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 1

1.1 Related Work 2

1.2 Outline 3

2 Terminology and Definitions 5

2.1 Sampling Bias 5

2.2 The Problem of Overfitting 5

2.2.1 Model and Hypothesis 5

2.2.2 Bias and Variance Tradeoff 6

3 Initial Data Analysis 9

3.1 Political Participation and Resilience 10

3.1.1 The Big Five Dimensions 10

3.1.2 Demographic and Personal Data 13

3.2 Survey Mismatch 13

3.2.1 Likert-Type Scale 13

3.2.2 Imbalanced Data 14

4 Feasibility of Learning 17

4.1 Learning from Positive and Unlabeled Data 19

4.1.1 Recovering Model Performance 22

4.2 Artificial Data Synthesis 25

4.3 Synthetic Minority Oversampling Technique 25

5 Results 27

5.1 One-Class Classification 27

5.2 Traditional Machine Learning 27

5.3 Positive Unlabeled Learning 27

5.3.1 Fraction of Representatives 27

5.4 Maximal Representative Subsample 29

5.5 Predicting Political Participation 30

6 Conclusion **31**

Bibliography **32**

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. Data sets are rarely generated under ideal conditions with bias pervasive in almost all empirical studies.

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 distributions 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.

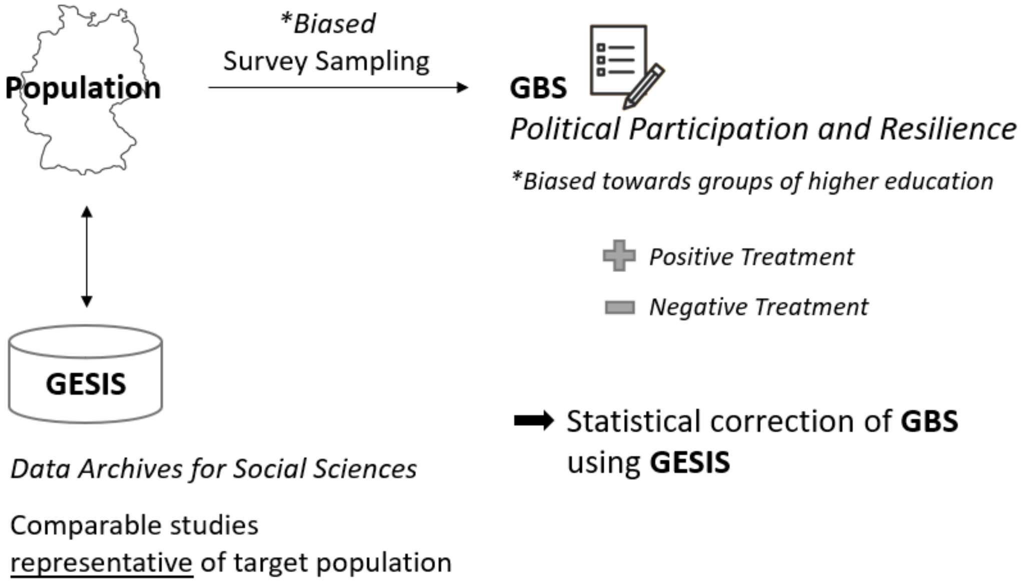

Population
*Biased
Survey Sampling
GBS
Political Participation and Resilience
*Biased towards groups of higher education
Positive Treatment
Negative Treatment
GESIS
Data Archives for Social Sciences
Comparable studies
representative of target population
Statistical correction of GBS using GESIS

Figure 1.1: 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.

Depending on the definition of MRS, there are two possible ways to tackle this problem:

1. Search algorithm with objective scoring function.
2. Try to avoid giving the synthesized data properties that makes it possible for a learning algorithm to distinguish synthesized from non-synthesized example such as if all the synthesized data comes from one of 20 car designs, or all the synthesized audio comes from only 1 hour of car noise. This advice can be hard to follow.

1.1 Related Work

Several approaches exist to deal with positive and unlabeled data. The most straightforward one is to assume that all the unlabeled data are negative and simply apply standard machine learning techniques (Neelakantan, Roth, and McCallum 2015). A second approach is to select some of the unlabeled examples that are very different from the positively labeled ones and label them as negative. A classier is then learned using the given positive examples and inferred negative examples (Liu et al. 2002; Li and Liu 2003; Yu, Han, and Chang 2004; Yu 2005; Li et al. 2009; Nguyen, Li, and Ng 2011). A third approach is to employ an evaluation metric that only uses positive, or positive and unlabeled data (Muggleton 1996; Lee and Liu 2003; Claesen et al. 2015a). Using this metric, one can tune for the best class weights or regularization settings (Lee and Liu 2003; Liu et al. 2005; Mordelet and Vert 2014; Claesen et al. 2015c). A nal approach is to explicitly consider the class prior. It can be used to either adapt algorithms to incorporate this information during learning (Denis 1998; Liu et al. 2003; Zhang and Lee 2005; Denis, Gilleron, and Letouzey 2005; Elkan and Noto 2008) or as a preprocessing step to assign weights to the unlabeled examples (Elkan and Noto 2008). Because the class prior is often not known, several methods were proposed in the last decade to estimate it from the positive and unlabeled data (Elkan and Noto 2008; du Plessis and Sugiyama 2014; du Plessis, Niu, and Sugiyama 2015; Jain, White, and Radivojac 2016; Jain et al. 2016; Ramaswamy, Scott, and Tewari 2016)

Both theoretical and practical results show that the out-of sample error increases proportionally to the distribution shift [2, 20]. To compensate for the degradation in performance, many techniques have been designed to reduce the effects of covariate shift.

1.2 Outline

I will frequently refer to decision tree theory. You will need a basic understanding of what they are to follow this text. Decision trees have performaned slightly outperformed comparable discriminative.

The remainder of this thesis is organized as follows. Section 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. Section 3 discusses the feasibility of learning in a binary classification setting. To estimate model performance in the absence of labeled negatives, standard evaluation metrics are adapted using an initial ranking model. Discriminative ensemble models are trained on positive and unlabaled data in Section 4 to label instances as representative that cannot be distinguished from the out-of-sample distribution. The resulting maximal representative subset of GBS is presented in Section 5 and compared to GBS and GESIS regarding political participation and resilience. Related work is discussed in Section 6 and Section 7 concludes.

