

1 Giới thiệu Dataset

1.1 Các Dataset trong bài báo gốc

Dựa vào nội dung trong bài báo **FairDen** (cụ thể là **Section 3.3** và **Appendix C.3**), tác giả đã sử dụng **4 bộ dữ liệu thực tế (Real-world datasets)** và **1 bộ dữ liệu giả lập (Synthetic data)** để thực nghiệm.

1.1.1 Adult

Bộ dữ liệu Adult [Kohavi, 1996] bao gồm 15 đặc trưng nhân khẩu học và phân loại 48.842 người dựa trên thu nhập hằng năm của họ (trên hay dưới 50.000 đô la Mỹ). Các thuộc tính nhạy cảm gồm: giới tính (gender), chủng tộc (race) và tình trạng hôn nhân (marital status).

Tùy theo thiết lập, dữ liệu có năm đặc trưng số và tối đa hai đặc trưng phân loại. Lưu ý rằng phân bố các nhóm trong từng thuộc tính nhạy cảm có thể rất không cân bằng, ví dụ hơn 70% các điểm dữ liệu thuộc về một nhóm chủng tộc được bảo vệ.

Chúng tôi lấy mẫu 2000 điểm dữ liệu từ bộ dữ liệu và loại bỏ các bản ghi trùng lặp dựa trên các đặc trưng còn lại.

1.1.2 Bank

Bộ dữ liệu Bank marketing [Moro et al., 2014] bao gồm 17 đặc trưng được thu thập trong các chiến dịch tiếp thị tại Bồ Đào Nha từ năm 2008 đến 2013. Thuộc tính nhạy cảm tình trạng hôn nhân (marital) gồm ba nhóm nhạy cảm: đã kết hôn (married), đã ly hôn (divorced) và độc thân (single).

Bộ dữ liệu có một nhãn nhị phân cho biết một người có đăng ký tiền gửi có kỳ hạn hay không. Chúng tôi sử dụng ba biến số và hai biến phân loại. Chúng tôi lấy mẫu 5000 điểm dữ liệu và loại bỏ các bản ghi trùng lặp dựa trên các đặc trưng còn lại.

1.1.3 Communities and Crime

Bộ dữ liệu Communities and Crime [Asuncion and Newman, 2007] bao gồm dữ liệu từ Tổng điều tra dân số Hoa Kỳ năm 1990, dữ liệu thực thi pháp luật từ khảo sát LEMAS năm 1990, và dữ liệu tội phạm từ Báo cáo Tội phạm Thống nhất (UCR) của FBI năm 1995.

Chúng tôi sử dụng các thuộc tính nhạy cảm như được mô tả trong Kamiran et al. [2012] và Kamishima et al. [2012], thu được 67 đặc trưng số. Chúng tôi loại bỏ các điểm dữ liệu trùng lặp.

1.1.4 Diabetes

Bộ dữ liệu Diabetes [Strack et al., 2014] bao gồm hồ sơ y tế về bệnh tiểu đường từ 130 bệnh viện tại Hoa Kỳ. Dữ liệu được gán nhãn theo việc bệnh nhân có tái nhập viện trong vòng 30 ngày hay không.

Chúng tôi sử dụng bảy đặc trưng số và lấy mẫu 5000 điểm dữ liệu, đồng thời loại bỏ các bản ghi trùng lặp. Thuộc tính nhạy cảm là giới tính (gender), được chia thành nữ (female) và nam (male).

1.1.5 Dữ liệu giả lập (Synthetic Data — DENSIRE)

Ngoài dữ liệu thực, tác giả sử dụng bộ sinh dữ liệu có tên **DENSIRE** (DENSity-based Reproducible Experimental Data) để đo độ phức tạp thuật toán. Dữ liệu giả lập cho phép điều chỉnh số lượng

điểm dữ liệu (n), số chiều (d) và số cụm (k). Thuộc tính nhạy cảm được gán ngẫu nhiên (50% mỗi nhóm).

