Data Wrangling for Fair Classification

Lacramioara Mazilu
University of Manchester
Manchester, United Kingdom
lacramioaramazilu@gmail.com

Norman W. Paton
University of Manchester
Manchester, United Kingdom
npaton@manchester.ac.uk

Nikolaos Konstantinou
University of Manchester
Manchester, United Kingdom
nkons@live.com

Alvaro A.A. Fernandes
University of Manchester
Manchester, United Kingdom
fernandesaaa@gmail.com

ABSTRACT

Whenever decisions that affect people are informed by classifiers, there is a risk that the decisions can discriminate against certain groups as a result of bias in the training data. There has been significant work to address this, based on pre-processing the inputs to the classifier, changing the classifier itself, or post-processing the results of the classifier. However, upstream from these steps, there may be a variety of data wrangling processes that select and integrate the data that is used to train the classifier, and these steps could themselves lead to bias. In this paper, we propose an approach that generates schema mappings in ways that take into account bias observed in classifiers trained on the results of different mappings. The approach searches a space of candidate interventions in the mapping generation process, which change how these mappings are generated, informed by a bias-aware fitness function. The resulting approach is evaluated using Adult Census and German Credit data sets.

KEYWORDS

fairness, bias, classification, data preparation, data wrangling

1 INTRODUCTION

Fairness in machine learning is important because machine learning supports decision making; problems resulting from bias have been widely recognised, and there are many results on fairness-enhancing interventions in machine learning [2]. Proposals have been made for taking fairness into account before, during and after learning.

Although proposals that focus on interventions before learning have considered a variety of techniques for selecting or modifying training data [7], most of this work assumes that the training data is already available. As such, the interventions take place after the data preparation steps that select and integrate data sets for analysis. As data scientists spend a considerable portion of their time on such steps, an opportunity seems to exist for making interventions earlier in the data processing pipeline.

In our previous work [12], we investigated how data wrangling processes could be adjusted to address dataset properties that can give rise to bias, specifically *sample size disparity* and *proxy attributes*. Such interventions are promising, in that they can be applied to unlabelled data, but there is no direct guarantee that addressing the underlying property will in fact reduce classifier bias.

In this paper, we investigate interventions in the data wrangling process that directly target classifier bias, when labels are available. Specifically, we assume a situation in which data preparation selects and combines datasets for use in training, and we intervene in data preparation to ensure that these selection and combination decisions are informed by the fairness of the resulting classifier. Similarly to other works, our focus is on the *pre-training* step, however, we do not rely on the assumption that the data is already gathered and that fairness is achieved through operations on the integrated data source. Our focus is on creating the integrated data source in a fashion that takes into consideration fairness measurements.

The contributions are as follows:

(1) The proposal of a strategy for fairness aware data preparation for classification.
(2) The realisation of the strategy as a search for fair data preparation plans.
(3) An evaluation of (2) for benchmark data sets that shows how the interventions can improve a specific fairness metric, namely demographic parity. Also, we show how the accuracy of the trained classifier is impacted by the interventions.

The remainder of this paper is structured as follows. Section 2 provides the context for this work by reviewing related results. Section 3 describes some technical background on data preparation required for later sections. Section 4 details the approach, in which a space of alternative wrangling strategies is explored. Section 5 presents an empirical evaluation, and Section 6 reviews progress against the objectives.

2 RELATED WORK

The problem considered in this paper is as follows: Assume we have access to data sets that can be used to populate a target schema that is suitable for training a classifier. Combine the data sets in such a way as to reduce the bias in the generated classifier.

Previous related work has tended to assume that there is an existing data set that is suitable for training the classifier. In such a case, the data is considered to have a sensitive attribute (such as gender or race), and the objective is to train a classifier that reduces bias with respect to some outcome, represented by labels in the data (such as a decision to hire or promote an individual).

This section reviews existing work on fairness pertaining to machine learning, looking at interventions that are designed to reduce bias that take place before, during and after training. The emphasis is on before training, as wrangling falls there.

