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, January 17, 2019

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. The concept of positive-unlabeled learning is then applied to further improve results.

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 situations 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 [1]. 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 conducted its survey by selecting people from Mainz, while trying to generalize to the entire German population. The survey data (GBS, n=579) 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 [3]. Datasets 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 [4].

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=4249) 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. 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.

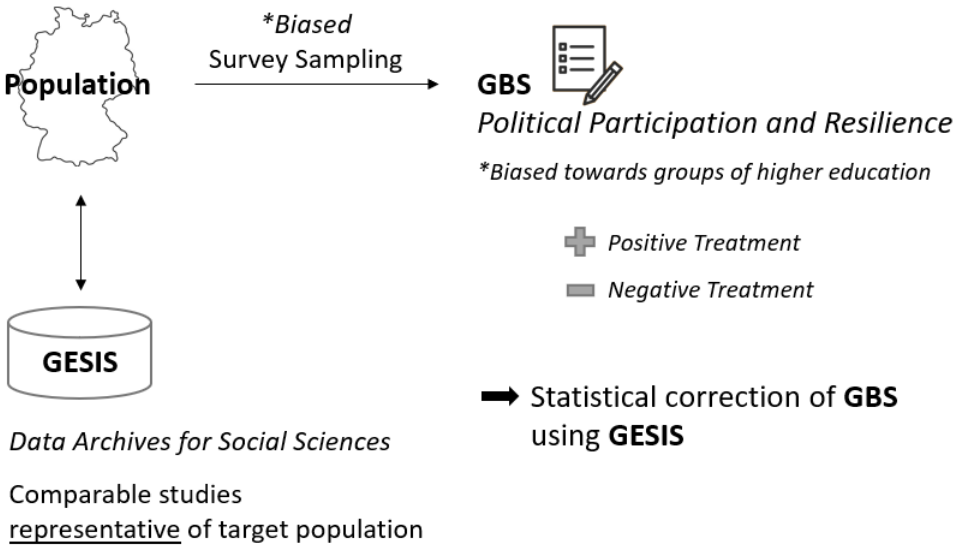

Figure 1.1: Multivariate auxiliary information GESIS linked to GBS so that expected bias can be detected and corrected for. In addition, GBS contains an attribute for positive or negative treatment of survey participants for further analysis.

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 $\frac{1}{2}$. The fraction of misclassified GBS instances is kept as proxy measure for the subsequent method of positive-unlabeled learning [12,13]. Positive-unlabeled learning

is a semi-supervised technique that does not make the simplifying assumption of GBS being positive or negative. These procedures can only work when most of the observations come from a feature space which is not specific to one of the surveys.

1.1 Related Work

Both theoretical and practical results show that the out-of sample error increases proportionally to the distribution shift. To compensate for the degradation in performance, many techniques have been designed to reduce the effects of covariate shift. The influence of sampling bias could be alleviated by weighting the instances according to their importance. Statistical adjustment might also be reached by developing survey weights for calibration estimators. However, these techniques require density estimation which is known to be a hard problem especially in high-dimensional cases [10, 21, 22]. In this work, discriminative learners characterize the importance without explicitly estimating the density ratios.

1.2 Outline

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. Section 3 defines key terminology and introduces positive-unlabeled learning. The resulting maximal representative subset of GBS is presented in Section 4. Finally, Section 5 and Section 6 will conclude by putting the results into perspective and providing an outlook on possible future research.

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 [1] 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.

[1] `https://github.com/laksannathan/mrs-thesis`

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 Section 2.2.

Fig. 2.2 lends itself to first thoughts about whether missing data elements depend on observable attributes or occur entirely at random. The graph is also able to detect functional dependencies in GESIS. Fig. 2.3 summarizes the main observations from GBS. The correlation matrix with ratio -1 to +1 in Fig. 2.4 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. Fig. 2.5 visualizes the response distribution of two BFI items for the "Conscientiousness" dimension, representing

both the high and low pole. "Agreeableness", "Extraversion", "Neuroticism" and "Openness" can be compared in the output folder of the project repository [2].

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 surveys [6].

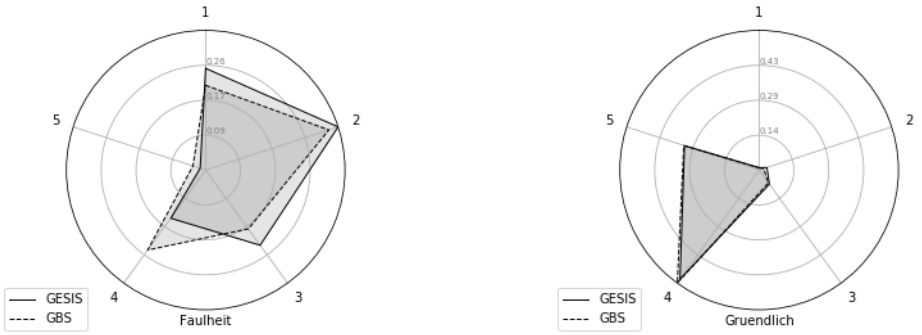

Figure 2.5: Conscientiousness is the degree of organization, self-regulation, and responsibility one exhibits. *"I see myself as someone who tends to be lazy."*(left). *"I see myself as someone who does a thorough job."*(right). The graphs are almost identical for the Likert item "Gruendlich". Respondents 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, considering the use and limitations of Likert scales [7].

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 [5]. 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. GESIS often contains textual values, which often cannot be clearly assigned to numeric

scales, so that the value counts have to be used to support the decision making. Table 2.2 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

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. The respondent may not respond and that type of nonresponse would also be considered missing values indicated by "-1" [8].

Raw Data	GESIS	1	2	3	4	5	6
3	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 that can be problematic for statistical learning.