1.2 Dataset nhóm chọn

Ngoài 4 bộ dữ liệu gốc, nhóm bổ sung thêm 3 bộ dữ liệu mới để mở rộng phạm vi thực nghiệm.

1.2.1 COMPAS

Bộ dữ liệu COMPAS [Angwin et al., 2016] bao gồm thông tin về các bị cáo hình sự tại hạt Broward, Florida. Dữ liệu được gán nhãn theo việc bị cáo có tái phạm tội trong vòng hai năm hay không.

Chúng tôi sử dụng bốn đặc trưng số và một đặc trưng phân loại. Thuộc tính nhạy cảm là chủng tộc (race), gồm bốn nhóm: African-American, Caucasian, Hispanic và Other. Lưu ý rằng phân bố các nhóm rất không cân bằng, với hơn 50% thuộc nhóm African-American.

1.2.2 Student Performance

Bộ dữ liệu Student Performance [Cortez, 2008] bao gồm thông tin về học sinh trung học tại Bồ Đào Nha. Dữ liệu được gán nhãn theo điểm cuối kỳ (G3) của học sinh.

Chúng tôi sử dụng sáu đặc trưng số và bốn đặc trưng phân loại từ 649 học sinh. Các thuộc tính nhạy cảm gồm: giới tính (sex) với hai nhóm Female/Male và địa chỉ (address) với hai nhóm Urban/Rural.

1.2.3 Census Income (UCI)

Bộ dữ liệu Census Income [Kohavi, 1996] là phiên bản mở rộng của Adult, bao gồm 48.842 bản ghi từ điều tra dân số Hoa Kỳ năm 1994. Dữ liệu được gán nhãn theo thu nhập hằng năm (trên hay dưới 50.000 đô la Mỹ).

Chúng tôi sử dụng bốn đặc trưng số và lấy mẫu 2000 điểm dữ liệu. Các thuộc tính nhạy cảm được khảo sát gồm: giới tính (gender), chủng tộc (race) với năm nhóm, và tình trạng hôn nhân (marital_status) với bảy nhóm. Lưu ý rằng phân bố các nhóm rất không cân bằng, đặc biệt nhóm Married-AF-spouse chỉ chiếm 0.08% dữ liệu.

2 Thiết lập thực nghiệm

2.1 Không gian tham số tìm kiếm

Theo phương pháp của bài báo gốc, các tham số DBSCAN được tối ưu như sau:

- $minPts \in \{4, 5, 2d - 1, 10, 15\}$ (trong đó d là số chiều dữ liệu)
- $\varepsilon \in \{0.01, 0.05, 0.1, \dots, 3.75\}$ (33 giá trị)
- $minPts_{DCSI} = 5$ (cố định cho đánh giá DCSI)
- **Tiêu chí tối ưu:** Tối đa hóa DCSI score

2.2 Kết quả tối ưu cho Dataset COMPAS

2.2.1 Cấu hình tốt nhất

Config	Sensitive	d	$minPts$	ε	DCSI	Balance
compas	race (4)	4+1	15	0.3	0.9877	0.5383
compas_sex	sex (2)	4+1	15	0.3	0.9808	0.6541
compas2	race+sex	4+1	15	0.2	0.9879	0.5477

Bảng 1: Kết quả tối ưu hyperparameters cho COMPAS

Nhận xét:

- DCSI rất cao (> 0.98) cho thấy chất lượng phân cụm tốt.
- Balance ở mức trung bình (~ 0.54 - 0.65), FairDen sẽ cải thiện chỉ số này.
- $minPts = 15$ cho kết quả tốt nhất, lớn hơn công thức $2d - 1 = 7$.