Before Approaches to reducing bias that influence the data used for training may remove attributes that correlate with the sensitive attribute, change the labels to reduce bias, choose samples of

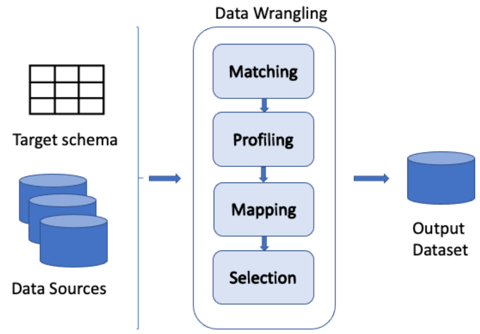

Figure 1: Data wrangling pipeline

the available data, or assign weights to different rows that impact their influence on the classifier [7]. Alternatively, pre-processing techniques may directly tackle properties of the underlying data set that may be problematic, such as coverage [10].

There has been less work on data wrangling, the focus of this paper; although some papers explicitly discuss data preparation, such work has generally been later in the pipeline than this paper. Valentim *et al.* [15] empirically compare several interventions, specifically the removal of the sensitive attribute, feature discretisation and sampling techniques. Accinelli *et al.* [1] consider coverage within data preparation, though primarily by selecting data rather than determining how the data should be wrangled.

In our previous work [12], we steer the generation of data preparation plans in ways that reflect risk factors for bias. The current paper follows a similar approach, in the sense that plan generation is steered, but directly calls the classifier on candidate plans, thereby steering the plan generation based on direct evidence of bias.

During Approaches to reducing bias within the learning algorithm are often associated with metrics that quantify some notion of bias, and the learning algorithm then explicitly trades off bias with accuracy [13]. For example, Hajian *et al.* [5] provides techniques for constraining the levels of bias exhibited by classification rules to a specified level.

After Approaches to reducing bias after learning operate by selectively changing labels at prediction time (e.g., [6, 8]).

Work on bias-aware data preparation can be seen as complementary to most other work on fair classification; benefits that arise with respect to fairness in upstream processes simply reduce the scale of the problem for later interventions.

3 TECHNICAL BACKGROUND

The approach to fair data preparation depends on identifying interventions during data wrangling that affect the behaviour of a wrangling pipeline. We build on the VADA data wrangling system [9], the relevant steps of which are illustrated in Figure 1. In VADA, given a target schema definition and the data sources, the system can generate schema mappings that populate the target. We particularly exploit the automatic generation of data preparation plans, to generate new ways of integrating the data. The steps that are relevant to this paper are:

Matching: identify relationships between source and target attributes, where the former may be suitable for providing data for the latter.

Profiling: identify candidate keys for the source relations, and (partial) inclusion dependencies that identify overlaps between attribute values in different sources.

Mapping: informed by the results of matching and profiling, generate candidate mappings in the form of views that can be used to populate the target table from the sources [11].

Selection: given the results of the mappings, select the top k results from the mappings that best satisfy some criterion. In this paper, the criterion used is *completeness* – prefer the mappings with the fewest missing values.

4 APPROACH

4.1 Overview of Approach

The approach to fair data preparation for classification involves exploring a space of candidate data preparation plans, assessing the fairness of the classifiers built using the results of these plans. The steps in the approach are illustrated in Figure 2, and are discussed below.

In the approach, a *Candidate Solution* captures a collection of interventions into the wrangling process that are deemed to be helpful for reducing bias. Specifically, the interventions involve removing matches or inclusion dependencies from consideration during data wrangling. For example, assume that m_1 is a match and i_1 is an inclusion dependency. It is possible that the data resulting from the wrangling pipeline in Figure 1 produces a fairer output data set when: *(i)* m_1 is available but i_1 is not; *(ii)* m_1 is not available but i_1 is available; *(iii)* m_1 and i_1 are both available; or *(iv)* neither m_1 nor i_1 are available. These different environments in which to wrangle constitute a search space of interventions that lead to the production of different data sets that may lead to more or less biased classifiers.