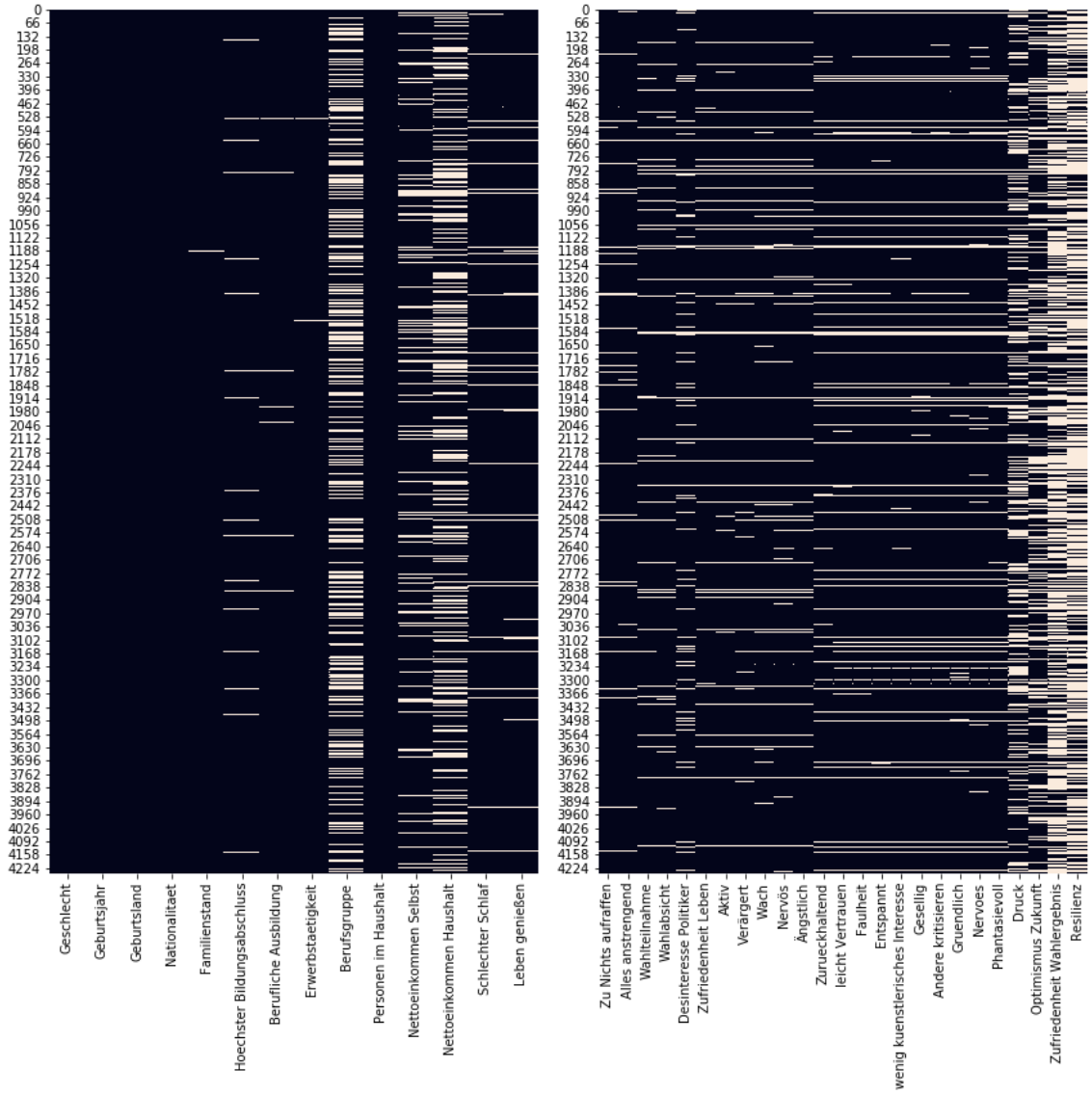

Figure 2.2: Missing values in GESIS. The attribute values of a participant are always known for "Geschlecht", "Geburtsland", "Geburtsjahr", "Nationalitaet", "Familienstand", "Personen im Haushalt". In contrast, the last three columns "Druck", "Optimismus Zukunft", "Zufriedenheit Wahlergebnisse" and "Resilienz" 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.

