Algorithmic Bias

from discrimination discovery to fairness-aware data mining

Sara Hajian
Francesco Bonchi @francescobonchi
Carlos Castillo @chatox

KDD 2016 Tutorial, August 13th, 2016. San Francisco, USA

Part I: Introduction and context

Part II: Discrimination discovery

Part III: Fairness-aware data mining

Part IV: Challenges and directions for future research

Discussion and further questions

Part I: Introduction and context

Part II: Discrimination discovery

Part III: Fairness-aware data mining

Part IV: Challenges and directions for future research

Discussion and further questions

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Discrimination in specific contexts

Discrimination and privacy

Resources

A dangerous reasoning

To discriminate is to treat someone differently

(Unfair) discrimination is based on group membership, not individual merit

People's decisions include objective and subjective elements

Hence, they can be discriminate

Algorithmic inputs include only objective elements

Hence, they cannot discriminate?

A stable but unequal marriage

National Resident Match Program (NRMP): US program to match interns to hospitals

V1: 1952 → 1997: algorithm favors hospitals and allows strategic manipulation

V2: 1998 → present: new algorithm is resistant to strategic manipulation

A. E. Roth and E. Peranson (1997). The effects of the change in the NRMP matching algorithm. JAMA, 278(9):729–732.

On the web: race and gender stereotypes reinforced

- Results for "CEO" in Google Images: 11% female, US 27% female CEOs
  - Also in Google Images, "doctors" are mostly male, "nurses" are mostly female
- Google search results for professional vs. unprofessional hairstyles for work

Image results: "Unprofessional hair for work"

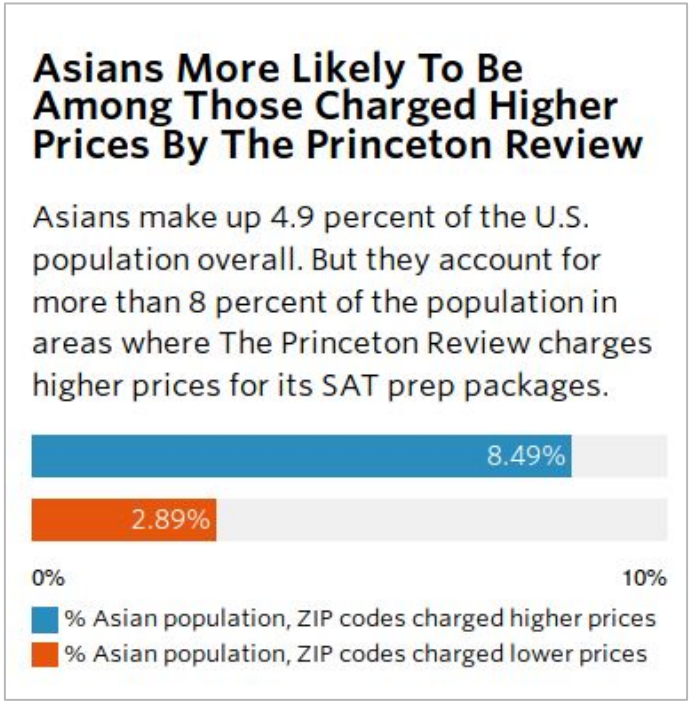
Image results: "Professional hair for work"

More on web discrimination later ...

M. Kay, C. Matuszek, S. Munson (2015): Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. CHI'15.

Geography and race: the "Tiger Mom Tax"

Pricing of SAT tutoring by The Princeton Review in the US doubles for Asians, due to geographical price discrimination



J. Angwin and J. Larson (2015). The tiger mom tax. ProPublica.

Judiciary use of COMPAS scores

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a 137-questions questionnaire and a predictive model for "risk of recidivism" and "risk of violent recidivism." The model is a proprietary secret of Northpointe, Inc.

Prediction accuracy of recidivism for blacks and whites is about the same (63% and 59%), but errs by being too lenient with whites and too harsh with blacks:

- Blacks that did not reoffend
  were classified as high risk twice as much as whites that did not reoffend
- Whites who did reoffend
  were classified as low risk twice as much as blacks who did reoffend

Pro Publica, May 2016. *https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm*

Supreme Court of Wisconsin (US) — July 13th, 2016

The court ruled that judges are allowed to use COMPAS scores, but they:

- **must** receive them accompanied by disclaimers and criticisms
- **can** use them as a factor to give non-prison alternatives to prison
- **can** use them as a factor to impose terms and conditions in parole
- **cannot** use them to make sentences longer or shorter

Wall Street Journal, July 13th, 2016. http://blogs.wsj.com/law/2016/07/13/court-judges-can-consider-predictive-algorithms-in-sentencing/

Self-perpetuating algorithmic biases

Credit scoring algorithm suggests Joe has high risk of defaulting

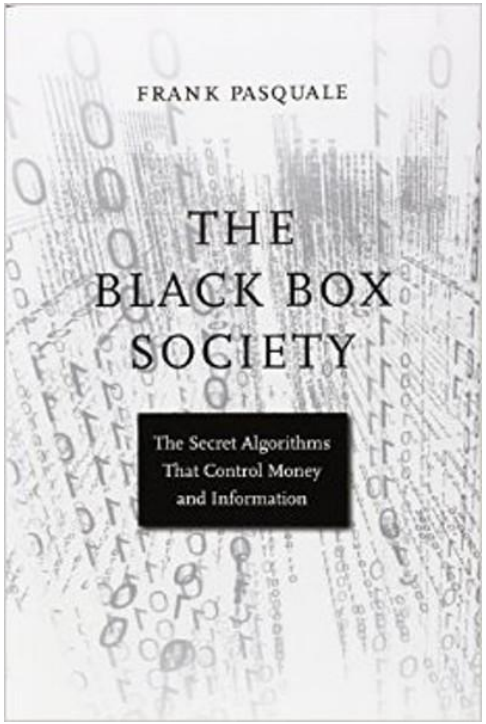
Hence, Joe needs to take a loan at a higher interest rate

Hence, Joe has to make payments that are more onerous

Hence, Joe's risk of defaulting has increased

The same happens with stop-and-frisk of minorities further increasing incarceration rates

To make things worse ...



Algorithms are "black boxes" protected by

- Industrial secrecy
- Legal protections
- Intentional obfuscation

Discrimination becomes invisible

Mitigation becomes impossible

F. Pasquale (2015): The Black Box Society. Harvard University Press.

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Discrimination in specific contexts

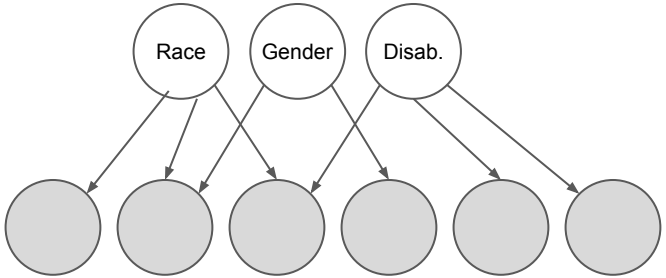
Discrimination and privacy

Resources

Some sources of algorithmic bias

Data as a social mirror

Protected attributes redundantly encoded in observables

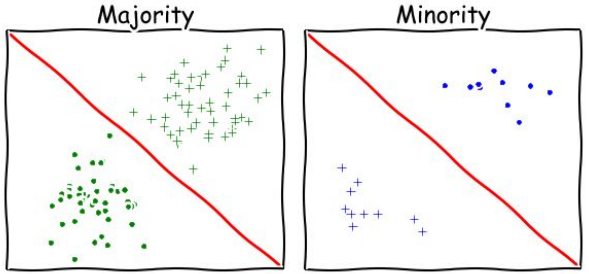


Correctness and completeness

Garbage in, garbage out (GIGO)

Sample size disparity: learn on majority

Errors concentrated in the minority class



M. Hardt (2014): "How big data is unfair". Medium.

Data mining assumptions might not hold

Data mining assumptions are not always observed in reality

- Variables might not be independently identically distributed
- Samples might be biased
- Labels might be incorrect

Errors might be concentrated in a particular class

Sometimes, we might be seeking more simplicity than what is possible

T. Calders and I. Žliobaitė (2013). Why unbiased computational processes can lead to discriminative decision procedures. Chapter 3 of: Discrimination and Privacy in the Information Society. Springer.