Assume we have a match $m : S.a \rightarrow T.a'$ relating attribute a from source S to attribute a' in target T. The removal of m from the wrangling process can have one of the following effects: *(i)* m is replaced by another match involving the same table, so that now $S.a'' \rightarrow T.a'$ and as a result $T.a'$ is populated differently; *(ii)* m is replaced by another match involving a different table, so that now $P.a \rightarrow T.a'$ and as a result $T.a'$ is populated differently; *(iii)* no alternative match is found, so $T'.a$ is populated with *null* values.

Assume we have an inclusion dependency $I = P.a \subseteq_\theta S.b$ between attributes in the sources P and S, where θ is the degree of overlap between the attributes a and b. The removal of I can have one of the following effects: *(i)* there is another inclusion dependency between P and S, which leads to the same tables being considered for joining, but on different attributes; *(ii)* there is another indirect way in which P and S can be joined through an intermediate table that can be used instead by the mapping generator; or *(iii)* the tables are no longer joined with each other.

To consider different ways of preparing the data, the approach is to explore the space of possible interventions, with a view to identifying combinations that lead to less biased results. The space of alternative interventions is explored using Tabu search [4], a local search algorithm that employs a diversity heuristic to avoid becoming stuck in local optima. When describing the approach, we will periodically refer to a running example.

Example 1. *Assume an example where a model is trained to predict which individuals will be hired, with the sensitive binary attribute gender, which can have the values male or female. Consider a target schema* $T(name, age, qualification, gender, hire)$, *which*

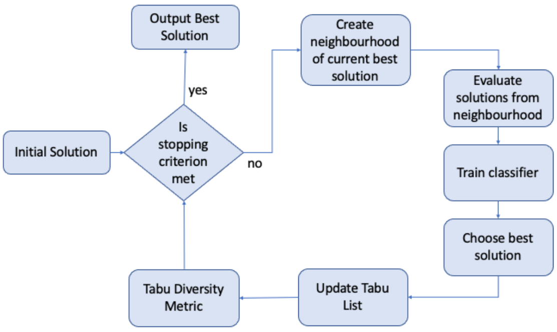

Figure 2: Steps in approach.

indicates the hiring decision of the person with the given name, age, qualification and gender.

4.2 The Steps

In this section, each of the steps in Figure 2 are described, in turn.

Stopping condition. Each iteration of the search uses the data wrangling pipeline in Figure 1 to create new data preparation plans, the data from which is used to train a classifier. As such, each iteration can be considered to be quite expensive, and certainly more expensive than evaluating a typical fitness function. As a result, the search cannot be allowed to carry out arbitrary numbers of iterations. The search terminates with its best solution when it has completed a specified number of wrangles (the number of executions of the data wrangling pipeline in Figure 1 during the search from Figure 2). The number is an input parameter that is usually chosen by considering the trade-off between runtime and bias reduction (an analysis is reported in our previous work in [12]).

Create neighborhood. A *Candidate Solution* is a set of interventions. An *intervention* is a match or inclusion dependency that is to be excluded from consideration when generating an integration using the wrangling pipeline in Figure 1. The *Current Candidate Solution* is the starting point for the creation of a neighbourhood. The neighbourhood is a set of *Candidate Solution*s, each of which is obtained by associating the *Current Candidate Solution* with a single *intervention* from a set of *Candidate Interventions.* The *Candidate Interventions* are the matches and inclusion dependencies used in the mappings created by running the wrangling pipeline in the context of the interventions from *Current Candidate Solution.*

Example 2. *Following on from Example 1, assume that there are sources $S_1(name, age, height)$ and $S_2(name, hire)$ and that there are matches m_1 from $S_1.name$ to $T.name$, m_2 from $S_1.age$ to $T.age$, m_3 from $S_2.name$ to $T.name$, m_4 from $S_2.hire$ to $T.hire$. Assume that a view has been produced for populating the target that contains the join $S_1 \bowtie_{S_1.name=S_2.name} S_2$, informed by an inclusion dependency i_1 between $S_1.name$ and $S_2.name$. In addition, assume that the Candidate Interventions are $\{m_1, m_2, m_3, m_4, i_1\}$. Further, assume that the Current Candidate is $\{m_1\}$. Then the neighbourhood will consist of the Candidate Solutions $\{m_1, m_2\}$, $\{m_1, m_3\}$, $\{m_1, m_4\}$ and $\{m_1, i_1\}$.*

