper assessment. State-of-the-art techniques in positive-unlabeled learningtackle this problem by treating the unlabeled sample as neg-atives and training a classier to distinguish between la-beled (positive) and unlabeled examples. Surprisingly,for a variety of performance criteria, non-traditional classi-ers achieve similar

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	0.500	0.130	GBS
	1.000	1.000	0.870	1.000	0.931	0.500	0.870	GESIS
W. Avg.	0.870	0.870	?	0.870	?	?	0.500	0.774

Table 3.1: Some descriptive statistics of location and dispersion for 2100 observed swap rates for the period from February 15, 1999 to March 2, 2007. Swap rates measured as 3.12 (instead of 0.0312). See Table **??** in the appendix for more details.

performance under traditional evalua-tion as optimal traditional classiers (Blanchard et al. 2010;Menon et al. 2015).

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

3.3.1 Recovering Model Performance

The most extensively studied and widely used performance evaluation in binary classication involves estimating the Receiver Operating Characteristic (ROC) curve. The ROC curve

plots the true positive rate (recall) of a classier as a function of its false positive rate (Fawcett 2006) over a range of decision thresholds. Furthermore, AUC has a meaningful probabilistic interpretation that is used to the ability of the classier to separate classesand is often used to rank classiers (Hanley and McNeil1982). the widely-accepted evaluation approaches us-ing ROC curves are insensitive to the variation ofraw prediction scores unless they affect the ranking.

Such perfor-mance estimation often involves computing the fraction(s)of correctly and incorrectly classied examples from bothclasses; however, in absence of labeled negatives, the frac-tions computed under the non-traditional evaluation are in-correct, resulting in biased estimates. Figure 1 illustratesthe effect of this bias by showing the traditional and non-traditional ROC curves on a handmade data set. Becausesome of the unlabeled examples in the training set are infact positive, the area under the ROC curve estimated whenthe unlabeled examples were considered negative (non-traditional setting) underestimates the true performance forpositive versus negative classication (traditional setting).This paper formalizes and evaluates performance estima-tion of a non-traditional classier in the traditional settingwhen the only available training data are (possibly noisy)positive examples and unlabeled data. hough the efcacy of non-traditional classiers has beenthoroughly studied (Peng et al. 2003; Elkan and Noto 2008;Ward et al. 2009; Menon et al. 2015), estimating their true performance has been much less explored. be recov-ered with the knowledge of class priors. results in biased empirical estimates of the classier performance

The ROC curve provides in-sight into trade-offs between the classiers accuracies onpositive versus negative examples over a range of decisionthresholds. Although model learning and performance evaluation in asupervised setting are well understood (Hastie et al. 2001),the availability of unlabeled data gives additional optionsand also presents new challenges. A typical semi-supervisedscenario involves the availability of positive, negative and(large quantities of) unlabeled data. Here, the unlabeled datacan be used to improve training (Blum and Mitchell 1998) orunbias the labeled data (Cortes et al. 2008); e.g., to estimateclass proportions that are necessary to calibrate the modeland accurately estimate precision when class balances (butnot class-conditional distributions) in labeled data are notrepresentative (Saerens et al. 2002). This is often the casewhen it is more expensive or difcult to label examples ofone class than the examples of the other.

Let f be the true distribution over the input space X from which unlabeled data is drawn. With distributions f1 and f0 of the positive and negative examples, respectively, it follows

that

$$f(x) = \alpha f_1(x) + (1 - \alpha) f_0(x)$$

with positive class prior $\alpha \in [0, 1], x \in X$.

Consider the binary classification problem from input $x \in X$ (BFI-10 and BRS data) to output $y \in Y$ (representative: '1', not representative: '0'). The learning objective is to discriminate between X_p drawn according to f_1 and X_u drawn according to f and recover its performance estimate in the traditional setting, i.e. evaluating the decision boundary between positive and negative data.