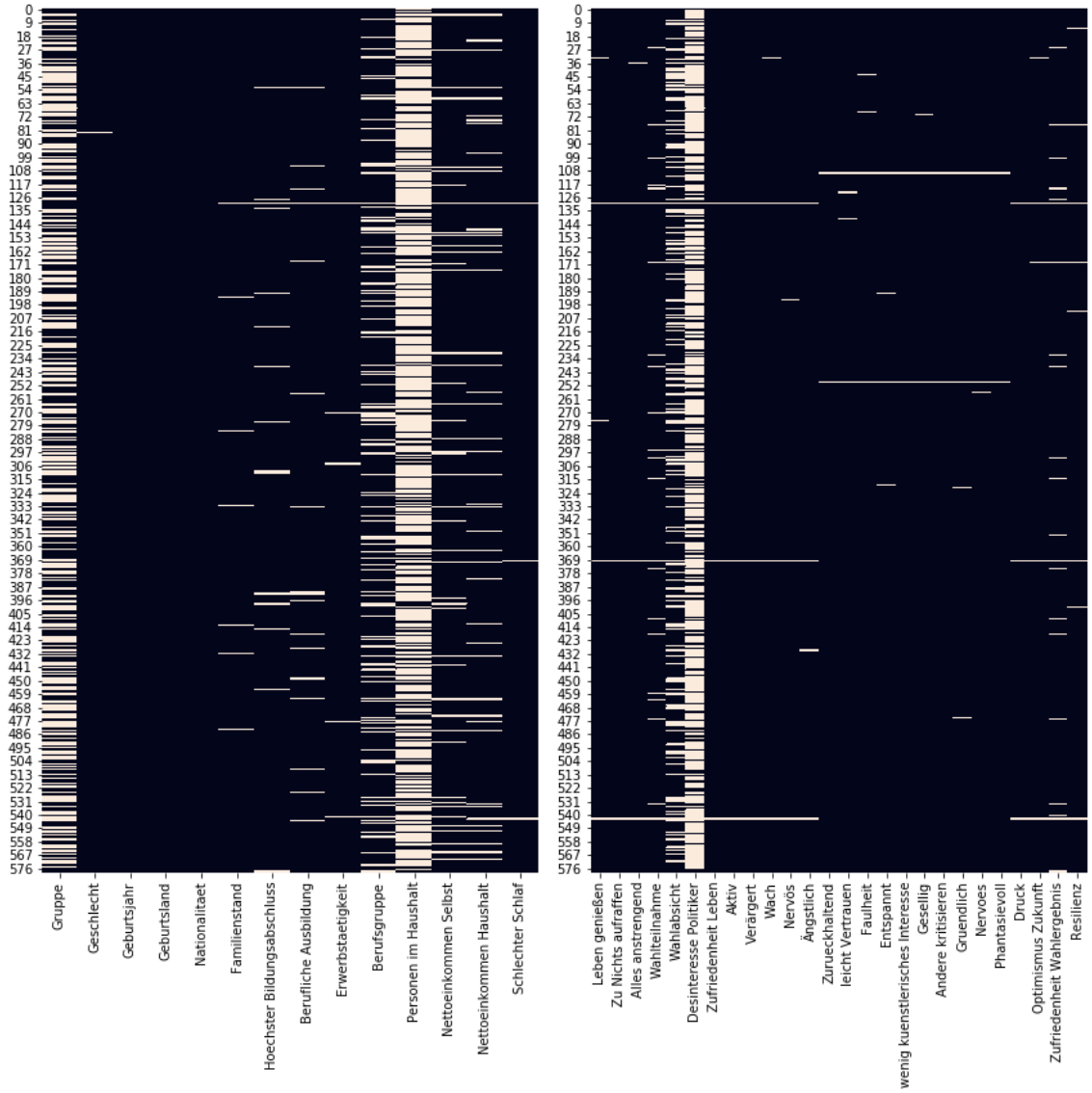

Figure 2.3: Missing values in GBS. There is one more attribute in "Gruppe". Not every participant received a positive or negative psychological treatment. Therefore, "Gruppe" is more likely to be missing than not. However, the absence of a value indicates no treatment rather than a missing positive or 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 needs 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.