Evaluate candidate solutions. Each of the *Candidate Solutions* in the neighbourhood in turn are passed to the data wrangling pipeline in Figure 1. The pipeline generates mappings, using automatically-derived matches and inclusion dependencies, but excluding those from the interventions of the *Candidate Solution*, thereby associating each *Candidate Solution* with a data set that populates the target schema. The resulting mappings are evaluated, thereby associating each *Candidate Solution* with a data set that populates the target.

Train classifier. Having generated a data set for each candidate solution, the next step is to train a classifier on each of the data sets, and compute the bias of each of the resulting classifiers. The overall approach is independent of the type of classifier used, but in the experiments we use the J48 implementation of the C4.5 decision tree learning algorithm [14].

For each data set, we carry out k-fold cross validation. As such, the classifier is trained k times using $k-1$ folds and evaluated on the remaining fold, each time with a different evaluation fold. Then, its fairness is assessed by averaging the bias for the evaluation folds. The overall approach is independent of the notion of bias used, but in the experiments we use *demographic parity* [3]. Demographic parity measures group fairness. Where the *Positive Rate* (PR) for a group is the fraction of the items in the group that have the given outcome (e.g., the fraction of the people in the group that are hired), the demographic parity for a fold i is computed as:

$$dp_i = abs(PR_i(G_1) - PR_i(G_2)) \tag{1}$$

where G_1 is one group (e.g., females) and G_2 is the other (e.g., males). Where $dp = 0$, there is no bias.

The overall demographic parity for a Candidate Solution is computed as:

$$dp_{sol} = \frac{\sum_{i=1}^{i=k} dp_i}{k} \tag{2}$$

Note that the role of training during the execution of the steps in Figure 2 is to obtain evidence to inform the search for fairer plans, and not to produce a final classifier.

Choose best solution. The best solution should have low bias. However, as the approach explores a space of interventions that are liable to lead to increasingly sparse results (due to the deletion of matches and inclusion dependencies), it is important not to end up with a solution that has low bias but poor accuracy as a result of the provision of sparse training data. So, to prefer solutions that retain more data that can be used as evidence by the classifier, the objective function for the search is:

$$obj = (w_1 * dp_{sol}) + (w_2 * r_{nulls}) \tag{3}$$

where r_{nulls} is the ratio of nulls in the data set, and w_1 and w_2 are weights, which are both set to 0.5 in the experiments, i.e., *completeness* and *fairness* are equally important in the end result. The search thus prefers solutions with low bias and high completeness.

Update Tabu list. The Tabu search maintains a collection of points in the search space that have already been visited. This list is updated here to include all solutions that have been evaluated in the current iteration.

The Tabu List is also used during the Create Neighborhood step, to avoid considering parts of the search space that have been explored before.

Example 3. *Following on from Example 2, assume that the current tabu list is $[\{m_1\}]$. Then, after the exploration of the Candidate Solutions in the neighbourhood, i.e., $\{m_1, m_2\}$, $\{m_1, m_3\}$, $\{m_1, m_4\}$*

and $\{m_1, i_1\}$, these will be added to the Tabu list so that the search is optimized by keeping track of the plans that have been explored, and thus that need not be explored again.

Tabu diversity mechanism. Tabu search is essentially a greedy local search algorithm, with a single *Current Candidate Solution*, which is normally replaced at each iteration by the best solution in its neighbourhood in accordance with the objective function. To avoid being trapped in local optima, a diversity mechanism can be used to jump to a new location in the search space. In our implementation of Tabu, when the best solution has not been improved for several iterations, the new *Best Candidate Solution* is set to the highest scoring of a subset of plans from the Tabu List.