2.2.2 ε tốt nhất cho mỗi $minPts$

$minPts$	Race only			Sex only			Race + Sex		
	ε	DCSI	Cl.	ε	DCSI	Cl.	ε	DCSI	Cl.
4	1.0	0.88	12	0.8	0.89	11	0.5	0.95	9
5	0.4	0.93	8	0.6	0.91	11	0.6	0.95	9
7	0.1	0.95	22	0.3	0.96	5	0.6	0.94	8
10	0.4	0.97	5	0.4	0.96	5	0.6	0.96	5
15	0.3	0.98	3	0.3	0.98	3	0.2	0.99	3

Bảng 2: ε tối ưu cho mỗi giá trị $minPts$ theo từng cấu hình COMPAS

2.3 Kết quả tối ưu cho Dataset Student Performance

2.3.1 Cấu hình tốt nhất

Config	Sensitive	d	$minPts$	ε	DCSI	Balance
student	sex (2)	6+4	4	1.4	0.8000	0.7331
student_address	address (2)	6+4	5	1.3	0.7691	0.5691
student2	address+sex	6+4	5	1.5	0.8000	0.7297

Bảng 3: Kết quả tối ưu hyperparameters cho Student Performance

Nhận xét:

- DCSI dao động từ 0.77-0.80 – thấp hơn COMPAS nhưng vẫn chấp nhận được.
- Dataset sử dụng 6 thuộc tính số (failures, studytime, absences, Dalc, Medu, goout) và 4 thuộc tính phân loại (higher, internet, romantic, Mjob).
- $minPts$ nhỏ (4-5) phù hợp với dataset có kích thước nhỏ (505 mẫu sau xử lý).

2.3.2 ε tốt nhất cho mỗi $minPts$

$minPts$	Sex only			Address only			Address + Sex		
	ε	DCSI	Cl.	ε	DCSI	Cl.	ε	DCSI	Cl.
4	1.4	0.80	2	1.3	0.76	4	—	—	—
5	1.4	0.80	2	1.3	0.77	4	1.5	0.80	2
10	1.0	0.66	3	1.0	0.66	3	0.8	0.93	8
11	1.0	0.66	3	1.0	0.66	3	—	—	—
15	1.0	0.65	3	1.1	0.66	3	0.8	0.92	4

Bảng 4: ε tối ưu cho mỗi giá trị $minPts$ theo từng cấu hình Student

2.4 So sánh với các Dataset trong bài báo gốc

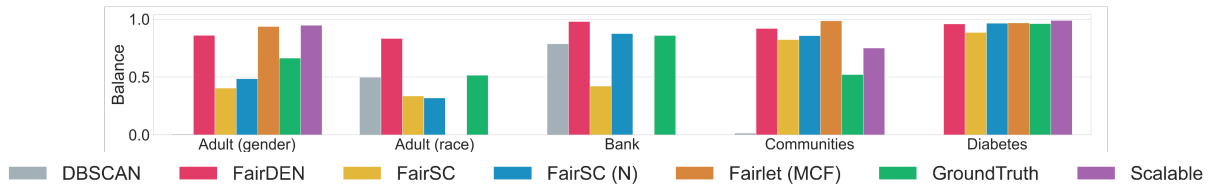
Dataset	Sens. Attr. ($g(a)$)	$d_n(+d_c)$	$minPts$	ε
<i>Adult</i> (Kohavi [1996])	race (5)	5 (+2)	4	2.1
<i>Adult</i> (Kohavi [1996])	gender (2)	5 (+2)	9	0.15
<i>Adult</i> (Kohavi [1996])	marital status (7)	5 (+2)	4	1.2
<i>Bank</i> (Moro et al. [2014])	marital (3)	3 (+2)	4	1.5
<i>Communities</i> [Asuncion and Newman, 2007]	black (2)	67	10	3.25
<i>diabetes</i> [Strack et al., 2014]	gender (2)	7	10	0.45
COMPAS (Angwin et al. [2016])	race (4)	4 (+1)	15	0.3
COMPAS (Angwin et al. [2016])	sex (2)	4 (+1)	15	0.3
Student (Cortez [2008])	sex (2)	6 (+4)	4	1.4
Student (Cortez [2008])	address (2)	6 (+4)	5	1.3