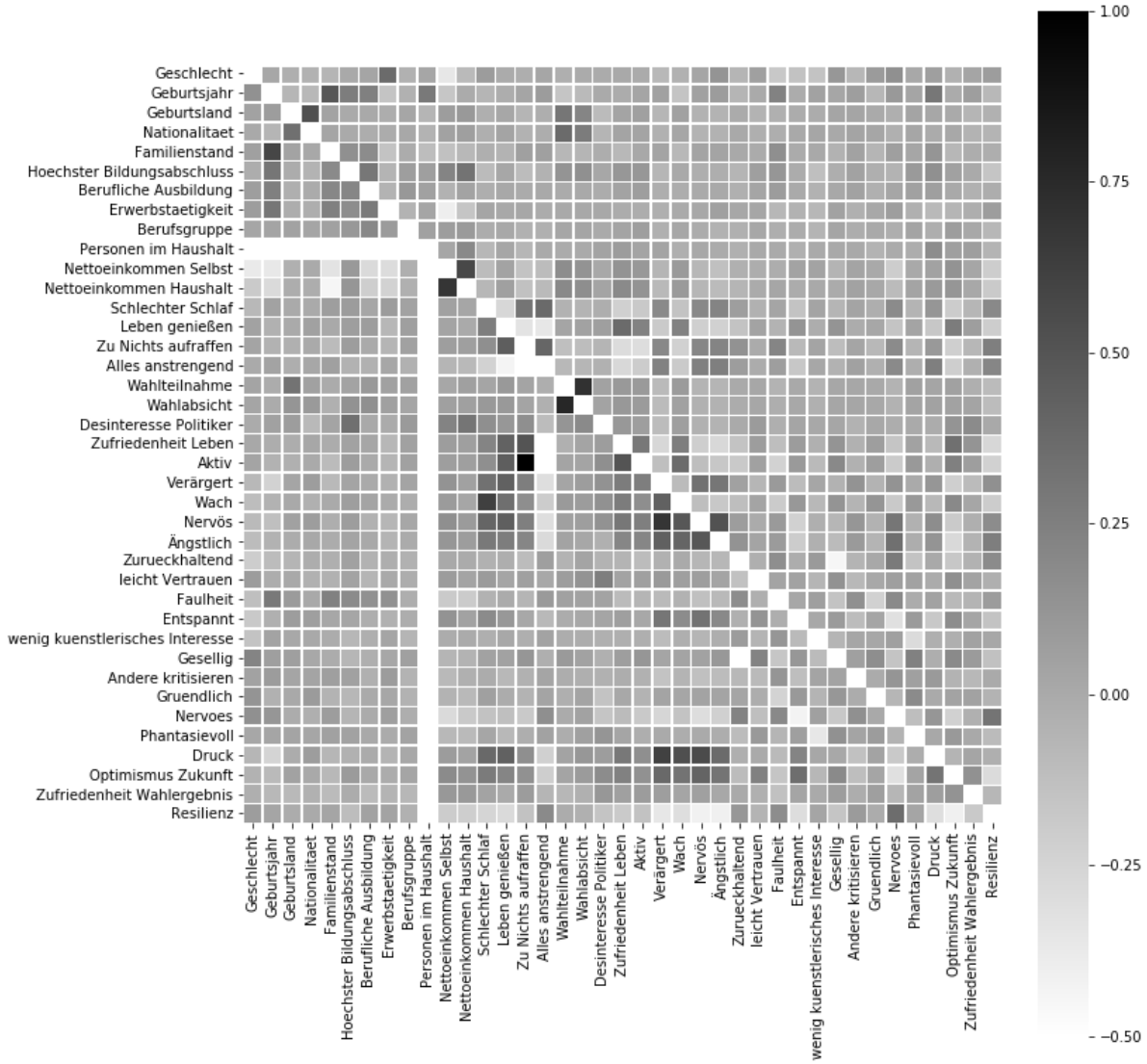

Figure 2.4: The upper right triangular matrix shows GESIS correlations while GBS correlations are shown in the lower left. The main diagonal should not be confused with white squares. These trivial combinations are simply excluded and not colored black. As can be seen "Personen im Haushalt" in GBS can 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. Entropy-based mutual information in "Wahlteilnahme" and "Wahlabsicht" have led to almost perfect classification performances in predicting political participation. "Wahlabsicht" is therefore removed.

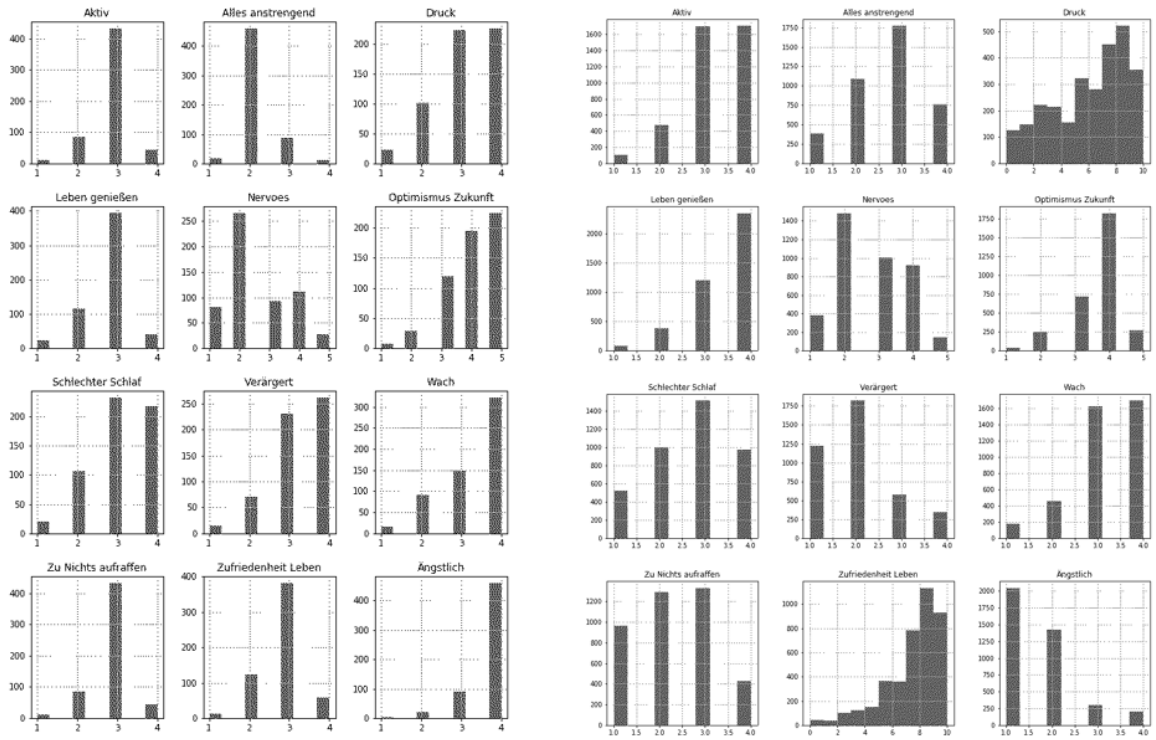

Figure 2.6: Overview of attributes that can not be used for further analysis due to data mismatch. GESIS histograms (right) show differently scaled values, some of which have been transformed already. Taking the inverse of "Aengstlich", "Alles anstrengend" and perhaps "Veraergert" for either the left 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. Depending on the discriminative algorithm, values that only appear in one of the two surveys are enough to perfectly classify instances. 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.

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 [18]. This paper defines key terminology of the theoretical aspects of survey analysis and machine learning. The MRS algorithm is then formulated as positive-unlabeled learning problem in addition to traditional machine learning.

3.1 Sampling Bias

Variables considered in the GBS study must accurately reflect the populations characteristics. That is, every element of the population has a known non-zero chance of being selected. This includes people, whether or not they choose to take part in the survey. Results can then be used to make estimates for not only the sample itself but also the target population. A sampling method is called biased if it systematically favors some people over others, e.g. by oversampling people with strong opinions and undersampling people who do not care much about the topic of the survey [4].

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 [26].

3.2 The Problem of Overfitting

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. 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 [26].

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. It often occurs when the algorithm has too many options to play with in approximating 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 [26]. The error rate on the training data, the resubstitution error, as turned out to be overly optimistic, is not a good indicator of performance on future data. A fair way to properly estimate error rates is to use cross-validation as a powerful general technique. The expected prediction error at any given data point x_0, the generalization error, can be decomposed as follows [25]:

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

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 [26].

3.3 Learning from Positive and Unlabeled Data

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

Training a binary classifier in a semi-supervised way from only positive and unlabeled sample points is called PU learning. A positive set P and mixed set U, which is assumed to contain positive and negative samples, are available for training. A variety of techniques exist to adapt supervised classifiers to the PU learning setting. PU learning is used for tasks such as matrix completion and multi-view learning. It can also be used in data mining to classify data streams, time series and to detect events, like co-occurences, in graphs [13, 20].

In semi-supervised learning, the most straightforward approach to deal with positive and unlabeled data, is to assume that all the unlabeled data are negative and simply apply traditional learning techniques. It has been shown that in many practical cases the posterior distributions in traditional and non-traditional setting provide the same optimal ranking of instances on a given test sample [13, 15].