2 TERMINOLOGY AND DEFINITIONS

A well-defined learning problem involves a number of design choices, including selecting the type of training experience, the target function to be learned, a representation for this target function, and an algorithm to learn from the source of training experience [2]. 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 perticipate, there is enough data of people of similiar demographics, socioeconomic status and psychological resilience to generalize from. This chapter defines key terminology and destinctions when learning from biased data. Basic design issues and approaches to supervised learning are covered, while conceptual elements of interest are introduced with regards to overfitting. The role of noise in the bias-variance decomposition will be analyzed and further broken down.

2.1 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.

2.2 The Problem of Overfitting

2.2.1 Model and Hypothesis

The model in supervised learning usually refers to the mathematical structure of how to make predictions y_i given x_i. The most common model is a linear regression model, where the prediction is given by a linear combination of weighted input features. The parameters, the weights of these features, are the undetermined part that need to be learned from data. Depending on the task, the prediction value can have different interpretations, i.e. regression or classification. The categorical outcome, ”did vote” or ”did not vote”, makes political participation a binary classification problem [7]. In machine learning, the terms hypothesis and model are often used interchangeably. This paper uses the following convention [3] as the terminology to describe ideas and concepts is not standardized:

- The phrase single hypothesis refers to a single probability distribution or function. An example is the polynomial $2x^2 + 3x + 1$.
- The word model refers to a set of probability distributions or family of functions with the same functional form. An example is the set of all quadratic functions.
- As a generic term, hypothesis refers to both single hypotheses and models.

With the definitions above, it is a hypothesis selection problem if both the degree of a polynomial and the corresponding parameters are of interest. The phrase single hypothesis refers to a single probability distribution or function. A machine learning model, the composite hypothesis, refers to a family of probability distributions or functions with the same functional form. An example is the set of all second-degree polynomials [9].

2.2.2 Bias and Variance Tradeoff

”Overfitting is the disease. Noise is the cause. Learning is led astray by fitting the noise more than the actual signal”[6]. To avoid overfitting you might deliberately exclude certain factors, increase sample size, stop the analysis early, or simply pick less complex algorithms. Regularization puts a break where additional iterations of algorithms start to harm the performance. Validation is another way to see what will actually happen out-of-sample

The minimum description length (MDL) principle, informally applied to all the sciences, is a formalization of Occam’s Razor in which the best hypothesis for a given set of data is the one that leads to the best compression of the data. MDL procedures automatically and inherently protect against overfitting and can be used to estimate both the parameters and the structure of a model. To decide among competing explanation of data given limited observations, the lesser complex model should always be preferred. Any regularity in data can be used to compress it, i.e. fewer symbols needed to describe the data. Equating learning with finding regularities implies that a model can compress data more, the more it learns the underlying patterns. Because many algorithms add an MDL function of the model to the objective function of the training, MDL methods can be interpreted as searching for a model with good predictive performance on unseen data [4].

By definition, statistical inference is taking the results of applying some sort of construct or model to specific data and then speculating that it would continue to perform well beyond the original observation range. Given a set of training samples (x_i, y_i) find a single hypothesis h that "fits the data well": $y_i = h(x_i)$ for most i. The equation is characterized by a trade-off between goodness-of-fit and complexity of the hypothesis:

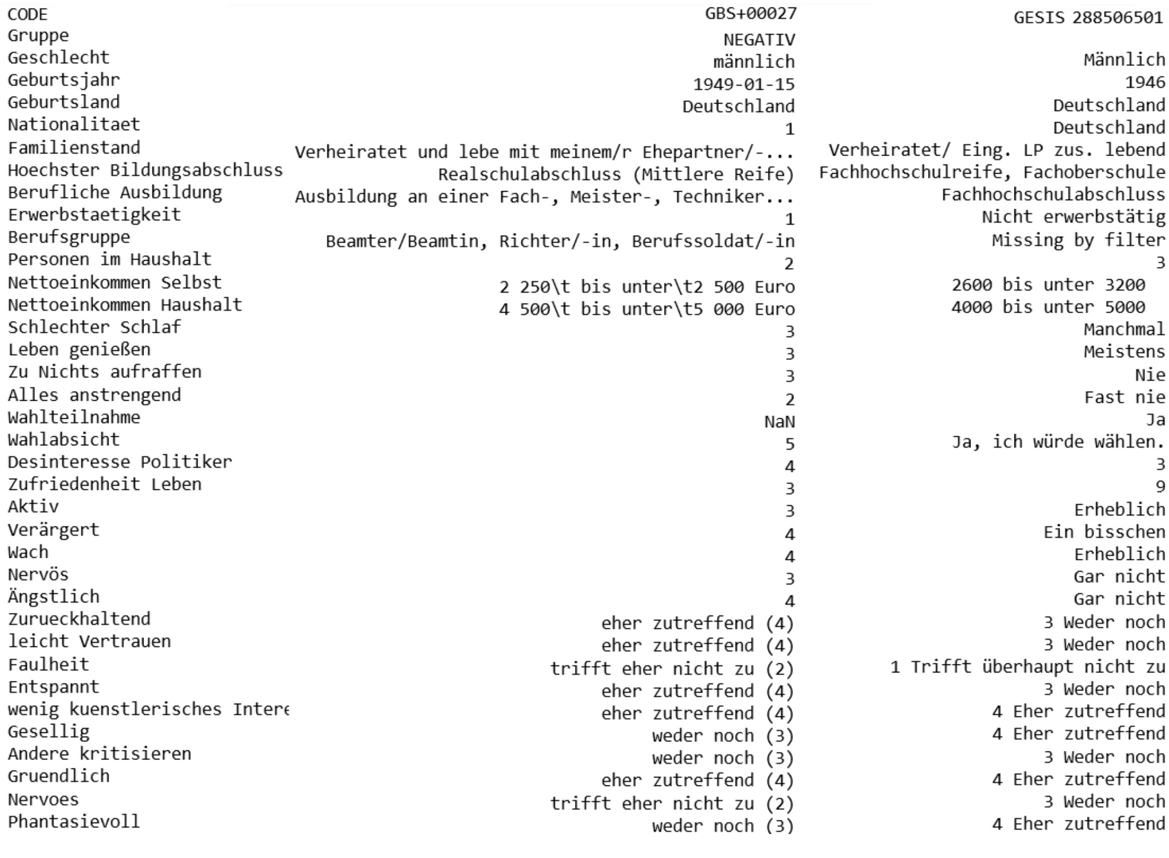
- if h is too simple, $y_i = h(x_i)$ may not hold for many values of i;
- if h is too complex, it fits the data very well but will not generalize well on unseen data.

