Discrimination-aware Data Transformations

Chiara Accinelli

Advised by: prof. Barbara Catania
University of Genoa, Italy

Abstract
A deep use of people-related data in automated decision processes might lead to an amplification of inequities already implicit in the real world data. Nowadays, the development of technological solutions satisfying nondiscriminatory requirements is therefore one of the main challenges for the data management and data analytics communities. Nondiscrimination can be characterized in terms of different properties, like fairness, diversity, and coverage, and many approaches have been proposed so far for guaranteeing nondiscrimination through the satisfaction of such properties during specific steps of the data processing pipeline. In this PhD project, we are interested in investigating the impact of coverage-based constraints on data transformations. Coverage aims at guaranteeing that the input dataset includes enough examples for each (protected) category of interest, thus increasing diversity with the aim of limiting the introduction of bias during the next analytical steps. We propose *coverage-based queries* as a mean to achieve coverage constraint satisfaction on the result of data transformations defined in terms of selection-based queries. Both precise and approximate algorithms are designed to guarantee a good compromise between efficiency and accuracy. The applicability of the approach is evaluated by integrating it in a data processing Python toolkit.

Keywords
nondiscrimination, data transformation, coverage, rewriting

1. Introduction

Nowadays, we are surrounded by data that are increasingly exploited to make decisions that might impact people's lives. It is therefore very important to understand the nature of that impact at the social level and take responsibility for them. The design of data-driven decision-support systems ensuring a *responsible* and *ethical* use of data is therefore a must and it has been recognized that both data management and data analytic communities should contribute [1, 21]. Such systems should ensure on one hand *transparency* and *interpretability*, making the process and the decisions easy to understand, and on the other *nondiscrimination* with respect to all the reference groups of individuals, usually defined in terms of sensitive attributes, like, e.g., gender.

Nondiscrimination can be characterized in terms of different properties like *fairness*, i.e., lack of bias [16], *diversity*, i.e., the degree to which different kinds of objects are represented in a dataset [11], and *coverage* [7], guaranteeing a sufficient representation of any category of interest in a dataset. As first pointed out in [1] and remarked in, e.g., [11, 21], such properties should be achieved through a holistic approach, incrementally enforcing nondiscrimination constraints along all the stages of the data processing life-cycle, through individually independent choices rather than as a constraint on the final result: the sooner you spot the problem fewer problems you will get in the last analytical steps of the chain (see, e.g., the Google's gorilla classification incident [20]).

In this PhD project, we are interested in investigating the impact of coverage-based constraints on the, possibly intermediate, datasets generated through data preparation, with a special focus on data transformations. This topic is relevant since any data preparation step that transforms the input datasets might lead to a violation of the coverage of protected categories, affecting subsequent analytical tasks. Notice that the input dataset can correspond to either raw data that have not been transformed yet (and in this case, solutions like those proposed in [7, 8] can be used to determine how to modify the input dataset and collect new data) or the result of, potentially many, data transformation queries. We are interested in this second case.

As an example, suppose you are interested in analyzing data of the well known *Adult* dataset[1] (e.g., predicting through classification which individuals make over 50k a year), after filtering it according to specific criteria (e.g., only senior job positions should be considered, qualified in terms of selection conditions over age, weekly working hours, and education level). Suppose you would like to guarantee nondiscrimination with respect to the gender by training a model whose accuracy does not deeply depend on this attribute. It has already been recognized that the quality of the classifier might depend, among the others, also on the number of instances, i.e., the coverage, of each group in the dataset [18]. Thus, if the selection query returning senior job positions includes few female,

Published in the Workshop Proceedings of the EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023, Ioannina, Greece)
chiara.accinelli@dibris.unige.it (C. Accinelli)
0000-0002-6626-9470 (C. Accinelli)

[1] https://archive.ics.uci.edu/ml/datasets/Adult

the result of the classifier can be biased.

In order to solve this problem without going back to the data collection step, additional female individuals could be added to the dataset generated through query execution.

We tackled this issue by defining and processing *coverage-based queries*, i.e., selection-based queries that, given a set of coverage constraints, always return a result satisfying the input constraints while staying close to the original request. In order to avoid disparate treatment discrimination [9] and guarantee transparency, the initial query is rewritten into a new one, satisfying the coverage constraint while staying close to the original request.

The main research questions addressed by the PhD project can be summarized as follows:
(RQ1) How can coverage-based queries be defined and how can they be characterized?
(RQ2) How can coverage-based queries be efficiently processed?
(RQ3) How can coverage-based queries be integrated in data processing environments?