Two areas of concern: data and algorithms

Data inputs:

- Poorly selected (e.g., observe only car trips, not bicycle trips)
- Incomplete, incorrect, or outdated
- Selected with bias (e.g., smartphone users)
- Perpetuating and promoting historical biases (e.g., hiring people that "fit the culture")

Algorithmic processing:

- Poorly designed matching systems
- Personalization and recommendation services that narrow instead of expand user options
- Decision making systems that assume correlation implies causation
- Algorithms that do not compensate for datasets that disproportionately represent populations
- Output models that are hard to understand or explain hinder detection and mitigation of bias

Executive Office of the US President (May 2016): "Big Data: A Report on Algorithmic Systems,Opportunity,and Civil Rights"

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Data analysis techniques for discrimination discovery: multidisciplinary approaches in the economic, legal, and statistical domains

Parallels with privacy (conceptually)

Legal concepts

Anti-discrimination legislation typically seeks **equal access** to employment, working conditions, education, social protection, goods, and services

Anti-discrimination legislation is very diverse and includes **many legal concepts**

- **Genuine occupational requirement** (male actor to portray male character)
- **Disparate impact** and **disparate treatment**
- **Burden of proof** and **situation testing**
- **Group under-representation** principle

Discrimination: treatment vs impact

Modern legal frameworks offer various levels of **protection** for being discriminated by belonging to a particular class of: gender, age, ethnicity, nationality, disability, religious beliefs, and/or sexual orientation

Disparate **treatment**:

> Treatment depends on class membership

Disparate **impact**:

> Outcome depends on class membership
>
> Even if (apparently?) people are treated the same way

Disparate treatment: crossing legal red lines

In the US, "steering", i.e. offering more expensive financial services to disadvantaged individuals is forbidden [Steel and Angwin 2010]

In Europe, insurers cannot discriminate by gender, even if they have accurate, well-grounded, and transparent statistical models indicating costs are higher for women [European Court of Justice (2011)]

Disparate impact

Doctrine solidified in the US after [Griggs v. Duke Power Co. 1971] where a high school diploma was required for unskilled work, excluding black applicants

In algorithms, disparate impact can be caused by, e.g.:

(i) how data are collected
(ii) how data are labeled
(iii) which features are used
(iv) usage of proxies for a protected attribute
...

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. California Law Review 104.

Proving discrimination is hard

In very rare cases there might be a confession that someone discriminated someone else because s/he was women, black, homosexual, etc.

In rare cases the evidence is almost comically overwhelming (e.g., Chinese applicants not interviewed for a position requiring fluency in Mandarin UK, 1991)

In most cases it is hard to prove intentions. The burden of proof can be shared (as in the European Court of Justice) for discrimination cases:

the **accuser** must produce evidence of the **consequences**,
the **defendant** must produce evidence that the **process** was fair

Alternative: situational testing

Migration Policy Group and Swedish Centre for Equal Rights (2009): Proving discrimination cases - the role of situational testing.

Do-it-yourself situational testing (We are not encouraging you to try!)

"Gay" vs "Hetero" Job Interview (2015)

Hidden camera video about two job interviews for the same position.

Olle has a truck driver license, experience, and is enthusiastic about the job. "But" he mentions he is saving money to buy an apartment with his boyfriend.

Konrad says he has no experience, spends the whole day playing video games, and shows no enthusiasm for the job.

Who do you think got the job offer?

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Discrimination in specific contexts

Discrimination and privacy

Resources

Principles for quantifying discrimination

In the context of an algorithm generating a prediction:

- Predictions for people with similar non-protected attributes should be similar
- Differences should be mostly explainable by non-protected attributes

Two basic frameworks for measuring discrimination:

- Discrimination at the **individual level**: consistency or individual fairness
- Discrimination at the **group level**: statistical parity

I. Žliobaitė (2015): A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

Consistency or individual fairness

Consistency score

$$C = 1 - \sum_i \sum_{yj \in knn(yi)} |y_i - y_j|$$

Where $knn(y_i)$ = k nearest neighbors of y_i

A consistent or individually fair algorithm is one in which similar people experience similar outcomes

Note that similar people to an individual in a protected group may also belong to that protected group, and perhaps they are all treated equally badly

Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In Proc. of the 30th Int. Conf. on Machine Learning. 325–333.

Statistical parity focuses on proportions

Example:

"Protected group" ~ "people with disabilities"

"Benefit granted" ~ "getting a scholarship"

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

Intuitively, if
a/n_1, the fraction of people with disabilities that **does not** get a scholarship
is much **larger** than
c/n_1, the fraction of people without disabilities that **does not** get a scholarship,
then people with disabilities could claim they are being discriminated.

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Simple discrimination measures

These measures compare the protected group against the unprotected group:

- Risk difference = RD = $p_1 - p_2$
- Risk ratio or relative risk = RR = p_1 / p_2
- Relative chance = RC = $(1-p_1) / (1-p_2)$
- Odds ratio = RR/RC

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

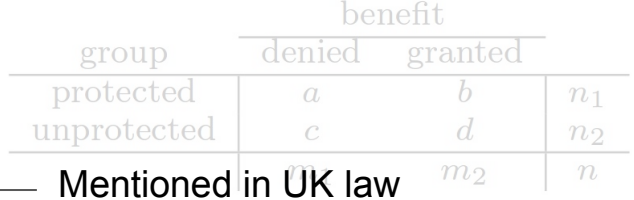
$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Simple discrimination measures

These measures compare the protected group against the unprotected group:

- **Risk difference** = RD = $p_1 - p_2$
- **Risk ratio** or relative risk = RR = p_1 / p_2
- Relative chance = RC = $(1-p_1) / (1-p_2)$
- Odds ratio = RR/RC



Mentioned in UK law

Mentioned by EU Court of Justice

US courts focus on selection rates:
$(1-p_1)$ and $(1-p_2)$

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Extended discrimination measures

These measures compare the **protected** group against the **entire population**:

- Extended risk difference = $p_1 - p$
- Extended risk ratio or extended lift = p_1 / p
- Extended chance = $(1-p_1) / (1-p)$

group	benefit: denied	benefit: granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$

More on the *extended lift* later ...

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Other measures of discrimination

Differences of mean

Difference of regression coefficients

Rank tests

Mutual information (between outcome and protected attribute)

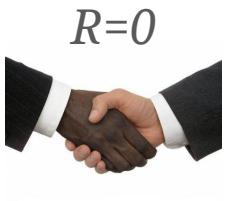
Unexplained difference (residuals of predictions built with non-protected attributes)

Consistency (comparison of prediction with nearest neighbors)

I. Žliobaitė (2015): A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

Example: regression coefficients

$$P = \beta_0 + \beta_1 R + \beta_2 Y + \beta_3 E + \beta_4 W + \varepsilon$$



P: price of a product on eBay, determined by …

R: buyer and seller have same race. Y: seller has a good rating. E, W: other.

β_1 positive means in-group discrimination, higher for low competition markets

Mean price paid: $6.37 if seller and buyer are same race, $5.93 (-7%) otherwise

J. Nunley, M. Owens, and S. Howard (2011): The effects of information and competition on racial discrimination: Evidence from a field experiment. Journal of Economic Behavior & Organization 80(3).

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Discrimination in specific contexts

Discrimination and privacy

Resources

Labor economic perspective

Who can discriminate

employers, other employees, customers

Issues

differences in wage, (un)employment rate, segregation into industries

Typical tool

regression of wage as a function of qualifications or seniority

A. Romei and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, pp 582-638.

Racial profiling

Typical tool

difference in proportions

Investigative tools

situational testing; natural experiments (e.g. observe other motorists in a stop zone to see if police stops blacks more than whites)

A. Romei and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, pp 582-638.

Credit and consumer markets

A traditional area for discrimination studies

Example of indirect discrimination is **redlining**: the practice of denying credit to all inhabitants of a certain area (when area is correlated with a protected attribute)

Typical tool for mortgages: comparison of mortgage rejection rates, mortgage price distributions, mortgage default rates

A. Romei and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, pp 582-638.

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Discrimination in specific contexts

Discrimination and privacy

Resources

Privacy and data protection are related

German *Ausländerzentralregister* (AZR) registers all foreigners living in Germany

Austrian national complained in court that AZR goes against free and non discriminatory movement of European citizens inside Europe

European Court of Justice decision in 2008:

- For determining right of residence, AZR is legal
- For demographics and statistics, only anonymized AZR data may be used
- For criminal investigations, no AZR data may be used

R. Gellert, K. de Vries, P. de Hert, and S. Gutwirth (2012). A Comparative Analysis of Anti-Discrimination and Data Protection Legislations. Chapter 4 of Discrimination and Privacy in the Information Society. Springer.