3 INITIAL DATA ANALYSIS

A brief characterization of the data employed in this evaluation is given.

In order to diagnose to what extent an algorithm suffers from , it will be useful to have another dataset. The da, that is understanding what properties of the data differ between GBS and GESIS.

For completeness, we provide a description of the feature engineering process. We specify the list of transformations that are sequentially applied to each group of features in order to prepare the input for the nal model. All missing values are imputed based on domain-expert decisions. Outliers are handledmanually. Wesplittheinputdataintothreegroupsoffeatures: continuous, categorical and text.



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 3.1: GBS - GESIS attribute and value comparison.

Continuous Features First of all, we perform logarithmic transformation of the skewed continuous features in order to reduce the skewness. Second, westandardizethecontinuousfeaturesbyremovingth emeanandscalingto unit variance. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the dataset. Categorical Features We encode all categorical variables with values between 0 and nclasses-1 where nclasses is the number of distinct categorical values for a given variable. Next, we perform one-hot encoding.

This combination of class imbalance with non-stationary environments poses significant and interesting practical problems for classification

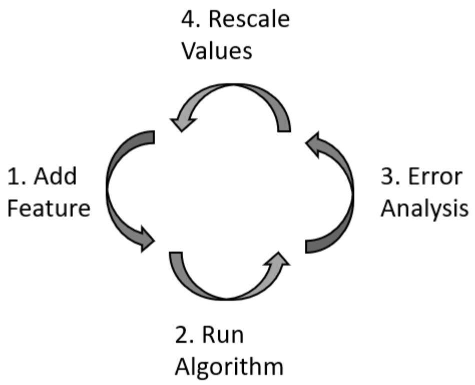


Figure 3.2: GBS - GESIS attribute and value comparison.

3.1 Political Participation and Resilience

3.1.1 The Big Five Dimensions

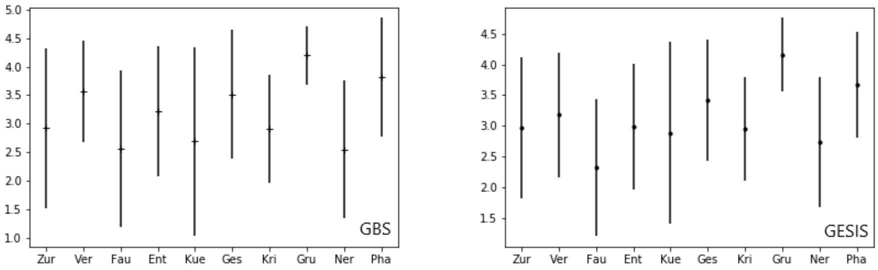


Figure 3.3: Mean and standard deviation of BFI-10 Items:*"A Short Scale for Assessing the Big Five Dimensions of Personality"*.

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, two BFI items for each dimension, representing both the high and low pole of each factor [Fig X]. 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.

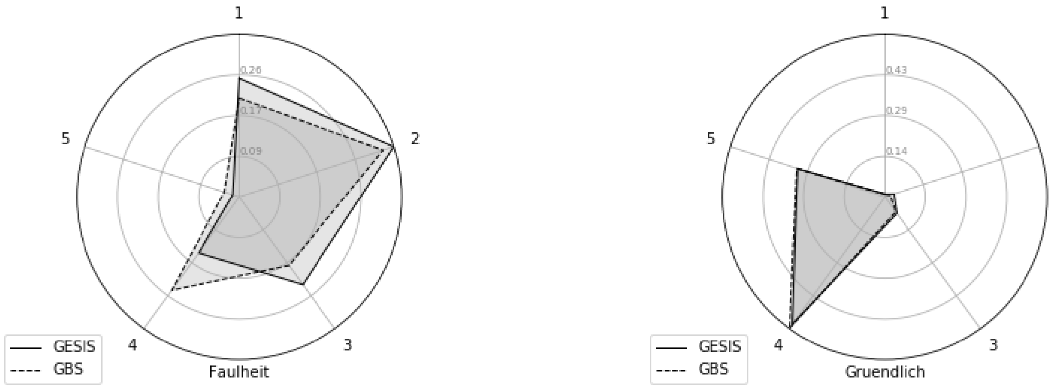


Figure 3.4: Conscientiousness: the degree of organization, self-regulation, and responsibility one exhibits. *"I see myself as someone who tends to be lazy."*(left). *"I see myself as someone who does a thorough job."*(right).

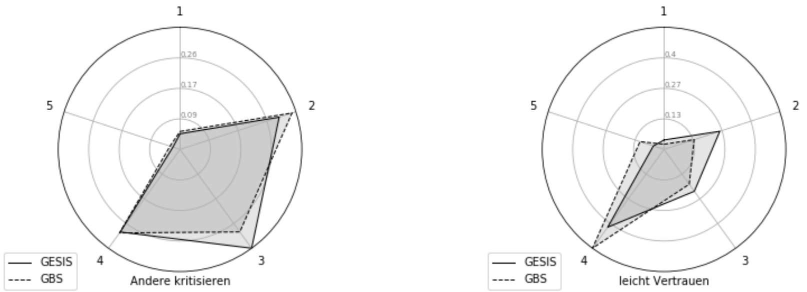


Figure 3.5: Agreeableness: *""I see myself as someone who tends to find fault with others"*(left). *"I see myself as someone who is generally trusting"*(right).

