Unbiased Binning for Fairness-aware Attribute Representation

Abolfazl Asudeh, Zeinab (Mila) Asoodeh,
Bita Asoodeh, Omid Asudeh

University of Illinois Chicago, University of Edinburgh, University of Utah
asudeh@uic.edu

52nd International Conference on Very Large Data Bases
Aug 31st - Sep 4th, 2026 – Boston, MA, USA

VLDB 2026 BOSTON

Outline

1. Motivation
2. Problem Formulation and Key Findings
3. Solution Overview: Unbiased Binning
4. Solution Overview: ε-biased Binning
5. Highlighted Experiments

Motivation

- *Attribute Binning:* Discretizing raw features into bucketized attributes – a common step before sharing a dataset
  - *Example:* equal-size binning ensures an equal number of tuples fall inside each bucket

Motivation

- *Attribute Binning:* Discretizing raw features into bucketized attributes – a common step before sharing a dataset
  - *Example:* equal-size binning ensures an equal number of tuples fall inside each bucket
- **Issue:** discretization can further increase the bias by clustering marginalized populations into lower-tier buckets
  1. *Example:* attribute income

Motivation

- *Attribute Binning:* Discretizing raw features into bucketized attributes – a common step before sharing a dataset
  - *Example:* equal-size binning ensures an equal number of tuples fall inside each bucket
- **Issue:** discretization can further increase the bias by clustering marginalized populations into lower-tier buckets
  1. *Example:* attribute income

Question

How to remove/reduce bias while minimally deviating from the original bucketization?

Outline

1. Motivation
2. Problem Formulation and Key Findings
3. Solution Overview: Unbiased Binning
4. Solution Overview: ε-biased Binning
5. Highlighted Experiments

k-binning

A partitioning of an attribute x into

- k ordered buckets $\mathcal{B} = \{B_1, \cdots, B_k\}$ by specifying $k-1$ boundaries $\{v_1, \cdots, v_{k-1}\}$.
- Size of a bucket: the number of tuples with their value of attribute x being in the range $(v_{j-1}, v_j]$.
  - Equal-size k-binning: $|B_j \cap \mathcal{D}| = \frac{n}{k}, \quad \forall j \in [k]$.

Example: a k-binning

$$\mathcal{B} = \left\{ \begin{array}{ll} B_1 & : \quad\quad\quad x \leq v_1, \\ B_2 & : \quad v_1 < x \leq v_2, \\ \cdots, & \\ B_k & : v_{k-1} < x \end{array} \right\}$$

Problem Formulation: **Unbiased Binning**

Unbiased Binning

Given a dataset $\mathcal{D}$, a binning $\mathcal{B} = \{B_i\}_{i=1}^k$ of an attribute x is unbiased if

$$\frac{|\mathcal{G}_l|}{|\mathcal{D}|} = \frac{|B_j \cap \mathcal{G}_l|}{|B_j \cap \mathcal{D}|}, \quad \forall l \in [\ell], j \in [k]$$

Problem 1

Given a dataset $\mathcal{D}$, an attribute $x \in X$, and a positive integer value k, partition x into k bins $\mathcal{B} = \{B_i\}_{i=1}^k$, such that $\mathcal{B}$ is unbiased, and $\left(\max_{i \in [k]} |B_i \cap \mathcal{D}| - \min_{j \in [k]} |B_j \cap \mathcal{D}| \right)$ is minimized.

Problem Formulation: ε-biased Binning

Unbiased binning may not exist, or it may cause a high Price of Fairness. On the other hand, a small amount of bias may be acceptable in practice.

Bias of a Binning

Bias β of a bucket B_j with respect to a group $\mathbf{g}_l$ as the difference in the ratio of $\mathbf{g}_l$ in B_j compared to its overall ratio:

$$\beta_{\mathcal{D}}(B_j, \mathbf{g}_l) = \left| \frac{|\mathcal{G}_l \cap B_j|}{|\mathcal{D} \cap B_j|} - \frac{|\mathcal{G}_l|}{n} \right| \tag{1}$$

Bias of a binning $\mathcal{B}$ is the maximum bias of its buckets for any group $\mathbf{g}_l$:

$$\beta_{\mathcal{D}}(\mathcal{B}) = \max_{l \in [\ell]} \max_{j \in [k]} \beta_{\mathcal{D}}(B_j, \mathbf{g}_l) \tag{2}$$

Problem 2

Given a dataset $\mathcal{D}$, find a binning $\mathcal{B} = \{B_i\}_{i=1}^{k}$ of an attribute x such that $\beta_{\mathcal{D}}(\mathcal{B}) \leq \varepsilon$.