The remainder of the paper is organized as follows. Section 2 compares our work with other existing approaches. Coverage-based queries are defined in Section 3 (RQ1) and solutions developed so far for their processing are described in Section 4 (RQ2) (some preliminary results on (RQ1) and (RQ2) can be found in [4, 5, 6]). Details about a Python data processing toolkit integrating the proposed techniques in Pandas[2] (RQ3) can be found in [3] but are not presented in the paper for space constraints. Finally, Section 5 concludes and presents some directions for further developments.

2. Related work

Discrimination-aware approaches have been proposed both with reference to data analysis (e.g., OLAP queries [15], set selection [22], ranking [23]) and data preparation (e.g, dataset repair during data acquisition [7, 16], with a special focus on coverage in [7, 8, 12], data cleaning [17] and data integration [14]).

Similarly to [7, 8, 12], we consider coverage as a mean to limit discrimination. However, rather than checking coverage over raw datasets and repair them in case of coverage unsatisfaction through new data acquisitions, we guarantee coverage satisfaction along data transformation chains defined in terms of selection-based queries.

Coverage-based queries, presented in this paper, change the result of an input query through rewriting rather than through the usage of ad-hoc query execution algorithms. This avoids a disparate treatment discrimination during selection-based query execution since a well defined criteria, i.e., a new selection-condition, is used to retrieve the new result, guaranteeing at the same time transparency. In this respect, coverage-based queries differ from other similarity-based query approaches, like fuzzy queries [13]. Other rewriting-based approaches have been proposed so far to tackle discrimination issues defined in terms of other properties and queries. Rewriting has been used for OLAP queries and causal fairness in [15] and, more recently, for range queries and fairness in [19]. As far as we know and according to [18], no other solutions addressing coverage-based rewriting in the context of selection-based queries have been proposed so far.

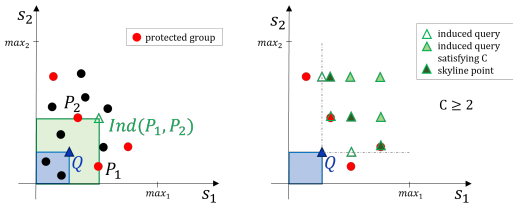
3. Coverage-based queries

Preliminaries. We consider data stored in *tabular datasets* (e.g., relations in a relational database, data frames in the Pandas environment). We assume that some discrete valued attributes $\mathcal{S} = S_1, ..., S_n$ of the input dataset are of particular interest (e.g., `gender` and `race`) since they identify protected groups and are called *sensitive attributes*. We focus on *selection-based data transformations* (or queries) over stored or computed (e.g, joined or aggregated) datasets, in analytical processes that might alter the representation (i.e., the coverage) of specific groups of interests, defined in terms of sensitive attribute values (e.g., SQL selections over relational data, data slicing operations in Pandas,[2] ColumnTransformers in Scikit-Learn).[3]

We consider boolean combinations of atomic selection conditions $sel_i \equiv A_i \theta v_i$, $v_i \in D_{A_i}$, $\theta \in \{=, <, \leq, \geq >\}$, A_i numeric attribute, $A_i \neq A_j$ $i, j = 1, \ldots, d$, that do not refer to, as usually assumed, sensitive attributes, i.e., $A_i \notin \mathcal{S}$. A selection-based query Q is thus denoted by $Q\langle v_1, ..., v_d\rangle$ or $Q\langle \overline{v}\rangle$, $\overline{v} \equiv (v_1, ..., v_d)$, and $\overline{v}$ is called *selection vector*. A *coverage constraint* has the form $\downarrow^{S_{l_1},\ldots,S_{l_h}}_{s_{l_1},\ldots,s_{l_h}} \geq k$ and specifies that the minimum number of instances with sensitive attribute S_{l_i} equal to s_{l_i}, $i = 1, ..., h$, in a query result has to be k. As an example, $\downarrow^{\texttt{gender}}_{\texttt{female}} \geq 10$ specifies that the result should include at least 10 female individuals. The group referred by a coverage constraint is called *protected group*.
Definition of coverage-based queries. Let C be a set of coverage constraints over a set of sensitive attributes S and $Q\langle \overline{v}\rangle$ a selection-based query. A *coverage-based query* ξ^C_Q for C and $Q\langle \overline{v}\rangle$ is a selection-based query that, given a dataset I, stretches the result $Q\langle \overline{v}\rangle(I)$ as little as possible so that the result satisfies the constraints in C. More precisely, when considering a dataset I: *(i)* ξ^C_Q returns the result of a query $Q\langle \overline{u}\rangle$ over I; $Q\langle \overline{u}\rangle$ is obtained from Q by only changing the selection constants that

[2] https://pandas.pydata.org/pandas-docs/stable/getting_started/intro_tutorials/03_subset_data.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html