Recall γ, false positive rate η and precision ρ are defined as: $\gamma = P[\hat{Y} = 1|Y = 1]$, $\eta = P[\hat{Y} = 1|Y = 0]$ and $\rho = P[Y = 1|\hat{Y} = 1]$, where $\hat{Y}$ is an estimate of the true class label Y. TPR γ can be estimated directly, because X_p was sampled from f_1, while this does not hold true for η given the absence of samples from f_0.

$$\gamma = \mathbb{E}\left[f_1[h(x)]\right] = \frac{1}{|X_p|} \sum\nolimits_{x \in X_p} h(x)$$
$$\hat{\eta}^{pu} = \mathbb{E}\left[f[h(x)]\right] = \frac{1}{|X|} \sum\nolimits_{x \in X} h(x)$$

The area under ROC curves $AUROC^{pu}$ so far could only be estimated for the positive versus unlabeled classification by plotting γ and $\hat{\eta}^{pu}$. To calculate AUC from AUC^{pu}, S. Jain et al. (2015) express η in terms of $\hat{\eta}^{pu}$ and $alpha$ and provide a full derivation from the probabilistic definition of the AUC with

$$\eta = \frac{\hat{\eta}^{pu} - \alpha\gamma}{1 - \alpha}$$

so that

$$AUC = \frac{AUC^{pu} - \frac{\alpha}{2}}{1 - \alpha}$$

proving

$$AUC > AUC^{pu} \iff AUC^{pu} > \frac{1}{2}$$

4 RESULTS

4.1 Maximal Representative Subsample

There is almost always not enough data available to partition it into separate training and test sets without losing signicant modelling or testing capability. In these cases, a fair way to properly estimate model prediction performance is to use cross-validation as a powerful general technique[5].

The overly optimistic resubstitution error, is not a good indicator of model performance. To evaluate the actual performance of a model, the given data samples need to be split. The proper procedure uses three sets: training data, validation data, and test data [2]. The holdout method is the most common approach to get a reliable performance estimation: A certain amount of data is reserved for testing while the remainder is used for the actual training. Because the method is very fast, it is useful to use when the algorithm is slow to train and the dataset is large. Training and test sets might not be representative of the same underlying distribution, e.g. class hardly represented in the test set.

The holdout estimate can be made more reliable by repeating the process with different subsamples. The error rates on the different iterations are averaged to yield an overall error rate. To further reduce the variance of the error estimate, each class is sampled with approximately equal proportions in both datasets, a technique called stratication. Figure X shows the results on the GFI-10 data.

4.1.1 Fraction of Positives

Estimating Positive Class Prior with One-Class SVMs. Using the One-Class SVM and its ability to capture the shape of the data set, hence performing better when the data is strongly non-Gaussian, i.e. with two well-separated clusters;