Agreeableness is the measure of one's cooperation, empathy, and willingness to trust and

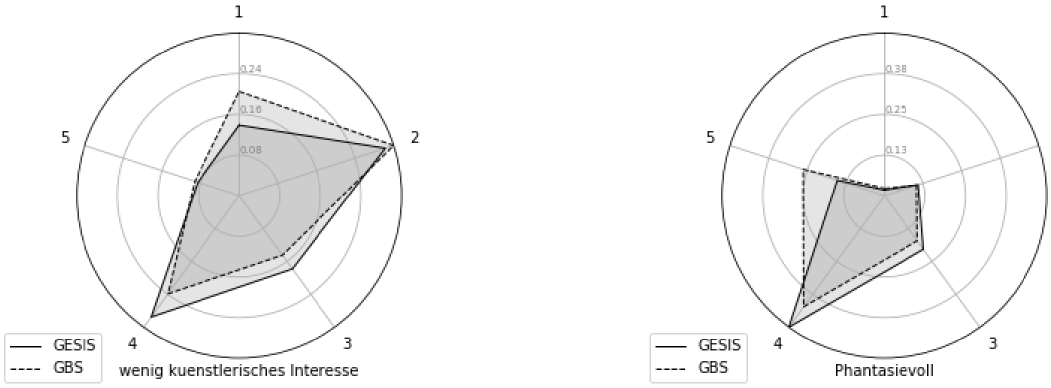


Figure 3.6: Openness: *" I see myself as someone who has few artistic interests"*(left). *"I see myself as someone who has an active imagination"*(right).

help others. Openness refers to openness to new experiences (i.e., whether one is timid/hesitant or eager about new objects or situations), level of inquisitiveness or curiousity, and level of preference for variety, novel stimuli and creativity. Extraversion refers to level of sociability, seeking and enjoyment of social contact, and energy and assertiveness in social situations. Extraverted people are outgoing and loquacious, while introverted people are reserved and favour solitude to social contact. Positive high affect (emotion) is linked to this concept. Neuroticism is characterized by easily experiencing disagreeable or negative emotions (anger, disappointment, frustration, etc.), and a poor coping response to those emotions

is still ongoing debate on whether to use a Likert scale item as categorical or numeric feature. The intervals between positions on the scale are monotonic but never so well-defined as to be numerically uniform increments. A "Strongly Agree (5)" response indicates more agreement than "Agree", but it does not show agreement that is five times stronger than "Strongly Disagree (1)".

There is an underlying measurement continuum, but This project treats the responses as if they fell on an interval scale.

Figure X. shows the response distribution of values for "Conscientiousness" of GBS and GESIS participants (see Appendix for a visualization of distribution shifts in "Agreeableness", "Openness", "Extraversion" and "Neuroticism").

Since GESIS and GBS analyse on a group level should be relatively insensitive to problems that may arise.

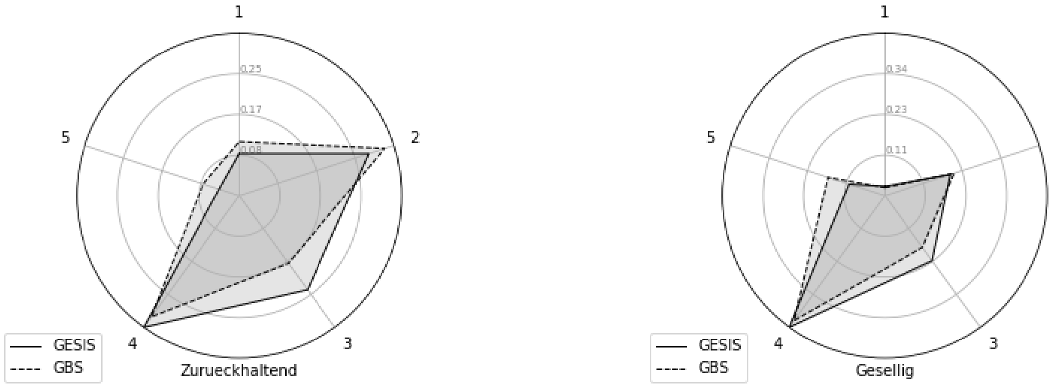


Figure 3.7: Extraversion: *"I see myself as someone who is reserved."*(left). *"I see myself as someone who is outgoing, sociable. "*(right).

In all these cases each aggregate measure (perhaps the mean) is based on many individual responses (e.g., n=50, 100, 1000, etc.). In these cases the original Likert item begins to take on properties that resemble an interval scale at the aggregate level. life satisfaction of states or countries, job satisfaction of departments,

The graphs are almost identical for the Likert item "Gruendlich". Kiviat diagrams are chosen over bar plots and histograms, as the data is not continuous but values are still related. 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.

3.1.2 Demographic and Personal Data

3.2 Survey Mismatch

3.2.1 Likert-Type Scale

John Tukey wrote otherwise (back in 1960) in a monograph "Data Analysis and Behavioral Science" (published in Collected Works v. III). One result he obtained is that if you're getting better than about 10percent test-retest agreement, your scale isn't narrow enough!

[1] It is confirmed that different numbers of rating bars in a subjective rating scale can have significant effects on the subjective measurement, thus the assessment of the Big Five dimensions of personality.

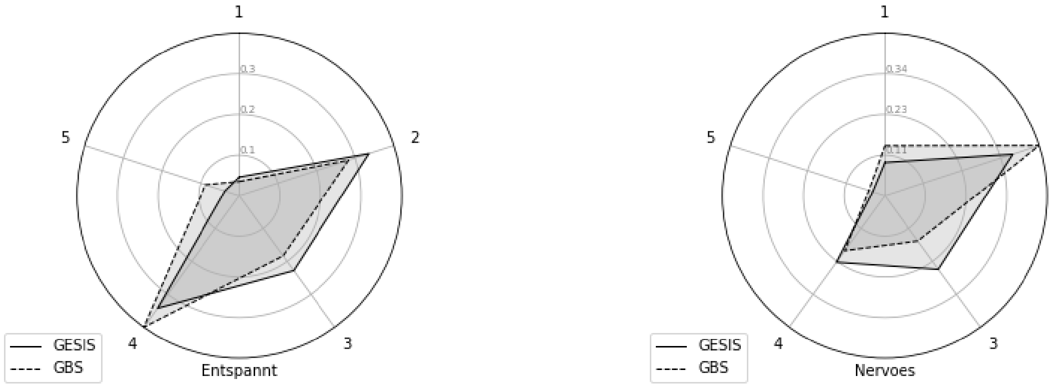


Figure 3.8: Neuroticism: *"I see myself as someone who is relaxed, handles stress well. "*(left). *"I see myself as someone who gets nervous easily."*(right)..