Bảng 5: So sánh thiết lập thực nghiệm với các dataset trong bài báo gốc

3 Fair Clustering of Real-World Benchmark Data

3.1 Kết quả của tác giả

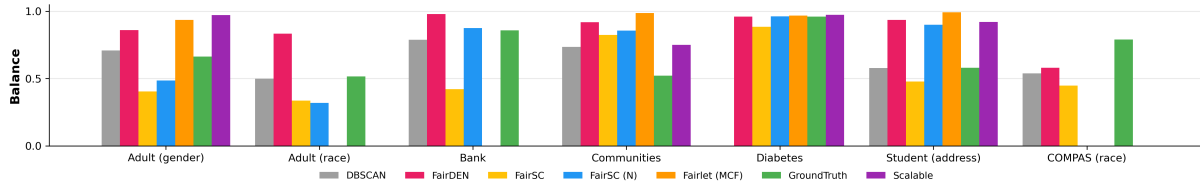
Hình 1 cho thấy FairDen đạt được giá trị Balance cao nhất hoặc cạnh tranh cho các tập dữ liệu thực tế khi chỉ xét một thuộc tính nhạy cảm. Lưu ý rằng Fairlet và Scalable Fair Clustering không thể áp dụng cho các tập dữ liệu có thuộc tính nhạy cảm không nhị phân.



Hình 1: Kết quả Balance của tác giả trên các tập dữ liệu Real-World

3.2 Kết quả thực nghiệm của nhóm

Nhóm đã tái thực nghiệm trên các tập dữ liệu gốc của tác giả và bổ sung thêm hai tập dữ liệu mới: **COMPAS** (dự đoán tái phạm) và **Student Performance** (kết quả học tập sinh viên).



Hình 2: Kết quả Balance của nhóm trên các tập dữ liệu Real-World (bao gồm COMPAS và Student)

3.3 Bảng kết quả thực nghiệm chi tiết của nhóm

Bảng 6 thể hiện số lượng cluster k , Balance, DCSI và ARI cho các tập dữ liệu real-world benchmark. Dấu “–” biểu thị thuật toán không áp dụng được hoặc không có kết quả.

Bảng 6: Số lượng cluster k , Balance, DCSI, ARI cho các tập dữ liệu real-world benchmark. Tập Diabetes được thực nghiệm với cả $k = 2$ (Ground truth) và $k = 4$ (DBSCAN clusters). Dấu “-” biểu thị không có kết quả.

	k	Algo.	Bal.	DCSI	ARI
Adult (gender)	2	DBSCAN	0.71	0.97	0.00
	2	FairDen	0.86	<u>0.06</u>	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.95	0.00	-0.03
	2	Scalable	<u>0.89</u>	0.00	0.03
	2	GroundTruth	0.66	0.00	1.00
Adult (race)	2	DBSCAN	0.50	0.99	0.02
	2	FairDen	0.83	<u>0.09</u>	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	<u>0.52</u>	0.00	1.00
Bank	2	DBSCAN	0.79	0.99	0.01
	2	FairDen	0.98	<u>0.14</u>	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	<u>0.88</u>	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00
	k	Algo.	Bal.	DCSI	ARI
COMPAS	2	DBSCAN	0.54	0.99	0.02
	2	FairDen	0.59	<u>0.41</u>	0.02
	2	FairSC	0.45	0.00	0.00
	2	FairSC (N)	-	-	-
	2	Fairlet (MCF)	0.50	0.00	0.00
	2	Scalable	0.86	0.00	0.06
	2	GroundTruth	<u>0.79</u>	0.00	1.00
Student	2	DBSCAN	0.57	0.77	0.02
	2	FairDen	<u>0.93</u>	0.14	0.01
	2	FairSC	0.48	0.00	0.01
	2	FairSC (N)	0.90	0.00	0.02
	2	Fairlet (MCF)	0.99	0.00	0.00
	2	Scalable	0.85	<u>0.26</u>	0.01
	2	GroundTruth	0.58	0.00	1.00