5 EVALUATION

5.1 Experimental Setup

Data Sets. To investigate the effectiveness of the approach described in Section 4, we have evaluated it in the context of data preparation scenarios derived from the following benchmark data sets:

- *German Credit*[1]*:* The role of the classifier is to determine if it is risky or not to give credit to a person. The dataset contains 1000 tuples, each with 21 attributes. The sensitive attribute is *Gender*, which contains the values *male* (69%) and *female* (*31%*). The class attribute is *Risky*, which contains the values *Yes* or *No*, stating if it is risky to give someone a loan or not, respectively.
- *Adult Census*[2]*:* The role of the classifier is to predict if a person makes over *\$50K* a year. The dataset contains 10,000 tuples[3], each with 14 attributes. The sensitive attribute is *Gender*, which contains the values *male* (66%) and *female* (*33%*). The class attribute is *Salary*, which contains the values $> 50K$ and $<= 50K$, stating if someone is predicted to earn more than $50K$ *per year* or less than that.

These are benchmark data sets, that have been used in other studies on fairness, e.g., [15–18]. However, each is supplied as a single table, whereas to experiment on data wrangling, several tables are required. So, for the experiment, each of the data sets is partitioned horizontally into groups of rows, which are then vertically partitioned. This subdivision of the tables creates different ways in which the target can be populated, and allows interventions to take place that apply to subsets of the data.

For *Adult Census*, we horizontally divided the dataset using the contained 42 *country* values, which we then divided into 2 vertical partitions, thus, amounting to 84 initial sources. For *German Credit*, we divided the initial dataset into 4 horizontal partitions, which were then divided each into 3 vertical partitions, resulting 12 initial sources.

In the experiments, the number of sources is varied; adding additional sources provides more ways of populating the target, and in turn more potential interventions to improve fairness. In order to create opportunities for different intervention plans, we created *alternative* sources with synthetic constant values for non-sensitive attributes. Thus, for each benchmark dataset, we created five scenarios, each with an increase of 25% *alternative* sources over the previous scenario, e.g., for *Adult Credit*, we created five scenarios with 84, 105, 127, 147, and 168 sources, each with a 0, 25, 50, 75, 100% increase over the initial number of sources (84). The *target* schema is that of the original benchmark data sets.

Configuring the search. There are a number of parameters that control the behaviour of the search. The maximum number of wrangles for both data sets is set to *80*; this is to obtain manageable experiment run times, while also allowing enough searching to generate plausible results. The value was obtained after a sensitivity analysis. In addition, up to the top 130 mappings are obtained from mapping generation; in practice, this is more than sufficient for the scenarios used, and often the mapping generator will produce significantly fewer mappings than this.

In the *German Credit* data set, the maximum number of tuples in the result of the wrangling process is set to *620* (the original data set contained *1,000* rows). This is to ensure that the results of the fair wrangling can be selective, leaving out results that are associated with more bias. For the same reasons, the maximum number of tuples in the *Adult Census* dataset is *6,616* out of *10,000.*

The results of the classifiers are obtained using 5-fold cross validation.

Classifier. The reported results are for the *J48 decision tree* classifier from the Weka[4]. The experiments were run with other classifiers, such as *Logistic Regression*, obtaining negligible variation between the experimental results, thus, we report only one set of results. Although the experiments involve binary classification tasks, the overall approach does not depend upon this.

Baseline. The baseline in the experiments is obtained by running the wrangling process without the approach from Section 4. Thus we are able to obtain a direct indication of the effectiveness of the interventions made. Note that there is no direct competitor in the literature with which to compare.

Experimental setup. The experiments were run over an Intel Core i5 with 2×2.7 GHz, and 8 GB RAM.