Key Findings

Objective	Algorithm	Optimal?	Practically Scalable?	Time Complexity	Space Complexity	Problem Scope
Unbiased Binning	UnbiasedBinning	✓	✓	$O(n \log n + m^2 k)$	$O(mk)$	Equal-size Binning
ε-biased Binning	DP	✓	✗	$O(n^2 k)$	$O(n^2)$	Equal-size Binning
	D&C	✗	✓	$O(n \log n)$	$O(n)$	General
	LS	✓	✓	$O(n \log n + n^{k-1})$	$O(n)$	General

Outline

1. Motivation
2. Problem Formulation and Key Findings
3. Solution Overview: Unbiased Binning
4. Solution Overview: ε-biased Binning
5. Highlighted Experiments

Boundary Candidates

* Simplification (for ease of presentation): suppose there are two groups (blue and red)

- Sort and index the tuples based on their values on x, such that $\forall i \in [n]$, $t_i[x] \leq t_{i+1}[x]$.
- Let r_n be the overall ratio of the blue tuples in the dataset, i.e., $r_n = \frac{|\mathcal{G}_1|}{n}$.
- An index $i \in [n]$ is a **boundary candidate** if its blue ratio is equal to the overall blue ratio.

Sorted Indices	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\frac{\textbf{Blue}}{\textbf{All}}$ **Ratios**	0	0	1/3 First	1/4 Pass	2/5	1/2	3/7	1/2	4/9	2/5	5/11	1/2	7/13	1/2	7/15	1/2
Boundary Candidates			Second	Pass		✓		✓				✓		✓		✓

Boundary Candidates

* Simplification (for ease of presentation): suppose there are two groups (blue and red)

- Sort and index the tuples based on their values on x, such that $\forall i \in [n]$, $t_i[x] \leq t_{i+1}[x]$.
- Let r_n be the overall ratio of the blue tuples in the dataset, i.e., $r_n = \frac{|\mathcal{G}_1|}{n}$.
- An index $i \in [n]$ is a **boundary candidate** if its blue ratio is equal to the overall blue ratio.

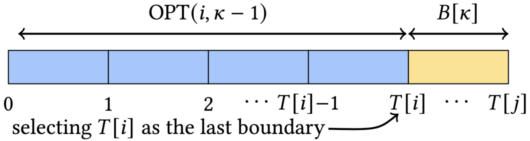
Sorted Indices	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\frac{\text{Blue}}{\text{All}}$ Ratios	0	0	1/3 First	1/4 Pass	2/5	1/2	3/7	1/2	4/9	2/5	5/11	1/2	7/13	1/2	7/15	1/2
Boundary Candidates			Second	Pass		✓		✓				✓		✓		✓

Theorem 1.

The boundaries of an unbiased binning $\mathcal{B}$ must only belong to the set of boundary candidates T.

Solution overview

1. Step 0: Sort $\mathcal{D}$ on x $O(n \log n)$
2. Step 1: Find the boundary candidates T
   $O(n)$ (two passes over the sorted lists)
3. Step 2: Dynamic Programming on T (details in the paper)
   $O(m^2 k)$; $m = |T|$



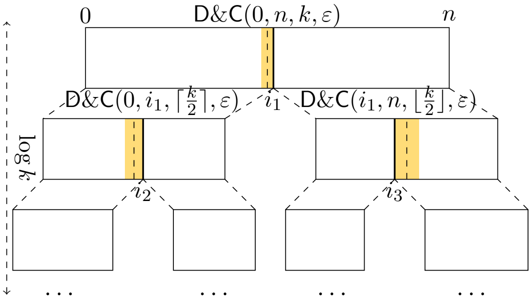
Outline

1. Motivation
2. Problem Formulation and Key Findings
3. Solution Overview: Unbiased Binning
4. Solution Overview: ε-biased Binning
5. Highlighted Experiments

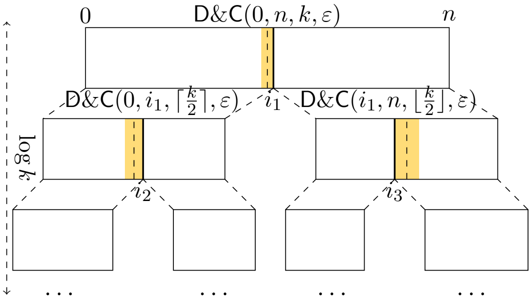
A practical solution

- Solution 1: Dynamic Programming for ε-biased binning
  - Theorem 1 (boundary candidates idea) does not hold for ε-biased binning.
  - $\Rightarrow$ Quadratic time and space $O(n^2)$ –> not scalable
- How about:
  1. A near-optimal solution?
  2. A practically efficient optimal solution?
- Key idea: Combine D&C with local search

D&C: a near-optimal solution



D&C: a near-optimal solution



Theorem 2.