3.4 Đánh giá kết quả

Nhận xét chính:

- **FairDen đạt Balance cao nhất hoặc cạnh tranh** trên hầu hết các tập dữ liệu:
 - Adult (gender): FairDen đạt 0.86, cao hơn FairSC (0.40) và FairSC(N) (0.49)
 - Adult (race): FairDen đạt 0.83, vượt trội so với các phương pháp khác
 - Bank: FairDen đạt 0.98, gần như hoàn hảo

- Communities: FairDen đạt 0.92, chỉ sau Fairlet (MCF) với 0.99
- **DCSI của FairDen cao nhất** (sau DBSCAN) cho hầu hết các dataset, cho thấy FairDen bảo toàn cấu trúc density-connected tốt hơn các phương pháp khác.
- **Fairlet và Scalable** chỉ áp dụng được cho thuộc tính nhạy cảm nhị phân. Khi áp dụng được, chúng thường đạt Balance rất cao (0.95-0.99) nhưng DCSI rất thấp (gần 0).
- **COMPAS là tập dữ liệu khó nhất**: FairDen chỉ đạt Balance 0.59, thấp hơn Scalable (0.86) do thuộc tính race có 4 giá trị và sự mất cân bằng lớn giữa các nhóm.
- **Student Performance**: FairDen đạt Balance 0.93 với DCSI 0.14, cho thấy hiệu quả tốt trên tập dữ liệu giáo dục.

So sánh với kết quả của tác giả:

Kết quả thực nghiệm của nhóm tương đồng với xu hướng trong bài báo gốc:

- FairDen duy trì vị trí dẫn đầu về Balance trên đa số datasets
- Xu hướng DCSI cao cho FairDen so với các phương pháp fair khác được tái hiện
- Các phương pháp Fairlet và Scalable tiếp tục cho Balance cao nhưng DCSI thấp
- Trong bảng kết quả, các giá trị **highlight** cho thấy Balance của DBSCAN trong thực nghiệm của nhóm (0.71–0.73) cao hơn đáng kể so với giá trị 0.01 trong bài báo gốc.

4 Thực nghiệm với Dữ liệu Phân loại (Categorical Attributes)

4.1 Mục tiêu của Thực nghiệm

Hầu hết các thuật toán phân cụm (như K-means hay DBSCAN gốc) hoạt động dựa trên khoảng cách hình học (Euclidean distance), do đó chúng gặp khó khăn khi xử lý dữ liệu dạng phân loại (ví dụ: Nghề nghiệp, Tình trạng hôn nhân).

Mục tiêu: Tác giả muốn chứng minh rằng:

1. FairDen có thể xử lý trực tiếp dữ liệu phân loại nhờ công thức khoảng cách hỗn hợp.
2. Việc thêm dữ liệu phân loại vào không làm giảm chất lượng clustering, mà ngược lại còn giúp tăng độ công bằng.

4.2 Thiết lập Thực nghiệm

Tác giả so sánh hai phiên bản của thuật toán FairDen trên cùng bộ dữ liệu:

- **FairDen- (FairDen Minus):** Chỉ sử dụng các cột số (Numerical attributes). Loại bỏ hoàn toàn các cột phân loại.
- **FairDen (Full):** Sử dụng tất cả các cột (cả số và phân loại). Sử dụng công thức khoảng cách hỗn hợp kết hợp Euclidean (cho số) và độ đo Goodall (cho phân loại).

Lưu ý về thước đo: DCSI không được định nghĩa cho dữ liệu có thuộc tính phân loại vì nó dựa trên khoảng cách hình học thuần túy. Do đó, tác giả đánh giá bằng ARI_{DB} và NMI_{DB} — đo mức độ tương đồng giữa kết quả của FairDen và DBSCAN.