The first approach handles unlabeled data as negatives and derives a solution that is purely discriminative, i.e. neither GBS nor GESIS distributions are modeled explicitly. Dis-

criminative learners will look for a decision boundary that separates participants according to the survey they have taken. Survey participants on a hold-out set are then classified as either GBS or GESIS, depending on which side of the decision boundary they fall. Predictions are made accordingly to create a ranking. The instance in GESIS with the lowest predicted probability marks the threshold for excluding every GBS instance that is ranked below.

The most extensively studied and widely used performance evaluation in binary classification involves estimating the Receiver Operating Characteristic (ROC) curve. AUROC or AUC, the area under the ROC curve, has a meaningful probabilistic interpretation that is used to rank classifiers as well as instances. The ROC curve visualizes the trade-offs between the classifier's performance on positive versus negative instances over a range of decision thresholds [9]. Evaluation approaches using ROC curves are insensitive to the variation of raw prediction scores unless they affect the ranking.

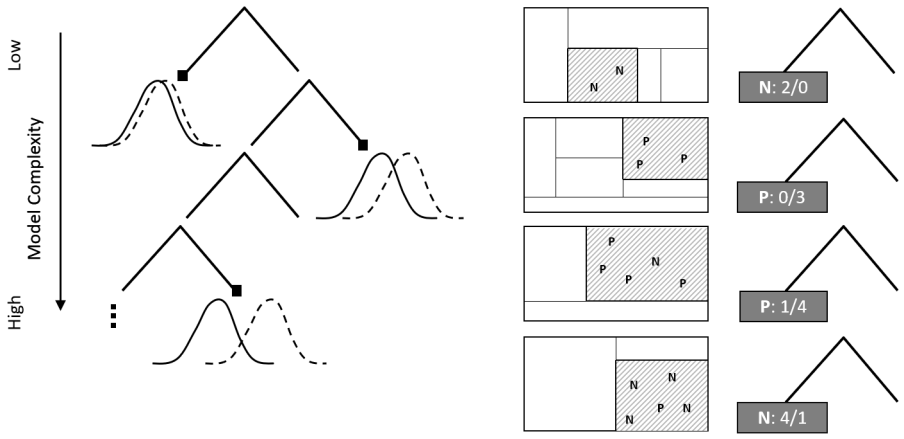

Figure 3.1: Tree-based algorithm as discriminative learner. Increasing model complexity results in a better separation of GBS and GESIS and an improvement of the AUROC on the training data. Predicted probabilities are based on the proportions of positives and negatives in subspaces. Test set instances of GBS in leaf nodes are removed if correctly classified. The terms over-represented and under-represented can be translated to four distinct leaf node possibilities. If all misclassified instances are excluded, this also changes GESIS distribution, of which we know it is representative. Subsampling means adapting GBS distributions only.

The subsequent method PU training has been summarised in Algorithm 1. Base models are trained on resamples from both GESIS and GBS, yielding different levels of fraction of positives in the resamples. The variation of this fraction between base model training sets induces variability between base models without increasing bias. Breiman [23] introduced bagging as a technique to construct strong ensembles by combining a set of base learners. To grow a suitably diverse ensemble of base classifiers, each time a random subsample U_t of U is selected, a classier is trained to discriminate P_t from U_t, and used to assign a predictive score to any element of U that has not been used for training. A final score is obtained by aggregating the predictions of the classifiers trained on subsamples that did not contain the instance itself. Instances are then excluded following the same ranking method [14, 16].

The assigned scores to the elements of GBS reflect our confidence that these elements are representative. However, one may argue that assigning a score to an unlabeled example that has been used as negative training example is problematic. In particular, if a model overfits, a false negative will hardly be given a high training score when used as a negative. Because some of the unlabeled data in the training set are in fact positive, the described procedure results in biased AUROC estimates. In contrast to PU learning itself, estimating the true performance of a non-traditional classifier has not been thoroughly studied. Assessing the performance of binary classifiers becomes a non-trivial task in the absence of known negatives. Estimating the fraction of positives enables us to calculate an upper and a lower bound for the AUROC using the ROC curve in the positive-unlabeled setting [14–16].

Let f be the true distribution over the input space X from which unlabeled data is drawn. With distributions f_1 and f_0 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 $x \in X$ 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 [15].

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}, [15] 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

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

In [14], Claesen et al. emphasize the relative importance of known negatives compared to known positives for the assessment of classifiers in PU learning.

4 RESULTS

This chapter presents the results of the two positive-unlabeled learning approaches for maximal representative subsampling. The implementations of the proposed PU learning techniques have been slightly adjusted, while the idea to improve the assessment of classifiers in a non-traditional setting was preserved. The traditional ROC curve , with logistic regression and support vector machines. The discussions begin with the difficulties of working and dealing with data from two different sources.

4.1 Decision-Trees for Data Processing

Since errors and inaccuracies slipped in quickly, preparing the data took a particularly long time. To ensure that the data has been read and mapped accordingly, various decision-trees have been trained right from the start to detect for mismatch. In fact, this is already an application of the discriminative learning procedure.

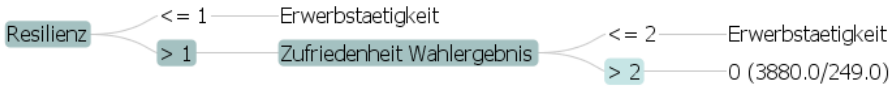

Figure 4.1: Trees were trained using either sci-kit learn or WEKA explorer to identify preprocessing mistakes. The outer right path of the decision-tree demonstrates the desired behavior of the constructed model. "Resilienz" and "Zufriedenheit Wahlergebnis" have been used to classify the majority of instances as GESIS. Both attributes have been measured on the same scale or mapped properly. Technically, this means that there was no better way to split the data at that specific node. Either the algorithm did not see any gain in continuing to split or the tree was fully grown first and then pruned back using out-of sample estimations.