Techniques for reducing the length of scales while maintaining psychometric quality. Figure X shows an example of such as a statement.

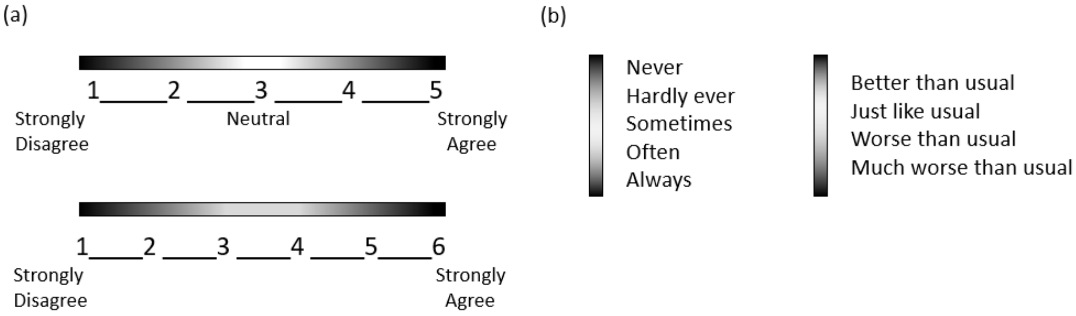


Figure 3.9: Example of a Likert item discrepancy. GBS uses an odd number of responses with a "neutral" option, such as "no opinion", "neither agree nor disagree" or some phrase to that effect. In contrast, there is an even number of responses for this item in GESIS encouraging participants to voice a positive or negative opinion.

In some cases, an additional "opt-out" option is provided for those respondents who truly cannot respond in GBS only.

3.2.2 Imbalanced Data

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 3.1: Different value scalings of attribute ”Desinteresse Politiker”.

4 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.

This leads to a binary classification task with GESIS as positive class and GBS as negative class, i.e. representative and non-representitve sample respectively. Discriminative learners will look for decision boundaries to distinguish the different views of GBS from GESIS in each of the four reference studies. False negatives are then more closely aligned with the target probability distribution. The process of classification is repeated until the learner starts fitting noise more than is warranted. To avoid overfitting, the learning objective needs to be refined as contingency tables lack proper interpretibility. An importance weighted adaption of cross-validation serves as model selection criterion. Given the imbalanced nature and size of GBS, learning is restrained to simpler algorithms with lesser degrees of freedom. The fraction of false positives in the result set of this procedure is kept as proxy measure for the subsequent method positive-unlabeled learning (PU learning). The development of classication models in this setting is often referred to as positive-unlabeled learning (Denis et al. 2005).

Some examples include sex, age, education level, socioeconomic status or marital status. Information collections with biased tendencies can't generate a representative sample.

Variables considered in the study must accurately reflect the populations characteristics.

Consider *attribute: income* of a subset of GBS participants. Statistical significance tests, e.g. Kolmogorov-Smirnov, Chi-Squared

Given a subset of GBS, similarity scores can be defined to evaluate the distance to reference distributions from GESIS. Kolmogorov-Smirnov tests or Chi-Squared assess the likelihood of an attribute of GBS There are $2^{|GBS|} = 2^{587}$ subsets of GBS. Evaluating every possible combination of GBS participants and its score is computationally intractable.

A well-dened learning problem where large polls (Umfragedaten) may contain valuable implicit regularities, requires a well-specied task, performance metric and source of training experience [2]. The MRS problem is now stated as a binary classification task with GESIS as positive class and GBS as negative class. Consider designing a computer program to learn to

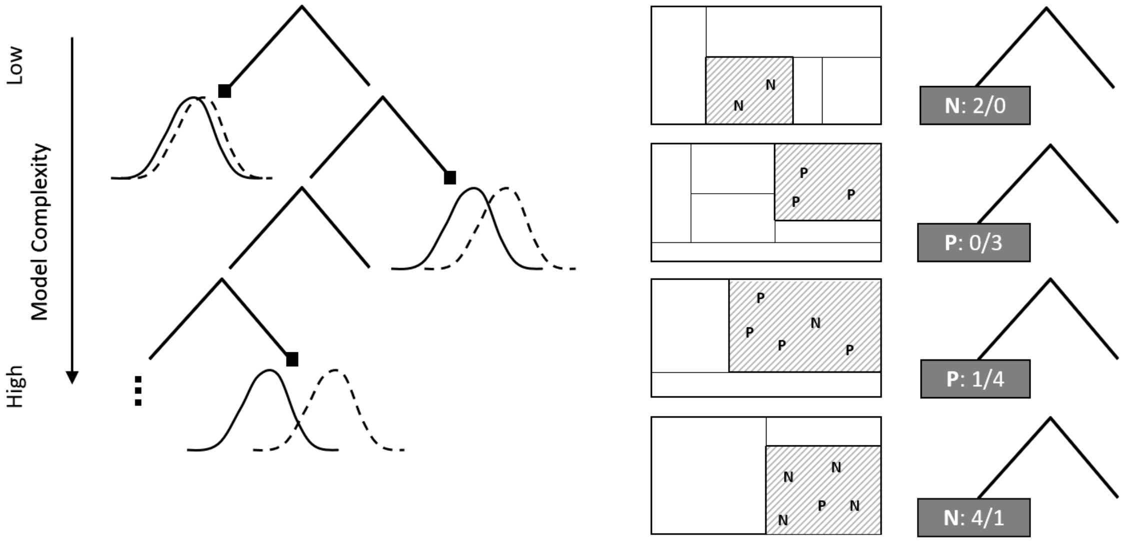


Figure 4.1: .

distinguish between . Using prior knowledge together with past experience to guide learning, a machine learning algorithm is fed with data from games that have been played by chess grandmasters. From this information, the program will learn to apply certain functions to specic board states and make decisions about which move to play next.