Privacy and data protection legislation differ

Privacy legislation cares about **one action** (storage of personal data)
independently of the consequences

Discrimination legislation cares about **one consequence** (unfair treatment)
independently of the mechanism

A connection between privacy and discrimination

Finding if people having attribute X were discriminated is like inferring attribute X from a database in which:

- the attribute X was removed
- a new attribute (the decision), which is based on X, was added

This is similar to trying to reconstruct a column from a privacy-scrubbed dataset

More on this relationship later …

Introduction and context

Motivation and examples of algorithmic bias

Sources of algorithmic bias

Legal definitions and principles of discrimination

Measures of discrimination

Discrimination in specific contexts

Discrimination and privacy

Resources

Some resources

- Presentations/keynotes/book
  - Suresh Venkatasubramanian: Keynote at ICWSM 2016
  - Ricardo Baeza: Keynote at WebSci 2016
  - Toon Calders: Keynote at EGC 2016
  - Discrimination and Privacy in the Information Society by Custers et al. 2013
- Groups/workshops/communities
  - Fairness, Accountability, and Transparency in Machine Learning (FATML) workshop and resources
  - Data Transparency Lab - http://dtlconferences.org/

Tools

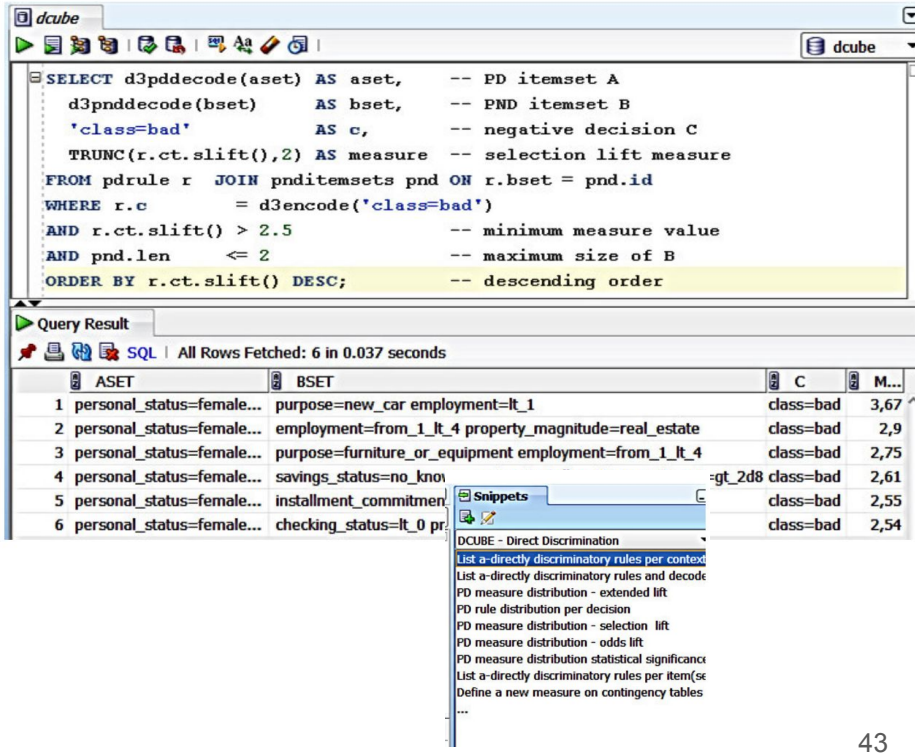
DCUBE (databases) http://kdd.di.unipi.it/dcube

S. Ruggieri, D. Pedreschi, F. Turini. Data mining for discrimination discovery. TKDD 4(2), May 2010

S. Ruggieri, D. Pedreschi, F. Turini. DCUBE: Discrimination Discovery in Databases. In SIGMOD 2010 Demos.

Adfisher (web ads)
https://github.com/tadatitam/info-flow-experiments

A. Datta, M.C. Tschantz, and A. Datta. Automated experiments on Ad privacy settings. In PETS, pp.92-112, 2015.



Part I: Introduction and context

Part II: Discrimination discovery

Part III: Fairness-aware data mining

Part IV: Challenges and directions for future research

Discussion and further questions

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

The discrimination discovery task at a glance

Given a large database of historical decision records,

find discriminatory situations and practices.

S. Ruggieri, D. Pedreschi and F. Turini (2010). DCUBE: Discrimination discovery in databases. In SIGMOD, pp. 1127-1130.

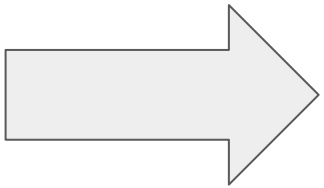
Discrimination discovery scenario

INPUT

Database of historical *decision records*

A *criterion* of (unlawful) discrimination

A set of *potentially discriminated* groups



OUTPUT

A subset of *decision records* and *potentially discriminated* people for which the *criterion* holds

D. Pedreschi, S. Ruggieri and F. Turini (2008). Discrimination-Aware Data Mining. In KDD, pp. 560-568.

Why is discrimination discovery hard?

Many different concepts regarding discrimination

- Including all the ones we mention in Part I of this tutorial

High dimensionality

- There are a huge number of possible contexts that may, or may not, be theater for discrimination.

Hidden indirect discrimination

- The features that may be the object of discrimination may not be directly recorded in the data
- The original data may have been pre-processed due to privacy constraints

The German credit score

A small dataset used in many papers about discrimination
(like Zachary's karate club for networks people)

N = 1,000 records of bank account holders

Class label: good/bad creditor (grant or deny a loan)

Attributes: *numeric/interval-scaled*: duration of loan, amount requested, number of installments, age of requester, existing credits, number of dependents; *nominal*: result of past credits, purpose of credit, personal status, other parties, residence since, property magnitude, housing, job, other payment plans, own telephone, foreign worker; *ordinal*: checking status, saving status, employment

German credit score dataset: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

Data mining provides a powerful tool for discovering discrimination in historical decision records

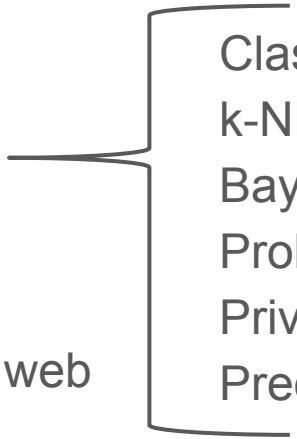
Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web



Classification rule mining	*Group* discr.
k-NN classification	*Individual* discr.
Bayesian networks	*Individual* discr.
Probabilistic causation	*Ind./Group discr.*
Privacy attack strategies	*Group* discr.
Predictability approach	*Group* discr.

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

Classification rule mining
k-NN classification
Bayesian networks
Probabilistic causation
Privacy attack strategies
Predictability approach

D. Pedreschi, S. Ruggieri and F. Turini (2008). Discrimination-aware data mining. In KDD'08.
D. Pedreschi, S. Ruggieri, and F. Turini (2009). Measuring discrimination in socially-sensitive decision records. In SDM'09.
S. Ruggieri, D. Pedreschi, and F. Turini (2010). Data mining for discrimination discovery. In TKDD 4(2).

Defining potentially discriminated (PD) groups

A subset of attribute values are perceived as potentially discriminatory based on background knowledge. Potentially discriminated groups are people with those attribute values.

Examples:

- Female gender
- Ethnic minority (*racism*) or minority language
- Specific age range (*ageism*)
- Specific sexual orientation (*homophobia*)

Discrimination and combinations of attribute values

Discrimination can be a result of several joint characteristics (attribute values) which are not discriminatory by themselves

Thus, the object of discrimination should be described by a conjunction of attribute values:

Itemsets

Association and classification rules

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database.