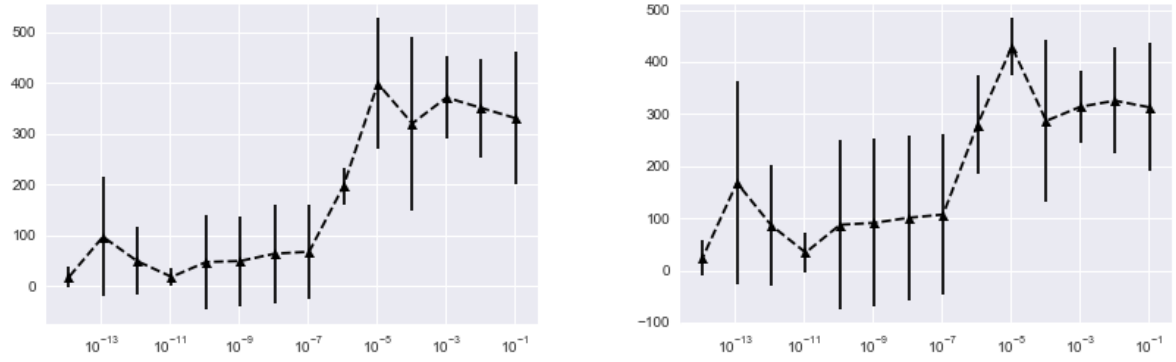
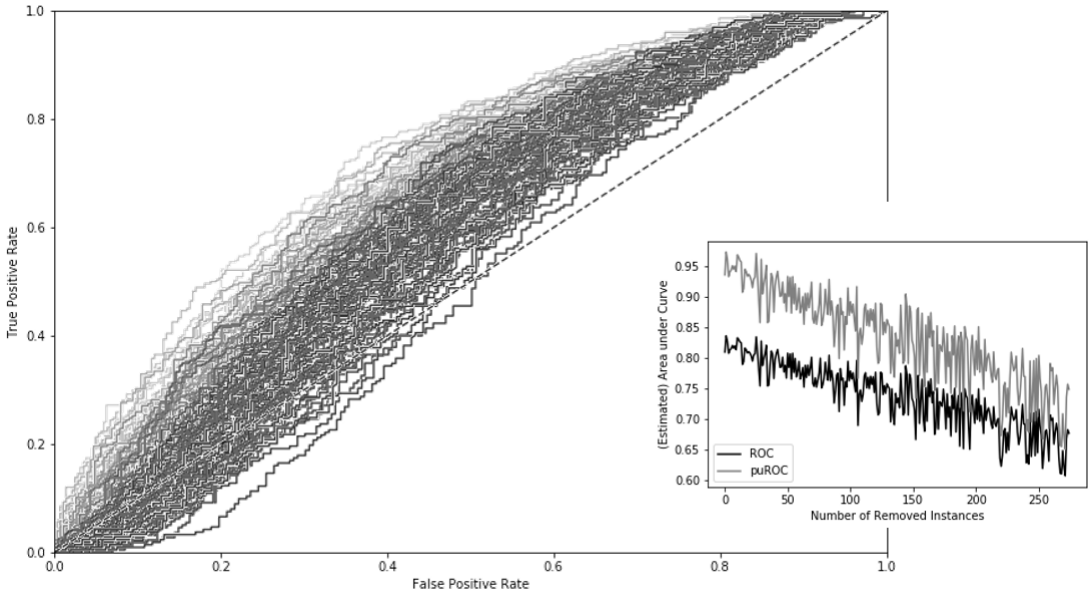


Figure 4.1: Tuning parameter nu that controls the trade-off between the fraction of non-representative samples and the number of support vectors in one-class SVM. More than 0.73 of GBS (right) are classified as representative with high confidence (low sdt) for the optimal value $nu = 10^{-5}$.

4.1.2 ROC and puROC Evaluation

4.2 Political Participation and Resilience

In modelling a political participation process, a computer program is designed to approximate the likelihood of a person going to vote on election day. Ideally, for every instance with unknown political interest and willingness to participate, there is enough data of people of similiar demographics, socioeconomics and psychological traits to generalize from.


True Positive Rate
False Positive Rate
(Estimated) Area under Curve
Number of Removed Instances
ROC
puROC