5.2 Results

German Credit Data. Figures 3 and 4 show the results of the experiment with the German Credit data set. For both figures, the horizontal axis reports the number of sources in each scenario. In Figure 3, the vertical axis represents the demographic parity bias (computed using Equation 2), while, in Figure 4, the vertical axis represents the accuracy of the classifier trained and evaluated on the output datasets. Both bias and accuracy represent the averages computed through cross validation.The following can be observed: *(i)* The fair wrangling approach performs as well as or better than the baseline for demographic parity in all scenarios. *(ii)* Both methods sometimes have the same demographic parity for different numbers of sources; this is because the same plan may be selected even though the number of available sources grows. *(iii)* The fairness interventions do not always provide improving results as more sources become available; this is because the search is not exhaustive, and thus, although the larger numbers of sources provide more opportunities for effective interventions to be discovered, there is no guarantee that the best intervention plans will be obtained. As mentioned in Section 5.1, the parameter for the amount of explored search space was set based on a sensitivity analysis similar to the one reported for

[1] https://www.kaggle.com/uciml/german-credit

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] Due to limited resources, we used stratified sampling to choose 10,000 tuples out of the 48,842 tuples in the original dataset. The used sample maintains the properties of the initial dataset.

[4] https://www.cs.waikato.ac.nz/ml/weka/

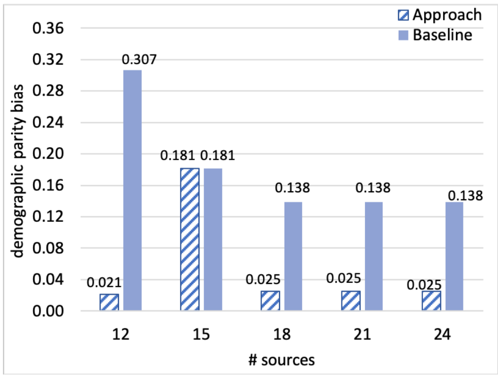

Figure 3: Demographic parity for the German Credit data set for different numbers of sources.

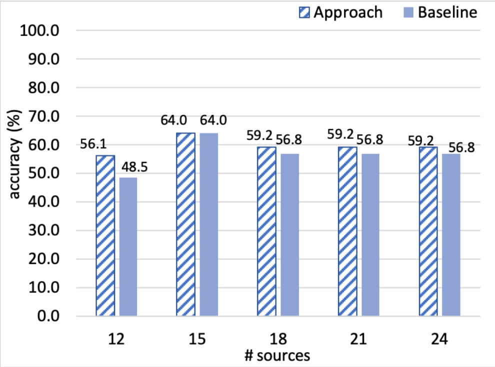

Figure 4: Classifier accuracy for the German Credit data set for different numbers of sources.

the experiments presented in our previous work in [12]. Furthermore, the additional sources may not always provide good opportunities to increase fairness. *(iv)* Accuracy did not suffer as a result of the interventions.

Adult Census Data. Figures 5 and 6 show the results of the experiment with the *Adult Census* data set. The following can be observed: *(i)* The results are similar to those for the *German Credit* data set; we do not repeat the recurring explanations here; *(ii)* The overall bias levels are smaller for the *Adult Census* data set than for the *German Credit* one, and in several cases bias is almost completely removed. *(iii)* The accuracy for the *Adult Census* data set is higher than that in the *German Credit* scenarios. This may be due to the difference in the amount of data that is used to train the classifier, i.e., *Adult Census* contains significantly more rows than that *German Credit* training set, which may lead to an underfit model for the *German Credit* scenarios.

Dataset	Runtime (in seconds) for scenarios with percent of synthetic sources				
	0%	25%	50%	75%	100%
German Credit	427	562	780	1,002	1,309
Adult Census	1,416	48	1,691	1,605	41

Table 1: Runtime for varying number of sources on each dataset

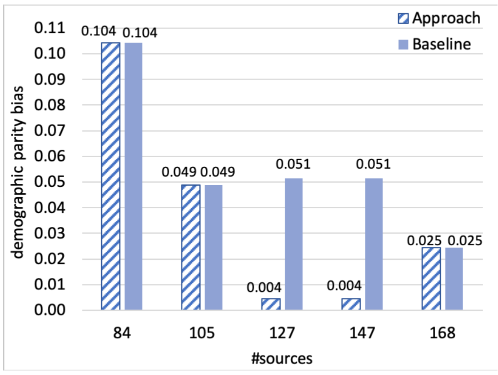

Figure 5: Demographic parity for the Adult Census data set for different numbers of sources.