(a) Input and induced queries (b) Skyline points

Figure 1: Coverage-based query properties

might depend on I; *(ii)* $\forall I\ Q\langle\overline{v}\rangle(I) \subseteq \xi_Q^C(I)$; *(iii)* all coverage constraints are satisfied by $\xi_Q^C(I)$; *(iv)* $Q\langle\overline{u}\rangle$ is *minimal*. Minimality means that any other query Q' satisfying conditions *(i)–(iii)* is such that either $card(Q'(I)) > card(Q\langle\overline{u}\rangle(I))$ or $card(Q'(I)) = card(Q\langle\overline{u}\rangle(I))$ and $Q\langle\overline{u}\rangle$ is syntactically closer than Q' to $Q\langle\overline{v}\rangle$, according to the Euclidean distance (defined in a unit space) between selection vectors.

Properties. A coverage-based query ξ_Q^C satisfies the following properties:

(*P1*) It can be represented in a *canonical form* in which each selection condition has the form $A_i \leq v_i$ or $A_i < v_i$; $Q\langle\overline{u}\rangle$ is then represented as point $\overline{u}$ in the d-dimensional space defined by selection attributes (see Q in Figure 1(a)).

(*P2*) Let $\xi_Q^C(I) \equiv Q\langle\overline{u}\rangle(I)$. We proved that $\overline{u}$ coincides with the upper right vertex of the minimum bounding box of at most d distinct points in I, Q, and the origin of the space (see the green triangle in Figure 1(a)) [2]. Such vertices are called *induced points* and the set of all induced points corresponds to the search space for coverage-based queries.

(*P3*) There is a relationship between $\overline{u}$ and the skyline of induced points corresponding to queries that, when executed over I, satisfy C; the dominance relation, needed for the skyline computation, is defined over selection attributes, assuming the lower the better (see Figure 1(b)). It can be proved that $\overline{u}$ coincides with the skyline point corresponding to the query with the minimal cardinality at the lowest distance from Q. Thus, $\overline{u}$ can be identified by combining skyline and top-1 computations (possibly mixed, as pointed out in [10]).

4. Coverage-based query processing

Properties P1, P2, and P3 suggest a naïve but inefficient approach for processing coverage-based queries, due to the size of the search space and skyline computation. We therefore improved such basic strategy under two directions, briefly described in the following. The designed algorithms, for each I, return one minimal solution[4] and do not rely on any index data structure, so that both stored and computed datasets can be considered.

A grid-based approximate approach. The first approach is *approximate* because, for each dataset I, it relies on a discretized search space and a sample-based approach for cardinality estimation, needed for constraint and minimality checking (property P3). It can be applied over any dataset for which a sample is available or can be easily computed on the fly.

The discretized search space is generated by considering the intersection points of a grid obtained by discretizing each axis (one for each selection attribute in Q, from the corresponding selection value in the query to the maximum value in the dataset), using standard binning approaches (e.g., equi-width and equi-depth). Each point on the grid corresponds to a selection-based query of type $Q\langle\overline{v}\rangle$, thus satisfying conditions *(i)* and *(ii)* of the reference problem. $Q\langle\overline{u}\rangle$ is then determined by visiting the discretized search space starting from Q, one point after the other, at increasing distance from Q (algorithm $CRBase$). The properties of the discretized search space and the canonical form are considered for pruning the space (algorithm $CRBaseP$), possibly increasing the number of points to be visited at different iterations (algorithms $CRBaseI$ and $CRBaseIP$).

Details on all the algorithm versions and an exhaustive experimental evaluation, on both synthetic and real-world datasets, have been presented in [5]. The obtained results depend on the density of the search space and show that: *(i)* equi-depth guarantees better performance over non-uniformly distributed data; *(ii)* a multi-level processing approach, like $CRBaseI$, greatly helps in reducing the curse of dimensionality when the query contains a high number of selection conditions; *(iii)* the processing performance linearly depends on the number of coverage constraints; *(iv)* a good level of accuracy can be obtained with relatively small samples; *(v)* coverage constraint satisfaction has an obvious impact on the rate of different groups of protected instances, i.e., on fairness.

An iteration-based precise approach. More recently, we started from the naïve approach, derived from P1, P2, and P3, to design a family of algorithms for the precise computation of coverage-based queries. The designed algorithms rely on the following considerations: (i) the induced query space can be computed in up to d iterations; the computation of new points at iteration i can be pruned by considering only points obtained at iteration $i-1$ that do not satisfy C; (ii) the iterated computation of induced points and skyline dominance checks can be interleaved so that the considered space at each iteration is further reduced; (iii) minimality can be checked either

[4] The proposed algorithms can be easily customized to return all minimal solutions or a specific one, according to some further optimality criteria.

during the skyline computation, to reduce the number of dominance comparisons, or after the skyline has been computed, limiting in this way the number of cardinality estimations; (iv) the grid-based approximate approach can be used as a filtering step, for further reducing the space before applying one precise algorithm. The proposed algorithms are currently under evaluation, on both synthetic and real datasets.