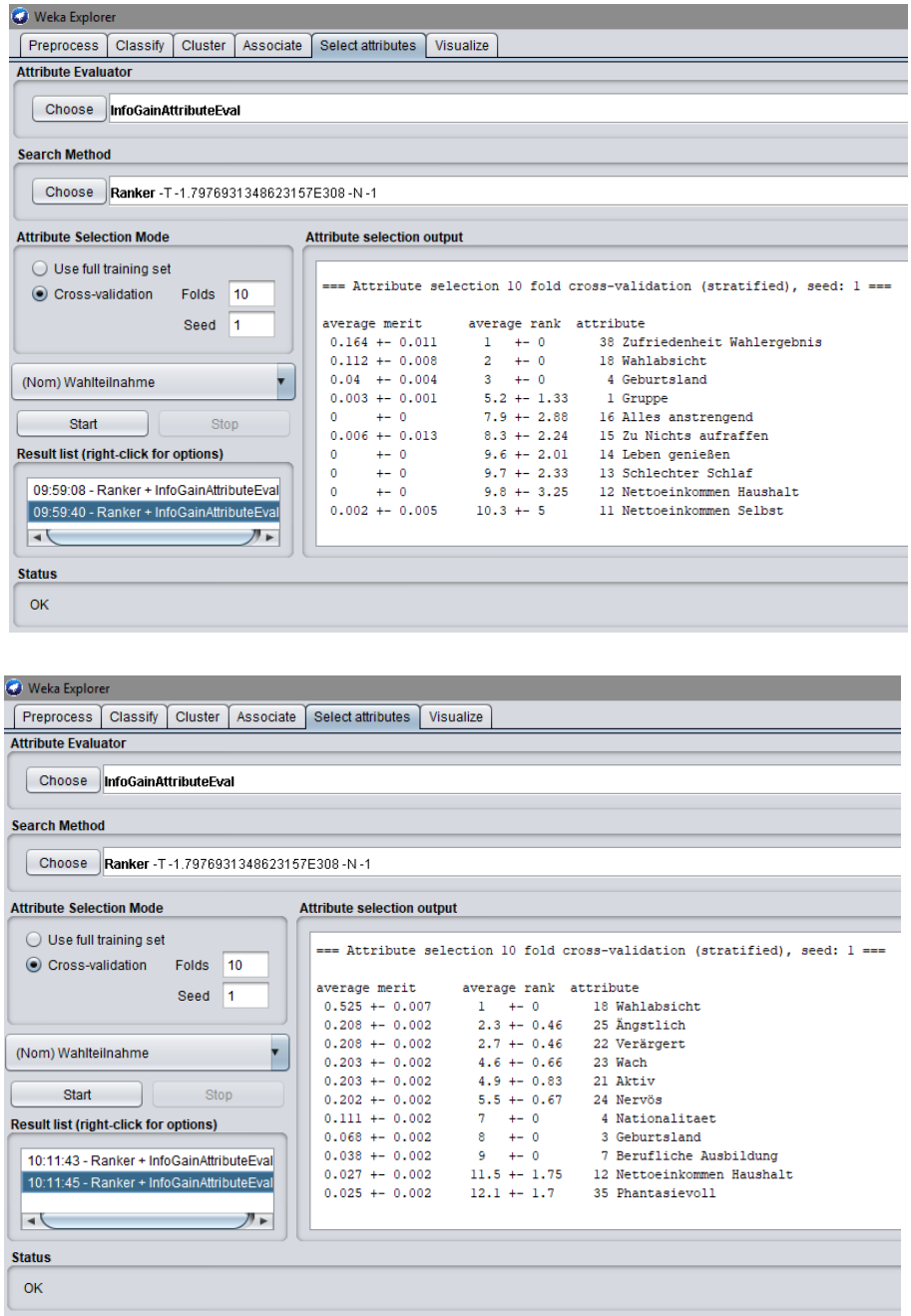


Figure 4.2: Feature importance GBS (n=579) and GBS MRS (n=280) in classification of political participation "Wahlteilnahme".

5 FUTURE WORK

Many dierent adaptations, statistics, and experiments have been left for the future due to lack of time, i.e. data matching and transformation with real data have been very time consuming. Controlled environments are needed to observe the behavior of the proposed algorithm.

For one thing, future work concerns deeper analysis of the proposed sampling method, in particular, experiments on synthezised data. The Synthetic Minority Over-sampling TEchnique (SMOTE [9]) is a very popular oversampling method that creates synthetic minority class instances. The SMOTE instances are linear combinations of two similar instances from the minority class (x and x^R) and are defined as: $s = x + u(x^R - x)$ with $0 \geq u \geq 1$. x^R is randomly chosen among the k nearest neighbors of x belonging to the minority class. SMOTE can be used to validate the MRS procedure by simulating the problem at hand.