In a classification rule, Y is a class item and X contains no class items.

$$X \rightarrow Y$$

Direct discrimination

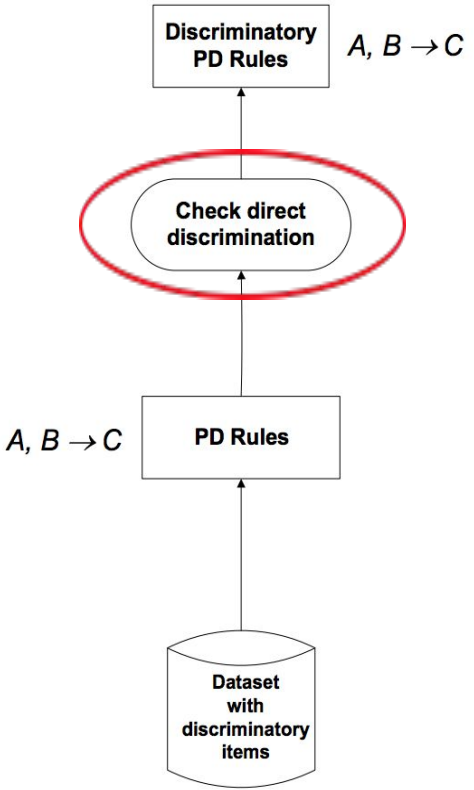
Direct discrimination implies rules or procedures that impose ‘disproportionate burdens’ on minorities

PD rules are any classification rule of the form:

A, B → C

where A is a PD group (B is called a "context")

Example:
gender="female", saving_status="no known savings"
→ credit=no

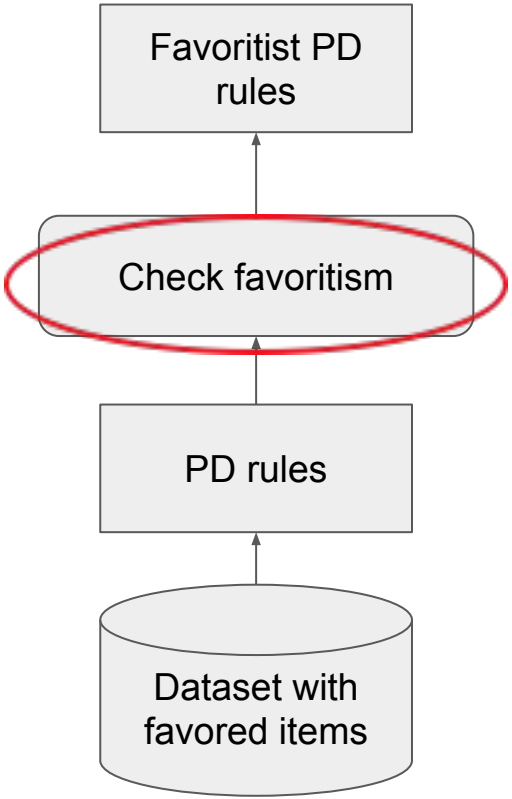


Favoritist PD rules

Is unveiled by looking at PD rules of the form

A, B → C

where C grants some benefit and A refers to a favored group.



Example:

gender="male", savings="no known savings"
→ credit=yes

Indirect discrimination

Indirect discrimination implies rules or procedures that impose ‘disproportionate burdens’ on minorities, though not explicitly using discriminatory attributes

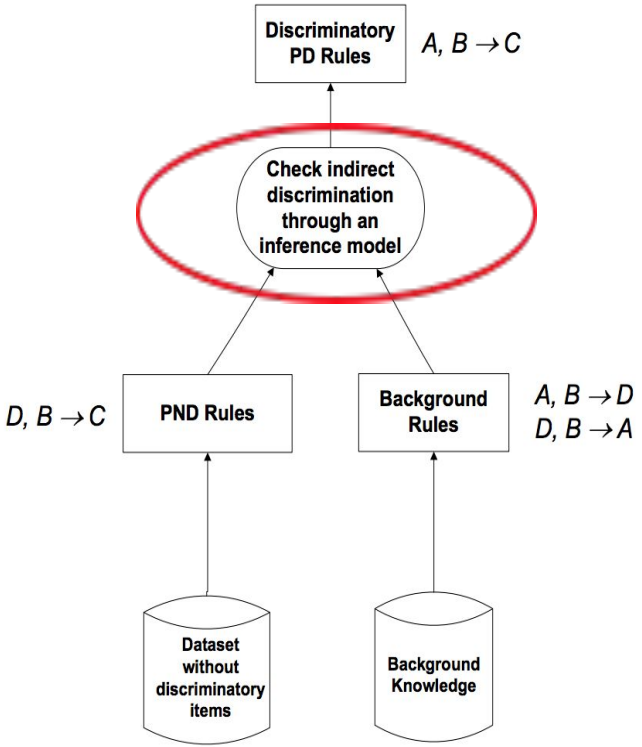
Potentially non-discriminatory (PND) rules may unveil discrimination, and are of the form:

D, B → C where D is a PND group

Example:

neighborhood="10451", city="NYC"

→ credit=no



Indirect discrimination example

Suppose we know that with high confidence:
(a) neighborhood=10451, city=NYC → benefit=deny

But we also know that with high confidence:
(b) neighborhood=10451, city=NYC → race=black

Hence:
(c) race=black, neighborhood=10451, city=NYC → benefit=deny

Rule (b) is background knowledge that allows us to infer (c), which shows that

rule (a) is indirectly discriminating against blacks

Evaluating PD rules through the extended lift

Remembering that conf(X → Y) = support(X → Y) / support(X)

We define the **extended lift with respect to B** of rule A, B → C as:

$$\text{elift}_B(A, B \rightarrow C) = \text{conf}(A, B \rightarrow C) / \text{conf}(B \rightarrow C)$$

The rules we care about are PD rules such that:

- A is a protected group (e.g. female, black)
- B is a context (e.g. lives in San Francisco)
- C is an outcome (usually negative, e.g., deny a loan)

The concept of α-protection

For a given threshold α, we say that PD rule A, B → C, involving a PD group A in a context B for an outcome C, is α-protective if:

$$elift_B(A, B \rightarrow C) = conf(A, B \rightarrow C) / conf(B \rightarrow C) \leq \alpha$$

Otherwise, we say that A, B → C is an α-discriminatory rule

Extension: strong α-protection is the symmetric version, where we ensure the numerator is greater than the denominator by using (1-conf)/(1-conf) if needed.

Relation of α-protection and group representation

For a given threshold α, we say that PD rule A, B → C,
involving a PD group A in a context B for a (usually bad) outcome C,
is α-protective if:

$$elift_B(A, B \rightarrow C) = conf(A, B \rightarrow C) / conf(B \rightarrow C) \leq \alpha$$

Note that:

$$elift_B(A, B \rightarrow C) = conf(B, C \rightarrow A) / conf(B \rightarrow A)$$

This means extended lift is the ratio between the proportion of the disadvantaged group A in context B for (bad) outcome C, over the overall proportion of A in B.

Direct discrimination example

Rule (a):

city="NYC"
→ benefit=deny
with confidence 0.25

Rule (b):

race="black", city="NYC"
→ benefit=deny
with confidence 0.75 **elift 3.0**

Additional (discriminatory) element increases the rule confidence up to 3 times.

According to *α*-protection method, if the threshold *α*=3 is fixed then the rule (b) is classified as discriminatory

Real-world example from German credit dataset

Fixing α=3:

(B) saving status = "no known savings" → credit = deny conf. 0.18

(A) personal status = "female div/sep/mar",
saving status = "no known savings" → credit = deny conf. 0.27 **elift 1.52**

Rule A is α-protective.

Real-world example 2 from German credit dataset

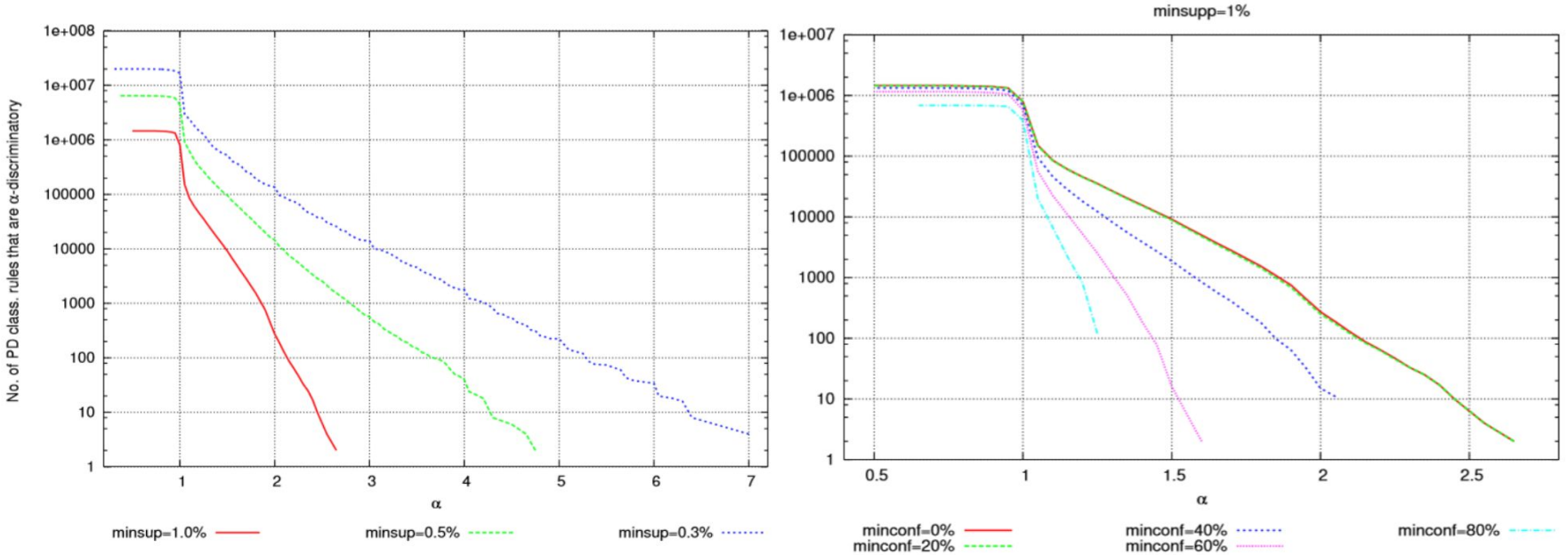
Fixing α=3:

(B) purpose = "used car" → credit = deny conf. 0.17

(A) age = "52.6+", personal status = "female div/sep/mar",
purpose = "used car" → credit = deny conf. 1.00 **elift 6.06**

Rule A is α-discriminatory.