Consider a randomly chosen survey participant, i.e. an instance of GBS or GESIS. If the poll indicates the or

Descriptive statistics can be used to

No practical amount of data can distinguish between two distributions, thus instances of GBS can not be proven to come from GESIS. However, discriminative learning allows to infer the conditional probability of *'instance of GBS/GESIS'* given the survey data within a probabilistic framework:

Discriminative learners will look for decision boundaries to distinguish the different views of GBS from GESIS. False negatives are then more closely aligned with the target probability distribution. The process of classification is repeated until the learner starts fitting noise more than is warranted. To avoid overfitting, the learning objective needs to be refined as contingency tables lack proper interpretibility. Given the imbalanced nature and size of GBS, learning is restrained to simpler algorithms with lesser degrees of freedom. The fraction of false positives in the result set of this procedure is kept as proxy measure for the subsequent method positive-unlabeled learning (PU learning). The development of classication models in this setting is often referred to as positive-unlabeled learning (Denis et al. 2005).

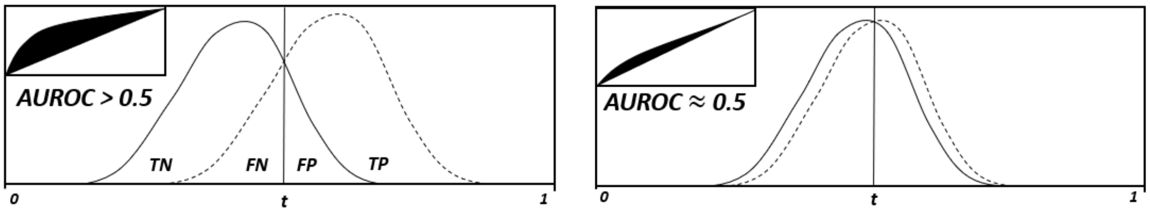


Figure 4.2: There is no caption for such a stupid figure.

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.

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.

For further, more technical reading, references to background papers are provided in [38].

Overfitting 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. By definition, statistical inference is taking the results of applying some sort of construct or model to specific data and then speculating that it would continue to perform well beyond the original observation range.

State-of-the-art techniques in positive-unlabeled learningtackle this problem by treating the unlabeled sample as neg-atives and training a classifier to distinguish between la-beled (positive) and unlabeled examples. Surprisingly,for a variety of performance criteria, non-traditional classi-ers achieve similar performance under traditional evalua-tion as optimal traditional classifiers (Blanchard et al. 2010;Menon et al. 2015).

4.1 Learning from Positive and Unlabeled Data

Breiman [9] introduced bagging as a technique to construct strong ensembles by combining a set of base models. Breiman [10] stated that the essential problem in combining classifiers is growing a suitably diverse ensemble of base classifiers which can be done in various ways [12]. In bagging, the ensemble models use majority voting to aggregate decisions of base models which are trained on bootstrap resamples of the training set. From a Bayesian point of view,

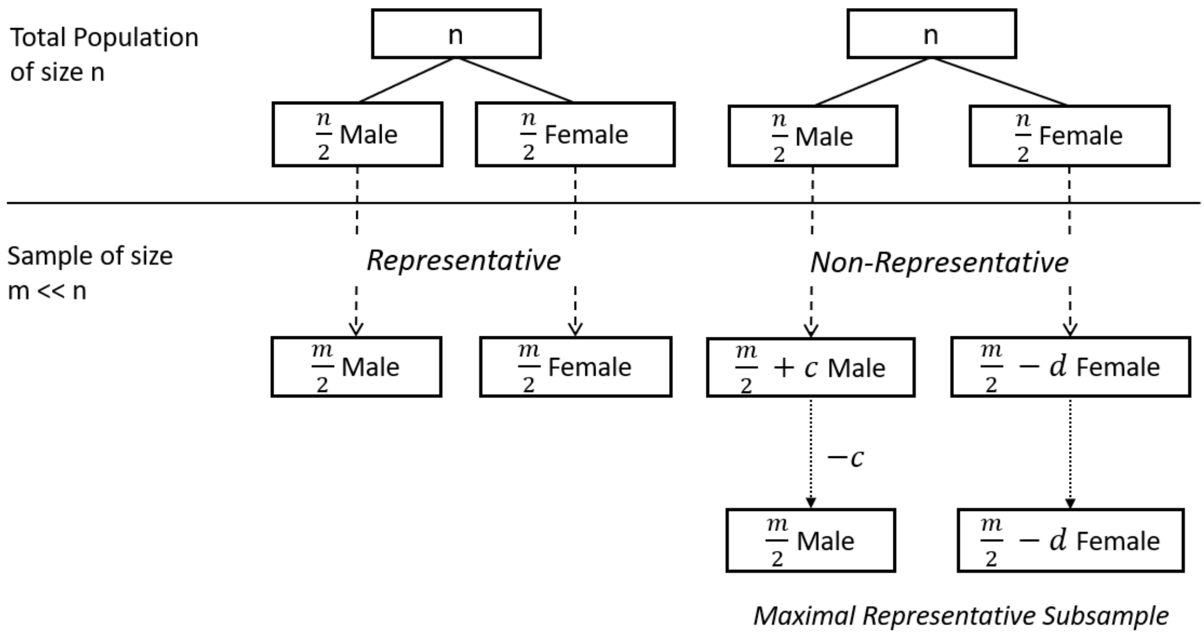


Figure 4.3: Consider samples of fixed size m with some constants c_m and d_m. Maximal representative sampling adjusts for nonresponse $-d$ of subgroup $Female$ by removing $+d$ of subgroup $Male$ from the sample.

bagging can be interpreted as a Monte Carlo integration over an approximated posterior distribution [40].

In his landmark paper, Breiman [9] noted that base model instability is an important factor in the success of bagging which led to the use of inherently instable methods like decision trees in early bagging approaches [19, 11]. The main mechanism of bagging is often said to be variance reduction [4, 10]. In more recent work, Grandvalet [24] explained that base model instability is not related to the intrinsic variability of a predictor but rather to the presence of influential instances in a data set for a given predictor (so-called leverage points). The effect of bagging is explained as equalizing the influence of all training instances, which is beneficial when highly influential instances are harmful for the predictors accuracy.

We have shown the effect 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 influence of training points as described by Grandvalet [24]. Variance reduction. Resampling a contaminated set yields different 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 tradeoff 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)

```
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 $f : U \to \mathbb{R}$

Initualize: $f(x) \leftarrow 0$ 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
Return
    $s(x) = f(x)/c(x)$
```

Algorithm 1: PU training procedure

	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 4.1: 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.