Fig. 4.1 visualizes a path of the latest J48 tree on the current set of features. The entire tree can be seen in the project repository. For the reasons mentioned in Chapter 2, tree-based algorithms are particularly suitable for quickly capturing incorrectly handled data. A

subgroup of all participants is said to be representative of the target population, regarding the particular complexity of the trained model. Model complexity, with respect to the right-most leaf node, gives high confidence to include the subgroup of 249 GBS instances in the MRS.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area	Class
LMT	0.971	0.655	0.916	0.971	0.943	0.878	0.979	GESIS
	0.345	0.029	0.619	0.345	0.443	0.878	0.519	GBS
J48	0.973	0.710	0.910	0.973	0.940	0.716	0.928	GESIS
	0.290	0.027	0.596	0.290	0.390	0.716	0.389	GBS

Table 4.1: J48 and LMT tree evaluation. These trees served as indicators for invalid mappings during development. The true negative rates of 0.029 and 0.027 indicate that there is just a small fraction of non-representative GBS instances. The constructed models are characterized by a high bias and low variance. Note that training was not an attempt to estimate the fraction of positives.

4.2 Maximal Representative Subsample

The initial idea was to train a one-class classifier (OCC) by learning from training data containing only GESIS instances. This technique is very similar to PU learning. To identify survey participants amongst all GBS instances, representative sampling of the negative class is not strictly required [30]. In one-class SVM, the support vector model infered the properties of representative cases and from these properties predicted which GBS participants are unlike the representative examples. Predictions from the One-Class SVM are uncalibrated scores that have not been normalized and therefore can not be compared to models based on different algorithms. Instead, multiple OCCs have been trained on different bootstrap samples from GESIS. The predicted probability was then derived by a majority decision. Fig. 4.2 shows the fraction of correctly classified instances by the OCCs as a function of the parameter ν on a hold-out set.

The OCC suffered from unavoidable variance in the learning task due to insufficient data points in the minority class GBS (n=579). There was clearly not enough data available to partition it into separate training and test sets without losing significant modelling or testing capability. The results of the ensemble were insufficient and unreliable. The question arises

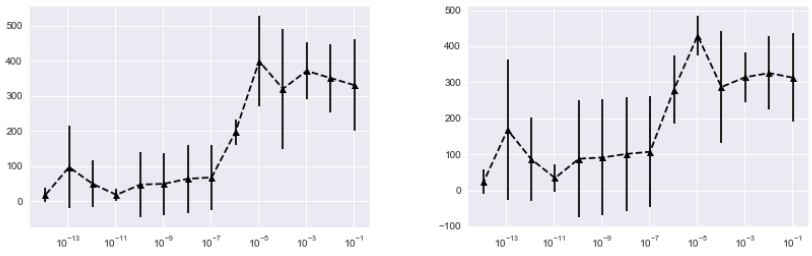

Figure 4.2: Tuning parameter ν that controls the trade-off between the fraction of non-representative samples and the number of support vectors in one-class SVM. More than 0.73 of GBS (right) are classified as representative with low confidence (high sdt.) for the optimal value $\nu = 10^5$. The plot on the left shows that not all GESIS participants have been classified as such.

whether learning is feasible at all. To address this problem, less complex algorithms have been used in the PU training procedure (Algorithm 1). In model selection, logistic regression slightly outperformed linear SVMs with respect to the AUROC as can be seen in Fig. 4.3. The optimized model then predicted the probability of every single instance in the combined dataset. The optimal threshold for the ROC curve with AUROC 0.62 gives an MRS consisting of a little less than half of the participants ($\frac{271}{579}$).

The result set of Algorithm 1 could have remained unchanged for further studies. However, it is the case that the AUROC of 0.62 has not been adapted for the positive-unlabeled setting as described in Section 3. This lies in the fact that no class prior of GBS was available. Subsequent runs have now been able to make use of the estimated fraction of latent positives. Instead of removing multiple misclassified instances, the implemented procedure trains classifiers iteratively to remove one instance at a time. This conceptually more intuitive when it comes to reducing sampling bias and seems less prone to overfitting. Fig. 4.4 demonstrates this procedure by plotting a ROC curve for the predicted probabilities of each classifier. In other words, the MRS is initialized with GBS itself and then subsampled step by step. Although this algorithm looks promising, further research must follow before it can be considered for MRS.

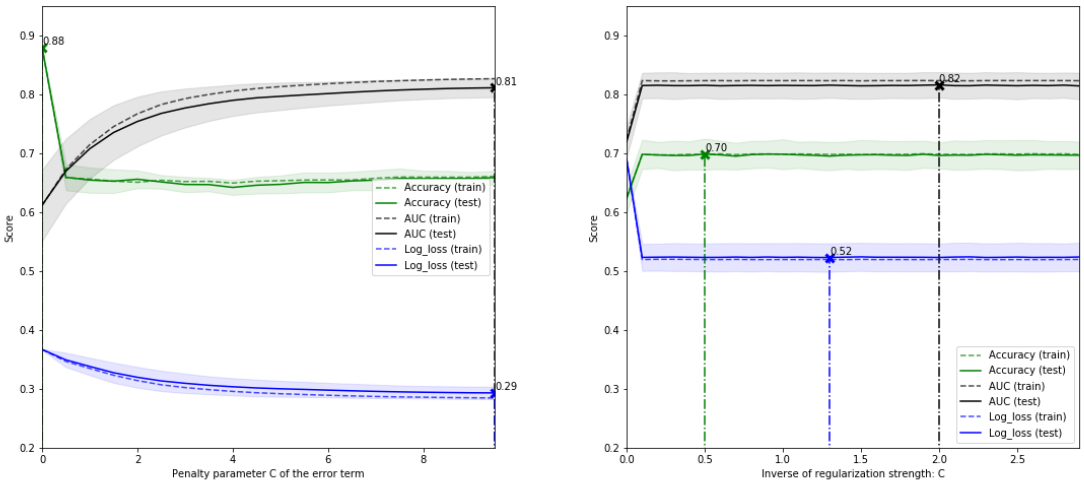

Figure 4.3: GridSearchCV: Gridsearch using repeated 10-fold stratified cross-validation on $\frac{1}{2}$ of the data to evaluate multiple scorers simultaneously. The baselines are 0.89 for accuracy, 0.82 for AUC/AUROC and 0.29 for logarithmic loss. The baselines outperform the trained models on the given metrics, due to the high imbalance in the data. Logistic regression (right) was slightly outperformed by Support Vector Machines (right).