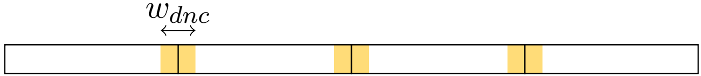
D&C finds an ε-biased binning unless none exists.

In summary

D&C quickly (in $O(n \log n)$) finds a near-optimal ε-binning.

LS: optimal and practically-efficient

1. Step 1: Find a near-optimal solution $\mathcal{B}^{app}$ using DnC
2. Step 2: using the objective value of $\mathcal{B}^{app}$ as an upper bound, apply a local search around the initial binning to find the optimal solution.



Time/Space complexity

LS has a space complexity of $O(n)$ and a time complexity of $O(n \log n + (w_{dnc})^{k-1}) \leq O(n \log n + n^{k-1})$.

Outline

1. Motivation
2. Problem Formulation and Key Findings
3. Solution Overview: Unbiased Binning
4. Solution Overview: ε-biased Binning
5. Highlighted Experiments

Impact on downstream tasks: Performance and Fairness Metrics Before and After Unbiased Binning

Logistic Regression

Metric	Baseline Binning	ε-biased Binning 0.8I	0.6I	$\sim 0.5I$	Improvement
Accuracy	0.67	0.7033	0.70	0.70	↑ Slightly improved
F1-score	0.25	0.2261	0.2623	0.18	↓ Slightly decreased
DI	0.9378	0.9804	0.9832	**0.9949**	✓ Very close to 1 (ideal)
SPD	-0.0545	-0.0180	-0.0151	**-0.0048**	✓ Almost perfect parity
EOD	-0.0794	-0.0195	-0.0215	+**0.0116**	✓ Major Improvement
AOD	-0.0341	-0.0124	-0.0113	**-0.0050**	✓ Almost perfect parity
INDV	0.974	0.973	0.968	0.958	↓ Slightly decreased

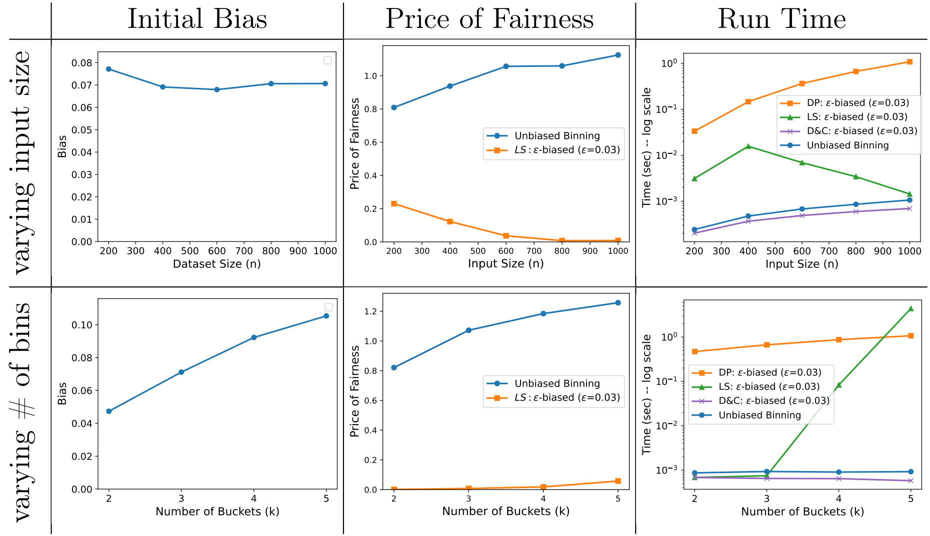
$I = [0.25, 0.093, 0.1]$ is the initial Bias of Baseline (Equal-size) Binning

DI: Disparate Impact; **SPD**: Statistical Parity Difference;

EOD: Equal Opportunity Difference; **AOD**: Average Odds Difference.

INDV: Consistency (Individual Fairness).

Highlighted experiments on German Credit dataset



Future Work

Thank you, REVIEWERS!

- Beyond individual attributes: higher-order correlations across multiple attributes.
- Task-aware Fair Binning
- Unbiased Binning as an In-process intervention. E.g., redefine the problem of "Optimal Binning" to maximize information while minimizing bias

Thank you, Question?

UIC

InDeX Lab
Innovative
Data
eXploration
Laboratory

THE UNIVERSITY OF EDINBURGH

VLDB 2026 BOSTON