4.1.1 Recovering Model Performance

methods to estimate true classicationperformance.

be recov-ered with the knowledge of class priors

results in biased empirical estimates of the classier performance now can be corrected with the knowledge of class priors 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;

The ROC curve provides in-sight into trade-offs between the classiers accuracies onpositive versus negative examples over a range of decisionthresholds. (pr-rc) curve, a plot of precision as a function of recall.The precision-recall evaluation, including summary statis-tics derived from the pr-rc curve, may be preferred to ROCcurves when classes are heavily skewed (Davis and Goad-rich 2006).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.

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).

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.

Though 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 trueperformance has been much less explored.

the widely-accepted evaluation approaches us-ing ROC or pr-rc curves are insensitive to the variation ofraw prediction scores unless they affect the ranking.

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.

The two main criteria considered in this work are the area under the ROC curve (AUC) and the F-Measure. 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). Another important performance criterion generallyused in information retrieval relies on the precision-recall

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

Recall γ, FPR η and Precision ρ are defined as:

$$\begin{aligned} \gamma &= P[\hat{Y} = 1 | Y = 1] \\ \eta &= P[\hat{Y} = 1 | Y = 0] \\ \rho &= P[Y = 1 | \hat{Y} = 1] \end{aligned}$$

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 $f0$.

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

The area under precision-recall curves $AUPR$ can be expressed using the approximated value for the fraction of positives α in X_u

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

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

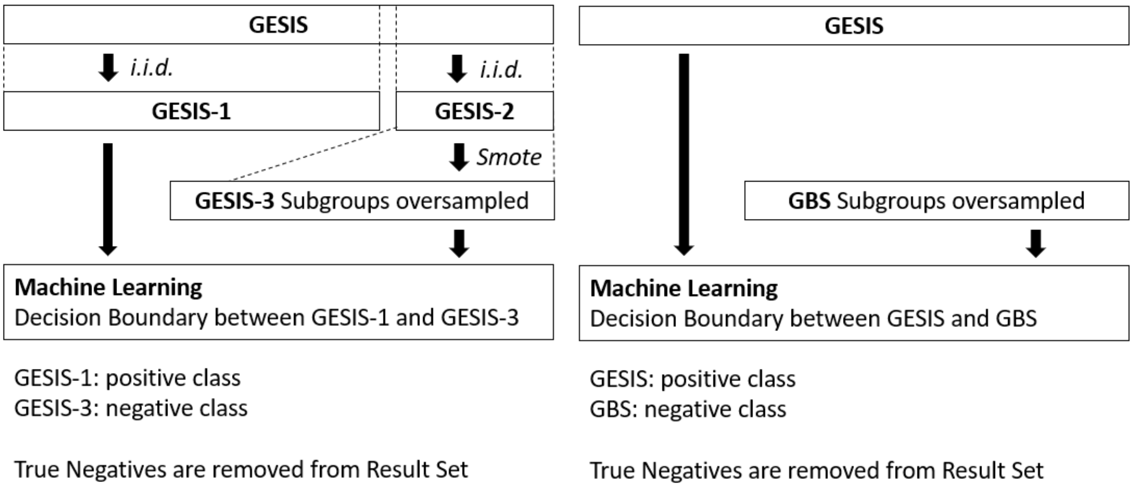
$$AUC > AUC^{pu} \iff AUC^{pu} > \frac{1}{2}$$

TODO: Text here

4.2 Artificial Data Synthesis

4.3 Synthetic Minority Oversampling Technique

Thischapterevaluatestheproposedstabilitymeasureandthecorresponding model selection approach by applying it to several synthetic problems and a real-life LGD modeling problem. The experimental setup is as follows. First


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
GESIS
GBS Subgroups oversampled
Machine Learning
Decision Boundary between GESIS and GBS

GESIS-1: positive class
GESIS-3: negative class

True Negatives are removed from Result Set

GESIS: positive class
GBS: negative class

True Negatives are removed from Result Set

5 RESULTS

5.1 One-Class Classification

5.2 Traditional Machine Learning

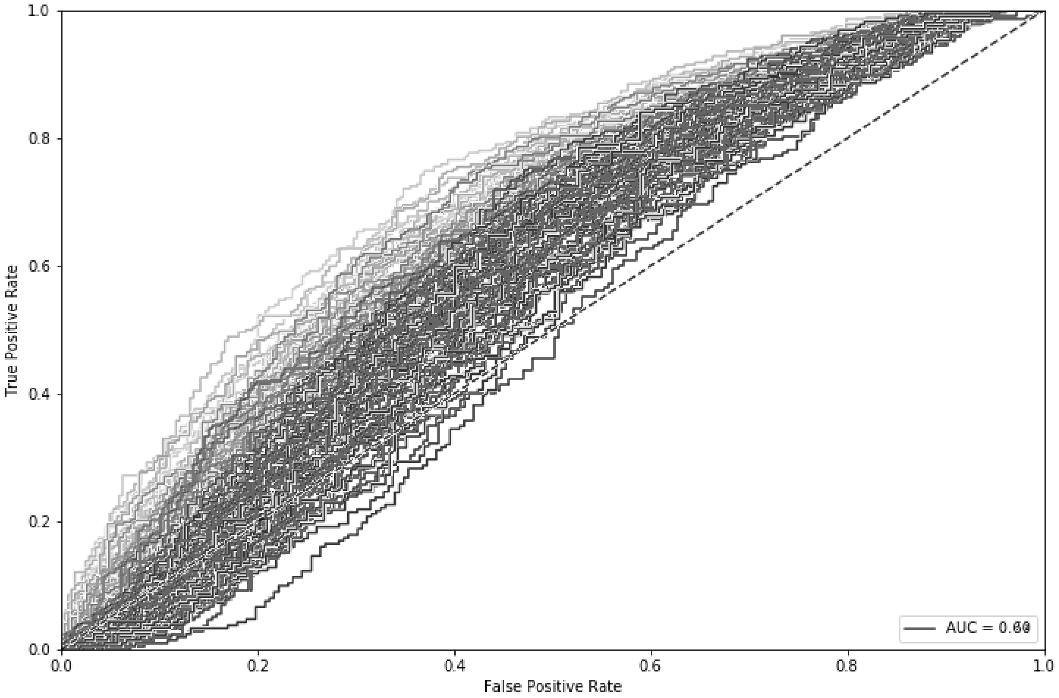


Figure 5.1: .