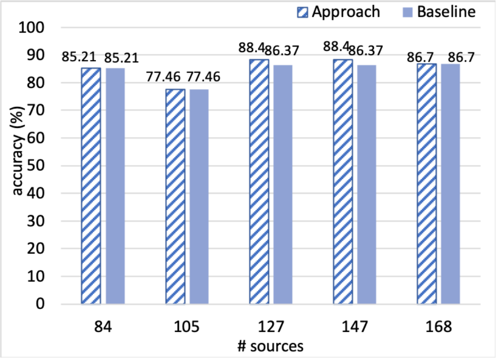

Figure 6: Classifier accuracy for the Adult Census data set for different numbers of sources.

Runtime. Table 1 shows the runtime for each dataset considering their different scenarios, i.e., with varying percents of synthetic data sources (as explained in Section 5.1). The runtime is expressed in seconds. Note that for the same percent, the number of sources for each dataset differs, e.g., for the 50% case, the *German Credit* dataset scenario contains 18 sources, while the *Adult Census* scenario contains 127 sources.

For the *German Credit* dataset, the runtime increases as the number of sources increases. This is due to the fact that *i)* the Tabu search space increases and there are more intervention plans to be explored, and *ii)* the wrangling process is longer in the cases with more sources because each of the wrangling components that are described in Section 3 depends on the number of input data sources, e.g., the mapping component considers combinations between all subsets of sources, thus, the number of subsets increases with the number of initial sources [11].

For the *Adult Census* dataset, it is interesting to notice that there is no clear pattern in the runtime values. This is due to the fact that for some of the scenarios, e.g., 25% and 100%, the search ends if the bias is below a set threshold. This threshold is set as, in the case of *demographic parity*, one cannot expect to have equal ratios for the *positive* labels for *male* and *female* values (in the sensitive attributes). Thus, we need to consider an acceptable difference. In these scenarios, we set the threshold at 0.05, i.e., $dp_{sol} < 0.05$ (see Equation 2). This threshold is set according to the used fairness measure, e.g., *demographic parity* in our case. For the Adult Census scenarios with 25% and 100% of synthetic sources, in both cases, the wangling process stops after the initial wrangle as it detects that the output dataset presents an *acceptable* bias value, thus, it does not prepare more interventions from the initial wrangle.

6 CONCLUSIONS

Our previous work tackled underlying causes of bias for unlabelled data [12]. This paper complements that work for labelled data, by wrangling in a way that directly responds to the bias observed when wrangling with different mappings. This work focuses on creating fair datasets by using *demographic parity* as a way to measure the bias. However, other metrics may be investigated, e.g., *equalized odds*, *equal opportunity,* etc., depending on the properties that are desired upon the output dataset.

We now revisit the claimed contributions from the introduction:

(1) *The proposal of a strategy for fairness-aware data preparation for classification.* The proposed strategy searches a space of interventions that impact on how data is integrated for analysis. The approach builds on the ability to automate aspects of the data preparation process. Interventions are carried out, integrations generated, and the resulting data is used to train classifiers. The classifiers are then tested for bias, and the most promising interventions investigated further.

(2) *The realisation of the strategy as a search for fair data preparation plans.* The strategy has been implemented as a Tabu search, over a space of interventions that include the exclusion of matches and inclusion dependencies. Removing selected matches and inclusion dependencies may reduce bias by removing problematic mappings, or by prioritising alternative ways of preparing the data.

(3) *An evaluation of (2) for benchmark data sets that shows how the interventions can improve a specific fairness metric, namely demographic parity. Also, we show how the accuracy of the trained classifier is impacted by the interventions.* Results using two evaluation scenarios show that the approach reduces bias in comparison with a non fairness-aware base case. Also, the evaluation shows that, compared to the baseline case, the accuracy of the classifier trained as a result of the interventions did not suffer.