α-Discriminatory rules in the German credit dataset



Data source: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Genuine occupational requirements

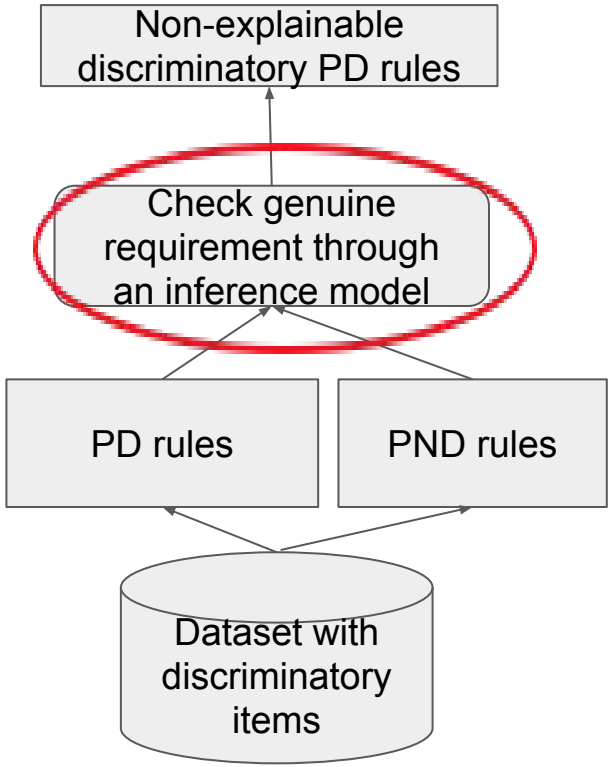
Supported by a PD rule of the form

$$A, B \rightarrow C$$

where C denies some benefit, we search for PND rules of the form

$$D, B \rightarrow C$$

such that D is a legitimate requirement, having the same effects of the PD rule



D. Pedreschi, S. Ruggieri and F. Turini (2009). Integrating induction and deduction for finding evidence of discrimination. In Proc. of International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.

Example: genuine occupational requirement

(a) [A] gender="female", [B] city="NYC" → [C] hire=no conf. 0.58

(b) [D] drive_truck="false", [B] city="NYC" → [C] hire=no conf. 0.81

(c) [A] gender="female", [B] city="NYC" → [D] drive_truck=false conf. 0.91

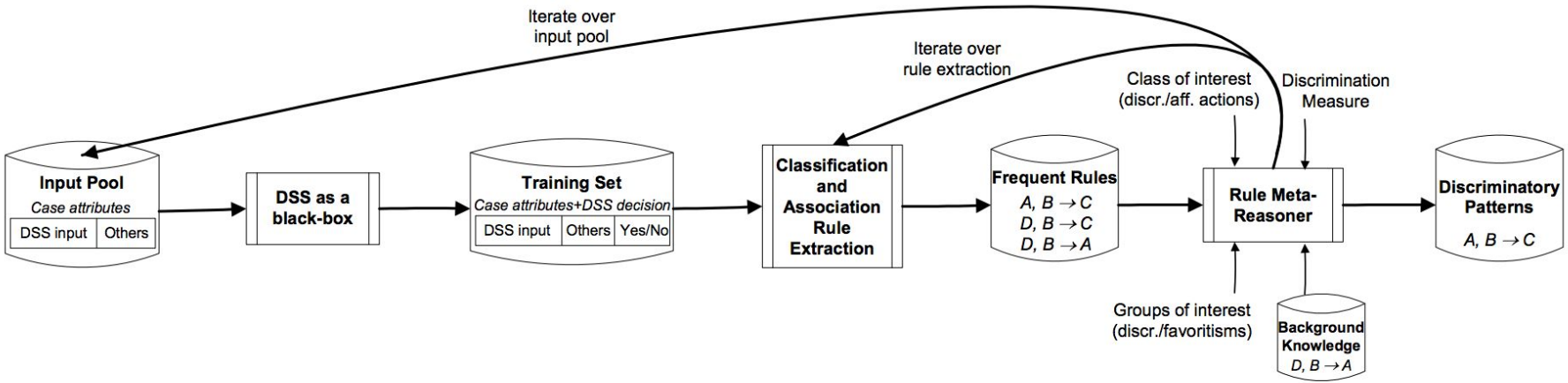
Let $p \in [0, 1]$. Classification rule (a) A, B → C with A being a PD attribute, is a p-instance of a PND rule (b) D, B → C, if:

- D is a legitimate ground for the decision (i.e., accepted by law),
- conf(D, B→ C) ≥ p · conf(A, B → C), and
- conf(A, B → D) ≥ p.

D. Pedreschi, S. Ruggieri and F. Turini (2009). Integrating induction and deduction for finding evidence of discrimination. In Proc. of International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.

Pipeline for analyzing discrimination

Reference model for analysing and reasoning on discrimination



D. Pedreschi, S. Ruggieri and F. Turini (2009). Integrating induction and deduction for finding evidence of discrimination. In Proc. of International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- **k-NN classification**
- Bayesian networks
- Probabilistic causation
- Privacy attack strategies
- Predictability approach

B. T. Luong, S. Ruggieri, and F. Turini (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. KDD'11

Limitations of classification rules approach

Legal limitations

- Undifferentiated groups are compared:
  - Do, e.g., women have the same characteristics of men they are compared with?
  - Or do they differ as per skills or other legally admissible reasons?

Limitations of classification rules approach (cont.)

Interpretational limitations

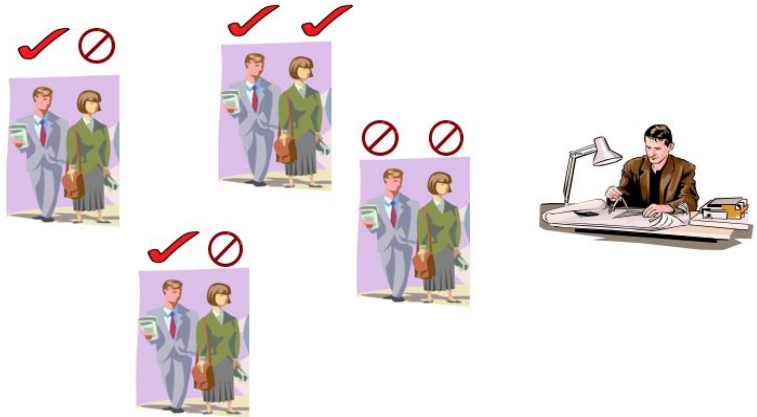
- Local contexts, possibly overlapping
- No global description of who is discriminated and who is not

Technical limitations

- Due to the use of frequent itemset mining (nominal attributes, nominal decisions) it requires discretization

Situation testing

- Legal approach for creating controlled experiments
- Matched pairs undergo the same situation, e.g. apply for a job
    - Same characteristics apart from the discrimination ground



k-NN as situation testing

Input: a dataset *R* of decision records

- For r ∈ *R*, dec(r) is the decision (discrete or continuous)
  - E.g., dec(r) is grant-benefit or deny-benefit
- P(R) is the set of protected-by-law groups, e.g., women
  - E.g., P(R) ={r ∈ *R* | r[gender]=female}
- U(R) = R \ P(R) is the rest of the dataset, e.g., men