5.3 Positive Unlabeled Learning

5.3.1 Fraction of Representatives

Estimating Positive Class Prior with One-Class SVMs. In a simple random sample, one can assume that observations are independent from each other. The complex sample design of

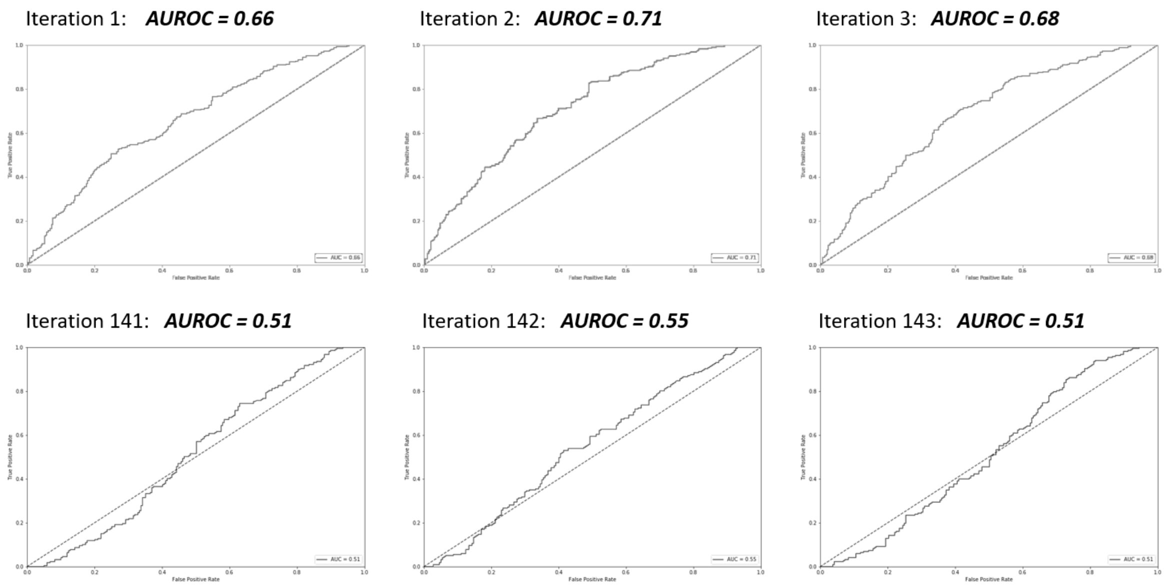


Figure 5.2: .

GBS however

, e.g. multi-stage samples from different survey periods, Complex sample design, such as multi-stage samples of schools, classes and students, students from one classroom are likely to be more correlated than those from another classroom.

we need to compensate for complex survey designs with features including, but not limited to, unequal likelihoods of selection, differences in response rates across key subgroups, and deviations from distributions on critical variables found in the target population from external sources, such as a national Census

most commonly through the development of survey weights for statistical adjustment. If complex sample designs are implemented in data collection but the analysis assumes simple random sampling, the variances of the survey estimates can be underestimated and the confidence interval and test statistics are likely to be biased (Heeringa, West, Berglund, 2010).

In a recent meta-analysis of 150 sampled research papers analyzing several surveys with complex sampling designs, it is found that analytic errors caused by ignorance or incorrect use of the complex sample design features were frequent. Such analytic errors define an important component of the larger total survey error framework, produce misleading descriptions of populations and ultimately yield misleading inferences (Aurelien, West, Sakshaug, 2016). It is thus of critical importance to incorporate the complex survey design features in statistical analysis.

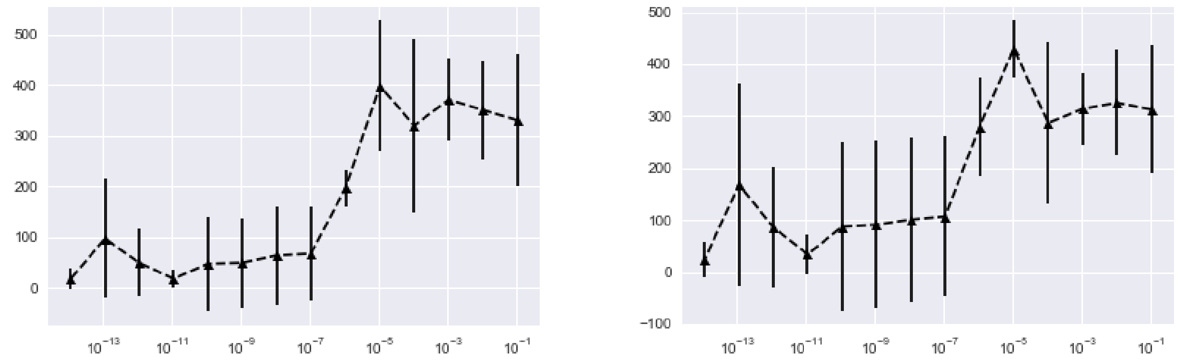


Figure 5.3: Tuning parameter nu that controls the trade-off between the fraction of non-representative samples and the number of support vectors in one-class SVM. More than 0.73 of GBS (right) are classified as representative with high confidence (low sdt) for the optimal value $nu = 10^{-5}$.

5.4 Maximal Representative Subsample

There is almost always not enough data available to partition it into separate training and test sets without losing significant 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 stratification. Figure X shows the results on the GFI-10 data.

5.5 Predicting Political Participation

XGBoost is an open-source software library for predictive modelling, created by Tianqi Chen in 2014 [1]. The name XGBoost stands for "Extreme Gradient Boosting" and implements "Gradient Boosting" as proposed in Greedy Function Approximation: A Gradient Boosting Machine by Friedman, while the term "extreme" refers to the engineering goal to maximize the resources used by the algorithm to achieve high accuracy, computational efficiency and scalability. What started off as a terminal application for a research project, has become a scalable end-to-end tree boosting system that has integrations with scikit-learn in Python, the caret package in R, as well as big data frameworks like Apache Spark and Hadoop. Since its introduction, XGBoost has been used in more than half of the winning solutions in Kaggle competitions, after it gained much popularity and attention in the community by winning the Higgs boson machine learning challenge.

6 CONCLUSION

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