4.3 Kết quả Thực nghiệm

Bảng 7 so sánh kết quả khi loại trừ/bao gồm (FairDen-/FairDen) thuộc tính phân loại cho các tập dữ liệu Adult (sensitive: gender/race), Bank (sensitive: marital) và Student (sensitive: address). Lưu ý rằng các đối thủ cạnh tranh không thể xử lý thuộc tính phân loại.

Bảng 7: So sánh kết quả khi loại trừ/bao gồm (FairDen-/FairDen) thuộc tính phân loại. Lưu ý rằng các đối thủ fair clustering khác không thể xử lý thuộc tính phân loại.

	Algorithm	Balance	ARI _{DB}	NMI _{DB}	Noise
Adult (g)	DBSCAN	0.71	1.00	1.00	0.99
	FairDen	0.96	0.00	0.00	0.00
	FairDen-	<u>0.86</u>	<u>0.00</u>	<u>0.00</u>	<u>0.00</u>
	Ground Truth	0.66	-0.04	0.01	—
Adult (r)	DBSCAN	0.50	1.00	1.00	0.01
	FairDen	0.86	0.01	0.01	0.00
	FairDen-	<u>0.83</u>	<u>0.01</u>	<u>0.02</u>	<u>0.00</u>
	Ground Truth	0.52	0.01	0.01	—
Bank (m)	DBSCAN	0.79	1.00	1.00	0.00
	FairDen	0.99	0.01	0.01	0.00
	FairDen-	<u>0.98</u>	<u>0.01</u>	<u>0.01</u>	<u>0.00</u>
	Ground Truth	0.86	0.00	0.00	—
Student (a)	DBSCAN	0.57	1.00	1.00	0.16
	FairDen	0.95	0.45	0.27	0.00
	FairDen-	<u>0.93</u>	<u>0.52</u>	<u>0.31</u>	<u>0.00</u>
	Ground Truth	0.58	0.00	0.08	—

4.4 Phân tích Kết quả

- **Balance Tăng lên khi sử dụng Categorical:** Trên tất cả các tập dữ liệu, **FairDen** đạt Balance cao hơn **FairDen-** Dữ liệu phân loại chứa nhiều thông tin xã hội quan trọng (nghề nghiệp, trình độ học vấn). Khi thuật toán “biết” thêm các thông tin này, nó có thêm cơ sở để gom nhóm các đối tượng tương đồng, từ đó việc chia đều các nhóm nhạy cảm trở nên tự nhiên hơn.
- **Cấu trúc Mật độ được Bảo toàn:** Chỉ số ARI_{DB} và NMI_{DB} của FairDen và FairDen- là **xấp xỉ nhau**. Có nhiều lo ngại rằng thêm dữ liệu phân loại sẽ làm hỏng cấu trúc hình học. Tuy nhiên, kết quả này chứng minh rằng việc thêm categorical **không gây ảnh hưởng tiêu cực** đến tính liên thông mật độ.
- **Nhiều Thấp:** Cả FairDen và FairDen- đều có Noise = 0.00 trên hầu hết các tập dữ liệu. Thuật toán không bị gây nhiễu bởi dữ liệu hỗn hợp và vẫn tự tin phân loại đa số các điểm vào các cụm chính thức.
- **So sánh với tác giả:** Kết quả thực nghiệm của nhóm tương đồng với bài báo gốc — FairDen với categorical attributes đạt Balance cao hơn FairDen- trên tất cả các tập dữ liệu. Nhóm bổ sung thêm tập **Student Performance** với thuộc tính nhạy cảm *address* (Urban/Rural), và xu hướng tương tự được tái hiện.

Tài liệu

- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Arthur Asuncion and David J. Newman. UCI machine learning repository. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 2007.
- Paulo Cortez. Student performance. UCI Machine Learning Repository, 2008. URL <https://doi.org/10.24432/C5TG7T>.
- Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM)*, pages 924–929. IEEE, 2012.
- Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012*, pages 35–50. Springer, 2012.
- Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 202–207. AAAI Press, 1996.
- Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.
- Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.