5. Conclusions and further developments

In this PhD project, we investigate the impact of coverage constraints on data transformations, as a mean for limiting bias in the next analytical steps. After defining coverage-based queries, we designed and experimentally evaluated both approximate and precise algorithms for their processing. The proposed solutions rely on query rewriting, a key approach for enforcing specific nondiscrimination constraints while guaranteeing transparency and avoiding disparate treatment discrimination.

Future work includes the integration of the proposed queries in a relational DBMS and the extension of the proposed solutions to consider further nondiscrimination constraints. To this aim, an interesting approach is to rely on a constraint-based optimization approach for specifying different types of constraints, possibly inherently different, as coverage and fairness [19], and determining the best data transformation rewriting.

References

[1] S. Abiteboul et al. Research directions for principles of data management. *Dagstuhl Manifestos*, 7(1):1–29, 2018.

[2] C. Accinelli. Discrimination-aware data transformations (doctoral dissertation, in preparation). University of Genoa, Italy, 2023.

[3] C. Accinelli, B. Catania, G. Guerrini, and S. Minisi. covRew: a Python toolkit for pre-processing pipeline rewriting ensuring coverage constraint satisfaction. In *Proc. EDBT*, pages 698–701, 2021.

[4] C. Accinelli, B. Catania, G. Guerrini, and S. Minisi. The impact of rewriting on coverage constraint satisfaction. In *Proc. EDBT/ICDT Workshops*, 2021.

[5] C. Accinelli, B. Catania, G. Guerrini, and S. Minisi. A coverage-based approach to nondiscrimination-aware data transformation. *ACM J. Data Inf. Qual.*, 2022.

[6] C. Accinelli, S. Minisi, and B. Catania. Coverage-based rewriting for data preparation. In *Proc. EDBT/ICDT Workshops*, 2020.

[7] A. Asudeh, Z. Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *Proc. ICDE*, pages 554–565, 2019.

[8] A. Asudeh, N. Shahbazi, Z. Jin, and H. V. Jagadish. Identifying insufficient data coverage for ordinal continuous-valued attributes. In *Proc. SIGMOD*, pages 129–141, 2021.

[9] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[10] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator. In *Proc. ICDE*, pages 421–430, 2001.

[11] M. Drosou, H. V. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in big data: A review. *Big Data*, 5(2):73–84, 2017.

[12] Y. Lin, Y. Guan, A. Asudeh, and H. V. Jagadish. Identifying insufficient data coverage in databases with multiple relations. *Proc. VLDB Endow.*, 13(11):2229–2242, 2020.

[13] Z. M. Ma and L. Yan. A literature overview of fuzzy database models. *J. Inf. Sci. Eng.*, 24(1):189–202, 2008.

[14] L. Mazilu, N. W. Paton, N. Konstantinou, and A. A. A. Fernandes. Fairness in data wrangling. In *Proc. of the Int. Conf. on Information Reuse and Integration for Data Science, IRI 2020*, 2020.

[15] B. Salimi, J. Gehrke, and D. Suciu. Bias in OLAP queries: Detection, explanation, and removal. In *Proc. SIGMOD*, pages 1021–1035, 2018.

[16] B. Salimi, B. Howe, and D. Suciu. Database repair meets algorithmic fairness. *SIGMOD Rec.*, 49(1):34–41, 2020.

[17] S. Schelter, Y. He, J. Khilnani, and J. Stoyanovich. Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In *Proc. of the Int. Conf. on Extending Database Technology, EDBT 2020*, pages 395–398, 2020.

[18] N. Shahbazi, Y. Lin, A. Asudeh, and H. V. Jagadish. A survey on techniques for identifying and resolving representation bias in data. *CoRR*, abs/2203.11852, 2022.

[19] S. Shetiya, I. P. Swift, A. Asudeh, and G. Das. Fairness-aware range queries for selecting unbiased data. In *Proc. ICDE*, 2022.

[20] T. Simonite. When it comes to gorillas, Google photos remains blind. Wired, Jan. 2018.

[21] J. Stoyanovich, B. Howe, and H. V. Jagadish. Responsible data management. *Proc. VLDB Endow.*, 13(12):3474–3488, 2020.

[22] J. Stoyanovich, K. Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In *Proc. EDBT*, pages 241–252, 2018.

[23] M. Zehlike, K. Yang, and J. Stoyanovich. Fairness in ranking, part I: score-based ranking. *ACM Comput. Surv.*, 55(6):118:1–118:36, 2023.