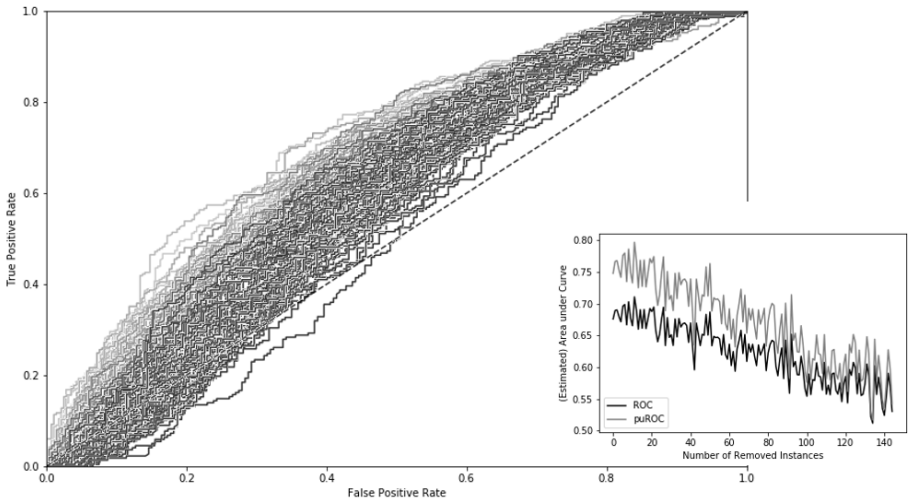

Figure 4.4: ROC and puROC evaluation for PU learning with Random Forests as base models [31]. An AUROC of approximately $\frac{1}{2}$ implies that there is no more evidence for covariate shift. Iterations stopped at an estimated decrease of $\frac{1}{5}$. The remaining instances are considered representative of the target population.

5 FUTURE WORK

Many different 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 synthesized data. The Synthetic Minority Over-sampling TEchnique (SMOTE [32]) 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 Fig. 5.1. It is only necessary to try to avoid giving the synthesized data properties that make 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.

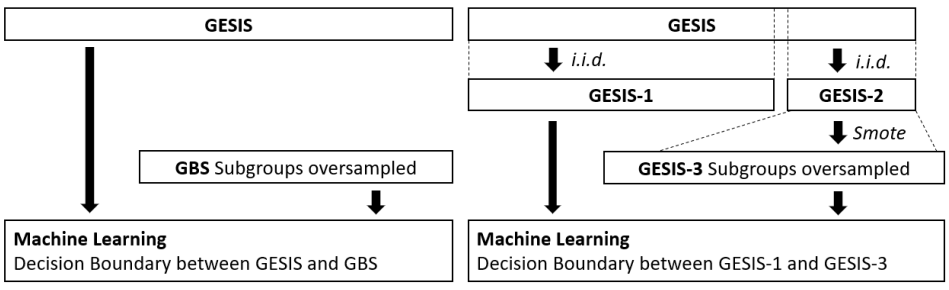

Figure 5.1: Artificial data synthesis to overrepresent subgroups of GESIS. True negatives are removed from the MRS with positive classes GESIS (left) and GESIS-1 (right). Oversampled instances can easily be marked as such for result set comparisons.

The main criterion 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 [27]. Precision and recall have been estimated in Section 3 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: $\rho = \frac{\alpha\gamma}{\hat{\eta}^{pu}}$.

Lastly, it is up to the research group to analyse political participation and resilience on the given MRS. Table 5.1 evaluates an individual attribute by measuring the amount of information gained about the class "Wahlteilnahme" given the attribute using 10 fold stratified cross-validation. Comparisons of the GBS dataset before and after subsampling might be of interest in this context as well. A closer look at the instances that have been excluded might support the initial claim that people of higher education are over-represented.

Init. avg. merit	Init. avg. rank	attribute	MRS avg. merit	MRS avg. rank
0.164 +/- 0.011	1.0 +/- 0.0	Zufrieden. Wahl.	0.006 +/- 0.007	7.3 +/- 3.5
0.112 +/- 0.008	2.0 +/- 0.0	Wahlabsicht	0.096 +/- 0.005	1.0 +/- 0.0
0.040 +/- 0.004	3.0 +/- 0.0	Geburtsland	0.037 +/- 0.005	2.0 +/- 0.0
0.003 +/- 0.001	5.2 +/- 1.3	Gruppe	0.002 +/- 0.001	3.6 +/- 0.6

Table 5.1: Feature importance in GBS (n=579) and GBS MRS (n=271) for classification of political participation "Wahlteilnahme". In modelling a political participation process, algorithms 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.

6 CONCLUSION

This paper dealt with the reduction of sampling bias in survey data for a psychological study. The availability of multivariate auxiliary information provided additional options and also presented new challenges.

Two PU learning variants were presented to reduce sampling bias in GBS. The first approach is consistent with the established practice that has evolved in machine learning of relating training and test distribution. Models are trained to detect for covariate-shift by predicting the origin of each instance. The second procedure can be best described as iterative reduction of AUROC for maximal representative subsampling. The fraction of positives from the previous method allowed for a better assessment of the actual AUROC. The difficulty in either of them consisted in not forcing value comparisons that would essentially introduce non-existent patterns.