Specific regions in the feature space are first over-sampled using SMOTE, before they are under-sampled by MRS. Experiments with multiple such synthesized data, with oversampling ratio ranging from high to low, might support the proposed procedure with greater evidence. The initial data sets are then compared to the result sets. GESIS is particularly well suited to artificially recreate the initial problem as visualized in Figure 5.1. It is only necessary to try to avoid giving the synthesized data properties that makes it possible for a learning algorithm to distinguish synthesized from non-synthesized example. This mechanism would for instance aid to compare classification results more easily.

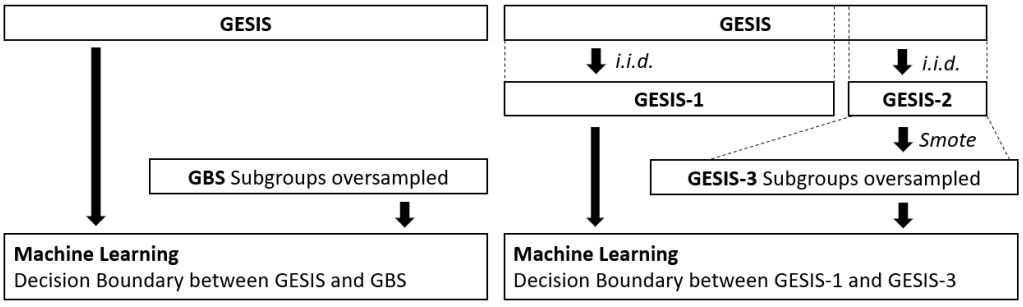


Figure 5.1: Artificial data synthesis to overrepresent subgroups of GESIS. True negatives are removed from the MRS with positive classes GESIS (left) and GESIS-1 (right). Oversampled instances can easily be marked as such for result set comparisons.

The main criteria considered in this work has been the area under the ROC curve. Having a single-number evaluation metric speeds decision-making when selecting among non-representative instances. It gives a clear preference ranking among all of them, and therefore a clear direction for progress. To enable a basis for a more informed exclusion of instances, another important performance criterion generally used in information retrieval could be added. The F-Measure, including summary statistics derived from the precision-recall curve, may be preferred to ROC curves when classes are heavily skewed (Davis and Goadrich 2006). Precision and recall have been estimated in section 3.3.1 in a positive-unlabeled setting. The area under precision-recall curves $AUPR$ can be expressed using the approximated value for the fraction of positives α in X_u: $\rho = \frac{\alpha\gamma}{\hat{\eta}^{pu}}$

6 CONCLUSION

Imbalanced setting where the minority class GBS is the unlabeled class of interest but handled as negative. evaluation in the absence of actual negatives is further enhanced by class imbalance. puF-Measure instead of (or in addition to) puROC could reduce the effects of class imbalance. inclusion of f1-measure would be interesting. finally, the label "representative" is actually a property of a sample and not its single instances. survey mismatch from gbs to gesis renders most attributes useless regarding entropy. forces valu comparisons essentially introduce non existent pattern. one class classification suffers from high variance in estimating the fraction of actually representative gbs data. assessment of binary classifiers in pu settings does not lead to more accurate roc curve estimations as the fraction of positives cannot be estimated. overrepresentative underrep. rep and nonrep. trhoughout the thesis are not properly defined.

BIBLIOGRAPHY

Breusch, T. S. and P. Schmidt (1988): "Alternative Forms of the Wald test: How Long is a Piece of String," *Communications in Statistics, Theory and Methods*, 17, 2789–2795.

Gallant, A. R. (1987): *Nonlinear Statistical Models*, New York: John Wiley & Sons.

DECLARATION OF AUTHORSHIP

I hereby confirm that I have authored this Bachelor's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Mainz, October 15, 2018

Laksan Nathan