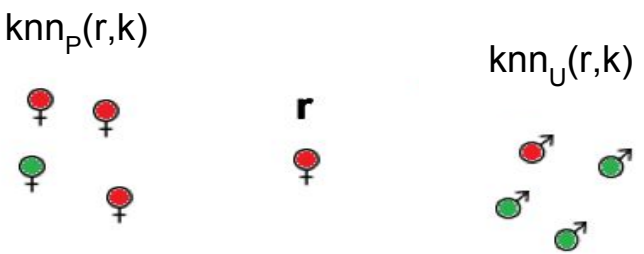
Relax the "identical characteristics" of situation testing to a "similar characteristics" by using a distance function *d*

- d(a,b) is defined over attributes that are legally admissible for the purpose of taking the decision

k-NN as situation testing (algorithm)

For r ∈ P(R), look at its k closest neighbors

- ... in the protected set
  - define p_1 = proportion with the same decision as r
- … in the unprotected set
  - define p_2 = proportion with the same decision as r
- measure the degree of discrimination of the decision for r
  - define diff(r) = p_1 - p_2 *(think of it as expressed in percentage points of difference)*



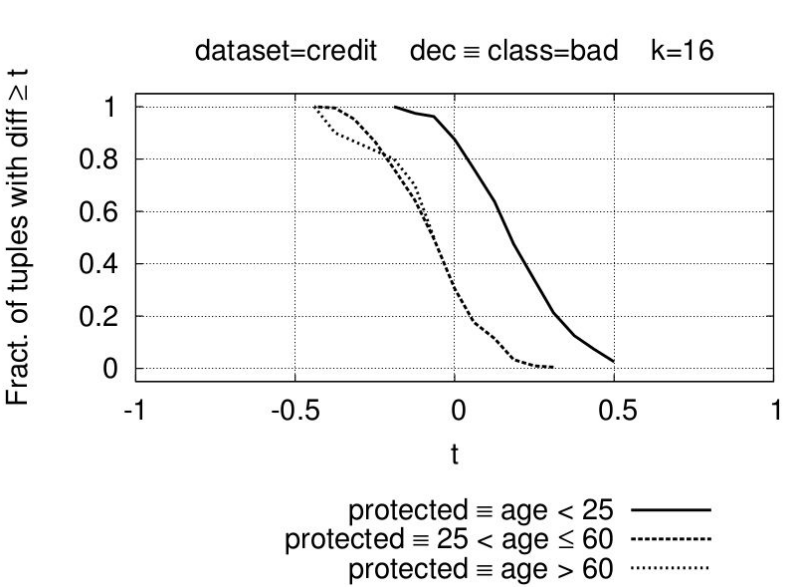
p_1 = 0.75

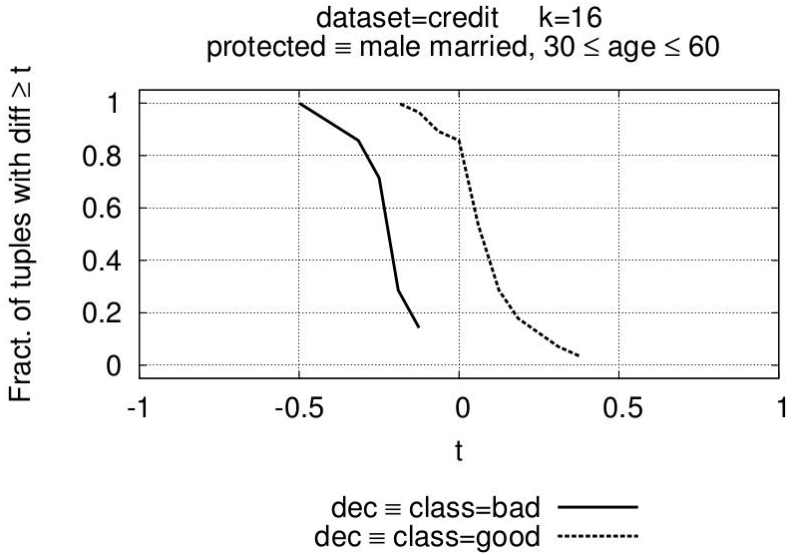
p_2 = 0.25

diff(r) = p_1 - p_2 = 0.50

k-NN as situation testing (results)

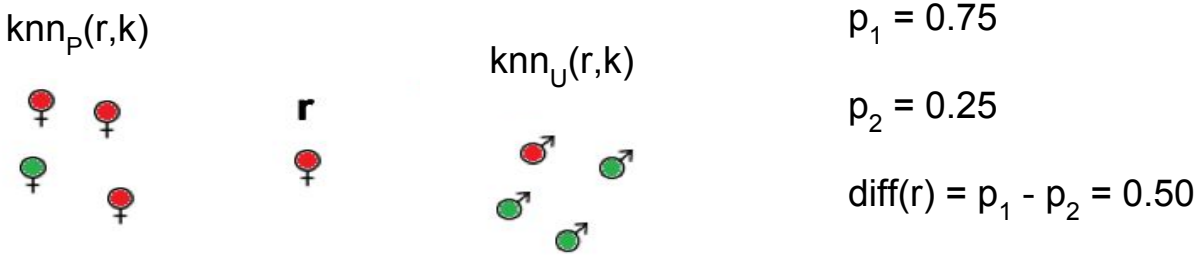
- If decision=deny-benefit, and diff(r) ≥ t, then we found discrimination around r





Characterizing discrimination using k-NN

- For r ∈ P(R), set a new attribute: "t-discriminated"
  - If dec(r) = deny-benefit and diff(r) ≥ t, t-discriminated(r) := TRUE
    - Otherwise t-discriminated(r) := FALSE
- Example: for t=0.3 the sample r below is classified as t-discriminated



$p_1 = 0.75$

$p_2 = 0.25$

$diff(r) = p_1 - p_2 = 0.50$

Characterizing discrimination using k-NN (cont.)

- To answer the question: under which conditions women were t-discriminated?
- We create a classifier with training set P(R), i.e. only protected people, and with class attribute *t-discriminated*

$$
\begin{array}{l}
\textbf{DiscoveryN}(\mathcal{R}, t)\ \{ \\
\quad \mathcal{L} = \emptyset \\
\quad \text{for } \mathbf{r} \in P(\mathcal{R})\ \{ \\
\quad\quad \text{if}(\ dec(\mathbf{r}) = \ominus \text{ and} \\
\quad\quad\quad\quad\quad\quad diff\,(\mathbf{r}) \geq t\ ) \\
\quad\quad\quad \mathbf{r}[\texttt{disc}] = \texttt{yes} \\
\quad\quad \text{else} \\
\quad\quad\quad \mathbf{r}[\texttt{disc}] = \texttt{no} \\
\quad\quad \mathcal{L} = \mathcal{L} \cup \{\mathbf{r}\} \\
\quad \} \\
\quad \text{build a classifier on } \mathcal{L} \\
\}
\end{array}
$$

Characterizing discrimination using k-NN (results)

- German credit dataset
  - protected = female non-single
  - 0.10-discriminated cases
- Decision tree model (C4.5)

```
num_dependents <= 1
|   credit_amount <= 2631: disc=yes (59.0/9.0)
|   credit_amount > 2631: disc=no (44.0/15.0)
num_dependents > 1: disc=no (6.0)

disc=yes:  Precision  0.847   Recall  0.769
```

Discriminated women had no dependents (children) and were asking for small amounts

- Classification rule model (RIPPER)

```
(credit_amount >= 3190) => disc=no (39.0/12.0)
(installment_commitment <= 2) and (residence_since >= 3)
                                        => disc=no (10.0/2.0)
 => disc=yes (60.0/9.0)

disc=yes:  Precision  0.85   Recall  0.785
```

Discriminated women were asking for small amounts and were either paying in many installments or had been resident for a short time

k-NN for discrimination prevention

- Goals of non-discriminating classifier:
  - Maximize classifier accuracy
    e.g., give credit to people who will pay
  - Minimize t-discriminated cases
    e.g., give credit to women if similar men would have been given credit
- Basic idea: t-correction of training set
  - Flip the labels from negative to positive for t-discriminated cases in the training set

$$
\begin{array}{l}
\textbf{PreventionN}(\mathcal{T},\ \mathcal{V},\ t)\ \{ \\
\quad \mathcal{T}' = \emptyset \\
\quad \text{for } \mathbf{r} \in \mathcal{T}\ \{ \\
\quad\quad \mathbf{r}' = \mathbf{r} \\
\quad\quad \text{if}(\ dec(\mathbf{r}) = \ominus \text{ and} \\
\quad\quad\quad\quad protected(\mathbf{r}) \text{ and} \\
\quad\quad\quad\quad\quad\quad diff(\mathbf{r}) \geq t\ ) \\
\quad\quad\quad\quad \mathbf{r}'[\mathtt{dec}] = \oplus \\
\quad\quad \mathcal{T}' = \mathcal{T}' \cup \{\mathbf{r}'\} \\
\quad \} \\
\quad \text{build classifiers on } \mathcal{T} \text{ and } \mathcal{T}' \\
\quad \text{compare them on } \mathcal{V} \\
\}
\end{array}
$$