Discriminative learners have not been able to distinguish 350 out of 579 GBS survey participants from GESIS participants. The subsequent method reduced the puAUC by 0.11, while the estimated AUC dropped from 0.75 to 0.64. The result set reflects the unbiased distribution of the target population more closely.

And yet subsampling always comes with a loss of information. Depending on the type of following research, it might be more appropriate to calibrate the GBS dataset using weights based on the calculated probabilities. An upcoming discussion with the responsible researchers from the chair of psychology will decide on further proceedings.

BIBLIOGRAPHY

[1] Kalisch R, Mueller MB, Tuescher O (2015) *Advancing empirical resilience research.* Behav Brain Sci. 38:e128.

[2] Rammstedt, B. and John, O.P. (2007). *Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German.* Journal of Research in Personality, 41, 203212.

[3] Candela J, Sugiyama M, Schwaighofer A, et al.: *Dataset Shift In Machine Learning.* MIT Press, Cambridge, Massachusetts,2009.

[4] West BT, Sakshaug JW, Aurelien GAS (2016) *How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data?* PLoS ONE 11(6): e0158120. https://doi.org/10.1371/journal.pone.0158120

[5] Heeringa, S.G., West, B. T., and Berglund, P.A. (2010). *Applied survey data analysis.* Boca Raton, FL: CRC Press.

[6] Likert, R. (1932). *A technique for the measurement of attitudes.* Archives of Psychology, 22(140).

[7] Jacoby, J, and Matell, M.S. (1971). *Three-point Likert scales are good enough.* Journal of marketing research, 8(4), 495-500.

[8] Sullivan, G. M., and Artino, A. R. (2013). *Analyzing and interpreting data from Likert-type scales.* Journal of Graduate Medical Education, 5(4), 541542.

[9] Hanley JA, McNeil BJ (1982) *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology 143:29-36.

[10] Jean-Claude Deville and Carl-Erik Sarndal, *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, Vol. 87, No. 418 (Jun., 1992), pp. 376-382

[11] Hidetoshi Shimodaira, *Improving predictive inference under covariate shift by weighting the log-likelihood function*, Journal of Statistical Planning and Inference 90 (2000) 227244.

[12] Francois Denis, Remi Gilleron, and Fabien Letouzey. *Learning from positive and unlabeled examples.* Theoretical Computer Science, 348(1):7083, 2005.

[13] Marc Claesen, Frank De Smet, Johan AK Suykens, and Bart De Moor. *A robust ensemble approach to learn from positive and unlabeled data using svm base mode*, Neurocomputing, 160:7384, 2015.

[14] Claesen, M.; Davis, J.; De Smet, F.; and De Moor, B. 2015. *Assessing binary classifiers using only positive and unlabeled data.* arXiv preprint arXiv:1504.06837.

[15] S. Jain, M. White, and P. Radivojac. *Recovering true classifier performance in positive-unlabeled learning.* In Proc. 31st AAAI Conference on Artificial Intelligence, AAAI '17, 2017.

[16] S. Jain, M. White, and P. Radivojac. *Estimating the class prior and posterior from noisy positives and unlabeled data.* In Proc. 30th Advances in Neural Information Processing Systems, NIPS '16, pp. 26932701, 2016.

[17] Bickel, S. et al. (2009). Discriminative Learning Under Covariate Shift. Journal of Machine Learning Research, 10, 21372155

[18] Tom Mitchell, *Machine Learning*, 3rd Edition, McGraw-Hill, 1997.

[19] K Sechidis, B Calvo, and G Brown. *Statistical hypothesis testing in positive unlabelled data.* In Machine Learning and Knowledge Discovery in Databases, pages 6681. Springer, 2014

[20] C. Elkan and K. Noto, *Learning Classifiers from Only Positive and Unlabeled Data*, in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008, pp. 213220.

[21] JM Brick, G. Kalton. *Handling missing data in survey research* , 1996, https://doi.org/10.1177/096228029600500302.

[22] Shimodaira, H. (2000). *Improving predictive inference under covariate shift by weighting the log-likelihood function.* Journal of Statistical Planning and Inference, 90, 227244.

[23] Leo Breiman. *Bagging predictors.* Machine learning, 24(2):123140, 1996.

[24] Trevor Hastie, Robert Tibshirani, Jerome H. Friedman: *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[25] Ian H. Witten, Eibe Frank: *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2011.

[26] Yaser S. Abu.Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin: *Learning From Data: A Short Course*, AMLbook.com, 2012.

[27] Jesse Davis , Mark Goadrich: *The relationship between Precision-Recall and ROC curves*, Proceedings of the 23rd international conference on Machine learning, p.233-240, 2006, Pittsburgh, Pennsylvania [doi¿10.1145/1143844.1143874]

[28] Walker, SH; Duncan, DB (1967). *Estimation of the probability of an event as a function of several independent variables.* Biometrika. 54 (1/2): 167178. doi:10.2307/2333860. JSTOR 2333860.

[29] Cox, DR (1958). *The regression analysis of binary sequences (with discussion).* J Roy Stat Soc B. 20 (2): 215242. JSTOR 2983890.

[30] Tax, D. (2001) *One-class classification: Concept-learning in the absence of counter-examples.* Doctoral Dissertation, University of Delft, The Netherlands.

[31] Leo Breiman (2001). *Random Forests.* Machine Learning. 45 (1): 532. doi:10.1023/A:1010933404324.

[32] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. *SMOTE: synthetic minority over-sampling technique.* J Artif Intell Res. 2002;16:341378.

DECLARATION OF AUTHORSHIP

I hereby confirm that I have written this Bachelor thesis independently and without using any sources other than those indicated. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Mainz, January 17, 2019

Laksan Nathan