ACKNOWLEDGMENTS

We are pleased to acknowledge the support of the UK Engineering and Physical Sciences Research Council, through the VADA Programme Grant (EP/M025268/1).

REFERENCES

[1] Chiara Accinelli, Simone Minisi, and Barbara Catania. 2020. Coverage-based Rewriting for Data Preparation. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020 (CEUR Workshop Proceedings, Vol. 2578)*, Alexandra Poulovassilis et al. (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2578/PIE4.pdf

[2] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *FAT**. ACM, 329–338. https://doi.org/10.1145/3287560.3287589

[3] Pratik Gajane. 2017. On formalizing fairness in prediction with machine learning. *CoRR* abs/1710.03184 (2017). arXiv:1710.03184 http://arxiv.org/abs/1710.03184

[4] Fred Glover. 1986. Future paths for integer programming and links to artificial intelligence. *Computers Operations Research* 13, 5 (1986), 533 – 549. https://doi.org/10.1016/0305-0548(86)90048-1 Applications of Integer Programming.

[5] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Trans. Knowl. Data Eng.* 25, 7 (2013), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

[6] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS* (Barcelona, Spain). Red Hook, NY, USA, 3323–3331.

[7] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *KAIS* 33 (2011), 1–33.

[8] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *ICDM*. USA, 924–929. https://doi.org/10.1109/ICDM.2012.45

[9] Nikolaos Konstantinou, Edward Abel, Luigi Bellomarini, Alex Bogatu, Cristina Civili, Endri Irfanie, Martin Koehler, Lacramioara Mazilu, Emanuel Sallinger, Alvaro Fernandes, Georg Gottlob, John Keane, and Norman Paton. 2019. VADA: An Architecture for End User Informed Data Preparation. *Journal of Big Data* 6:74 (2019). https://doi.org/10.1186/s40537-019-0237-9

[10] Yin Lin, Yifan Guan, Abolfazl Asudeh, and H. V. Jagadish. 2020. Identifying Insufficient Data Coverage in Databases with Multiple Relations. *Proc. VLDB Endow.* 13, 11 (2020), 2229–2242. http://www.vldb.org/pvldb/vol13/p2229-lin.pdf

[11] Lacramioara Mazilu, Norman W. Paton, Alvaro A.A. Fernandes, and Martin Koehler. 2019. Dynamap: Schema Mapping Generation in the Wild. In *Proceedings of the 31st International Conference on Scientific and Statistical Database Management* (Santa Cruz, CA, USA) *(SSDBM '19)*. Association for Computing Machinery, New York, NY, USA, 37–48. https://doi.org/10.1145/3335783.3335785

[12] Lacramioara Mazilu, Norman W. Paton, Nikolaos Konstantinou, and Alvaro A. A. Fernandes. 2020. Fairness in Data Wrangling. In *21st International Conference on Information Reuse and Integration for Data Science, IRI 2020, Las Vegas, NV, USA, August 11-13, 2020*. IEEE, 341–348. https://doi.org/10.1109/IRI49571.2020.00056

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *CoRR* abs/1908.09635 (2019). arXiv:1908.09635 http://arxiv.org/abs/1908.09635

[14] J. Ross Quinlan. 1996. Bagging, Boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI*, William J. Clancey and Daniel S. Weld (Eds.). AAAI Press / The MIT Press, 725–730. http://www.aaai.org/Library/AAAI/1996/aaai96-108.php

[15] Inês Valentim, Nuno Lourenço, and Nuno Antunes. 2019. The Impact of Data Preparation on the Fairness of Software Systems. *CoRR* abs/1910.02321 (2019). arXiv:1910.02321 http://arxiv.org/abs/1910.02321

[16] Vladimiro Zelaya, Paolo Missier, and Dennis Prangle. 2019. Parametrised Data Sampling for Fairness Optimisation. In *KDD XAI*.

[17] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML* (Atlanta, GA, USA) *(ICML'13)*. III–325–III–333.

[18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AAAI* (New Orleans, LA, USA). NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779