(Discrimination prevention/mitigation is the focus of Part III …)

k-NN for discrimination prevention (results)

	No pre-processing		0.10-correction	
classifier	accuracy	0.10-discr.	accuracy	0.10-discr.
C4.5	85.60%	4.24%	84.94%	1.07%
Naïve Bayes	82.46%	4.06%	82.33%	2.23%
Logistic	85.28%	6.61%	84.70%	0.61%
RIPPER	84.42%	5.24%	83.98%	3.94%
PART	85.20%	12.62%	84.00%	2.3%

Only the training set changes, the testing set is fixed

Accuracy shows a small decrease

0.10-discrimination reduces substantially

Dataset: Income Predictions on Census Data (aka "Census Income" dataset) https://archive.ics.uci.edu/ml/datasets/Adult

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- **Bayesian networks**
- Probabilistic causation
- Privacy attack strategies
- Predictability approach

K. Mancuhan and C. Clifton (2014). Combating discrimination using bayesian networks. In Artificial Intelligence and Law, 22(2).

Let's go back to the classification-based approach

It hinges on computing $\text{elift}_B(A, B \rightarrow C) = \text{conf}(A, B \rightarrow C) / \text{conf}(B \rightarrow C)$

What are these quantities?

conf(A, B → C)

is the confidence on an outcome given a protected attribute and a context

conf(B → C)

is the confidence on an outcome given just the context

Extending the *elift* idea

Let A be the protected attributes, B the context (unprotected attributes)

Suppose we have a way of computing P(benefit=deny|...)

Suppose we use it to classify people: P(benefit=deny|...) ≥ θ ⇒ deny_benefit

If P(benefit=deny|A, B) / P(benefit=deny|B) ≥ α,

then we could say that the classifier is α-discriminatory

Extending the *elift* idea (cont.)

Let A be the protected attributes, B the context (unprotected attributes)

Let R be "redlining" attributes, which are correlated with the protected attributes A

Suppose we have a way of computing P(benefit=deny|...)

Suppose we use it to classify people: P(benefit=deny|...) ≥ θ ⇒ deny_benefit

If P(benefit=deny|A, B, R) / P(benefit=deny|B) ≥ α,

then we could say that the classifier is α-discriminatory

Bayesian networks

- Bayesian networks estimate the probability *P(A, B, C)* by capturing the conditional dependencies between the attributes within the sets *A* and *B*.
  - Bayesian networks can be used to estimate *P(A, B,C)* probability and the *P(C|A,B)* class probability can be derived from the Bayes theorem
  - Bayesian networks capture correlations between attributes
  - Bayesian networks are appropriate to define a decision process
- The *elift* can be extended to *belift* by calculating the numerator and the denominator probabilities with Bayesian networks
  - It identifies changes in decision making due to the usage of protected and redlining attributes
  - It captures both direct and indirect discrimination

Bayesian *elift* (*belift*)

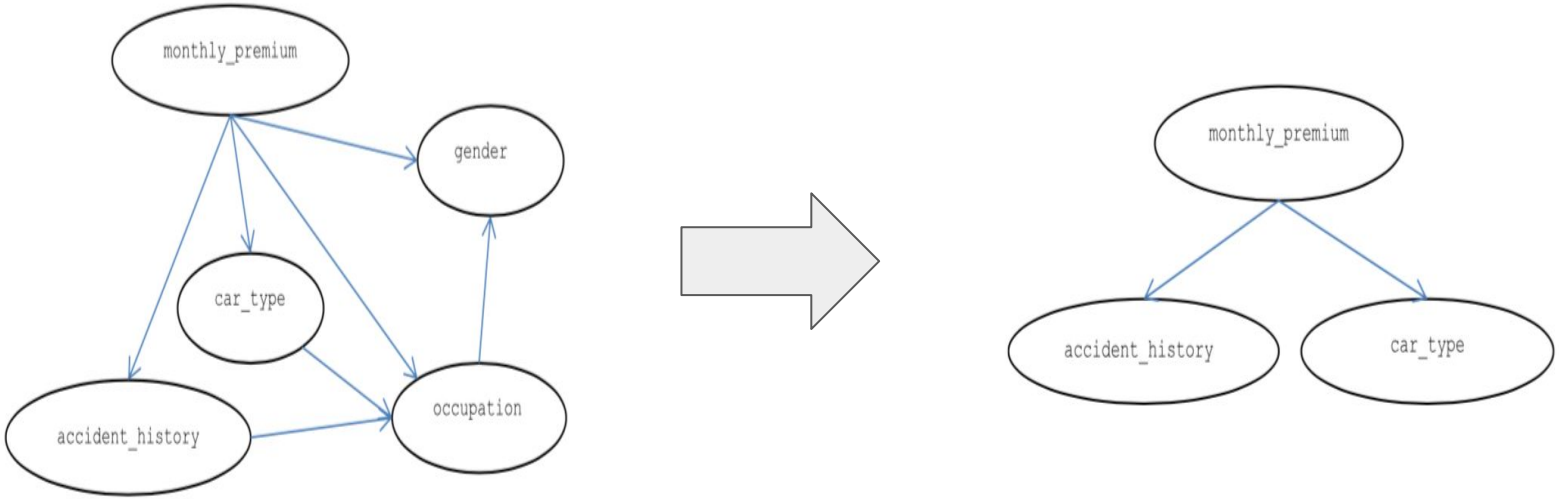
$$belift = \frac{P(C \mid a_1, a_2, \ldots, a_l, b_1, b_2, \ldots, b_m, r_1, r_2, \ldots, r_n)}{P(C \mid b_1, b_2, \ldots, b_m)}$$

This indicates how many times the usage of protected attributes (*A*) and the redlining attributes (*R*) increase the class probability for a given instance with respect to its probability using only non-protected attributes (B)

Bayesian networks for discrimination discovery

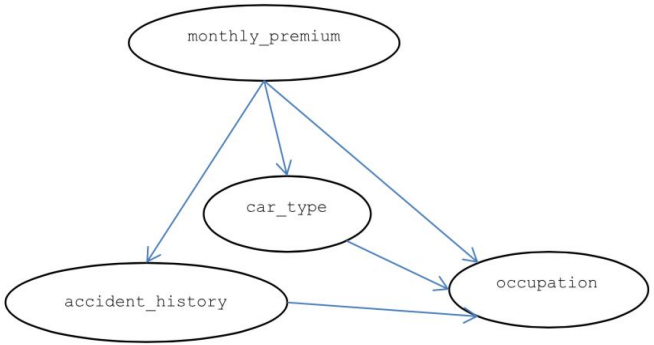
- Build a Bayesian network from a given set of instances D (*net*)
- Build a second Bayesian network (*relative_net*)
  - By removing protected attributes and any attribute directly connected to them in *net*
- For each instance i in D *net* and *relative_net*, compute probability of target class and apply decision threshold
  - If the decision is the same, ignore: instance was not discriminated
  - If the decision is different (i.e., **there is a "decision flip" when adding protected and redlining attributes**), and *belift* ≥ α, add instance i to the set of discriminated instances

Example: *net* and *relative_net*



Bayesian networks for discrimination prevention

- Build a Bayesian network from a given set of instances D (*net*)
- Build a non-discriminatory Bayesian network from net (*nd_net*)
  - **Remove** the protected attribute nodes, but **keep** the redlining attribute nodes
  - **Correct** the dataset by flipping the class label of discriminated instances
  - **Update** probability tables of *nd_net* using the corrected dataset
- Return the non-discriminatory Bayesian network (*nd_net*)



Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- Bayesian networks
- **Probabilistic causation**
- Privacy attack strategies
- Predictability approach

F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti (2015). Exposing the probabilistic causal structure of discrimination. arXiv:1510.00552.

Previous approaches

- Legal limitations
  - Any legally-valid proof of discrimination requires evidence of causality [Foster, 2004]
- Technical limitations
  - The state-of-the-art methods are essentially correlation-based
  - Spurious correlations can lead to false negatives and false positives
  - A Bayesian network would not be able to disentangle the direction of any causal relationship

Sheila R. Foster (2004). Causation in Antidiscrimination Law: Beyond Intent Versus Impact. Houston Law Review 41, p. 1469

To prove discrimination:

We need to assess discrimination as a causal inference problem from a database of past decisions, where causality can be inferred probabilistically

Suppes probabilistic causation theory (constraints)

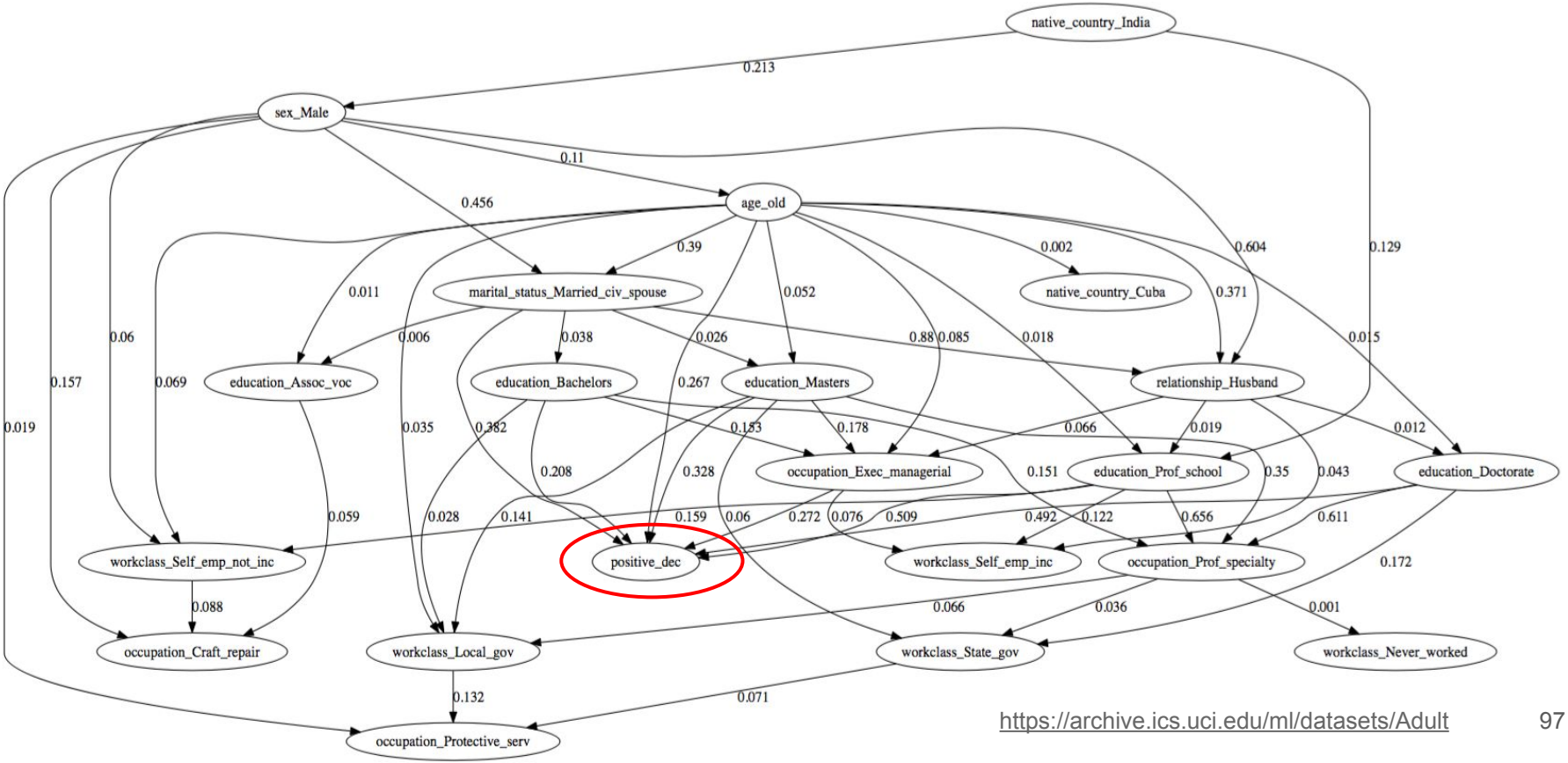
- Let *h* denote cause, *e* denote effect
- Temporal priority
  - Any cause must happen before its effect: $t_h < t_e$
- Probability raising
  - Cause must raise the probability of observing the effect: $P(e \mid h) > P(e \mid \neg h)$

The Suppes-Bayes Causal Network (SBCN)

- Represents the causal structures existing among the attributes in the data by a constrained Bayesian network:
  - each node represents an assignment attribute=value
  - each arc (u,v) represents the existence of a relation between u and v satisfying Suppes' constraints (temporal priority and probability raising)
  - each arc is labeled with a positive weight: p(u|v) - p(u|¬v)
- Reconstructed from the data using **maximum likelihood estimation (MLE)**, where we force the conditional probability distributions induced by the reconstructed graph to obey Suppes' constraints
  - *Network simplification* by extracting a minimal set of edges which best explain the data. This **regularization** is done by means of the Bayesian Information Criterion (BIC)

$$score_{\mathrm{BIC}}(D', G') = LL(D'|G') - \frac{\log s}{2}\dim(G').$$

SBCN example: Census Income dataset



https://archive.ics.uci.edu/ml/datasets/Adult

Graph representation allows interesting applications

Two largest communities detected using the Walktrap algorithm [Pons and Latapy 2006]

C_1
negative_dec, wc:Private, ed:Some_college, ed:Assoc_acdm, **ms:Never_married**, ms:Divorced, **ms:Widowed**, ms:Married_AF_spouse, oc:Sales, oc:Other_service, oc:Priv_house_serv, re:Own_child, re:Not_in_family, re:Wife, **re:Unmarried**, re:Other_relative, **ra:Black**, oc:Armed_Forces, oc:Handlers_cleaners, oc:Tech_support, oc:Transport_moving, ed:7th_8th, ed:10th, ed:12th, ms:Separated, ed:HS_grad,ed:11th, nc:Outlying_US_Guam_USVI_etc, nc:Haiti, **ag:young**, **sx:Female**, ra:Amer_Indian_Eskimo, nc:Trinadad_Tobago, nc:Jamaica, oc:Machine_op_inspct, ms:Married_spouse_absent, oc:Adm_clerical,
C_2
positive_dec, oc:Prof_specialty, wc:Self_emp_not_inc, ms:Married_civ_spouse, oc:Craft_repair,oc:Protective_serv, **re:Husband**, ed:Prof_school, wc:Self_emp_inc, **ag:old** , wc:Local_gov, **oc:Exec_managerial**, ed:Bachelors, ed:Assoc_voc, ed:Masters, wc:Never_worked, wc:State_gov, ed:Doctorate, **sx:Male**, nc:India, nc:Cuba

age (ag), education (ed), marital status (ms), native country (nc), occupation (oc), race(ra), relationship (re), sex (sx), workclass (wc)

P. Pons and M. Latapy (2006). "Computing communities in large networks using random walks." J. Graph Algorithms Appl. 10.2: 191-218.

Key idea: random walks

Start from the node we want to test (e.g. gender=female)

Let δ+ represent the positive outcome node (e.g., high salary)
δ- the negative outcome (e.g., low salary)

Perform *n* random walks starting from v

Let rw(v → δ-) be the number of random walks starting from v and ending in δ-

People with attribute v are discriminated if the fraction **rw(v→δ-) / n** is large

Result: top discriminated groups

Census Income Dataset

	$ds^-(v)$
relationship_Unmarried	1
marital_status_Never_married	0.996
age_Young	0.995
race_Black	0.994
sex_Female	0.98

German credit dataset

	$ds^-(v)$
residence_since_le_1d6	1
residence_since_gt_2d8	1
residence_since_from_1d6_le_2d2	1
age_gt_52d6	0.86
personal_status_male_single	0.791

Extension 1: genuine occupational requirements

Suppose v_{legal} is a genuine occupational requirement

e.g., having a truck driver license

The *fed* coefficient (fraction of explainable discrimination) is

$$rw(v \rightarrow v_{legal} \rightarrow \delta\text{-}) \,/\, rw(v \rightarrow \delta\text{-})$$

Result: fraction of explainable discrimination

Source node	Intermediate	$fed^{-}(v)$
race_Amer_Indian_Eskimo	education_HS_grad	0.481
sex_Female	occupation_Other_service	0.310
age_Young	occupation_Other_service	0.193
relationship_Unmarried	education_HS_grad	0.107
race_Black	education_11th	0.083

Extension 2: individual and subgroup discrimination

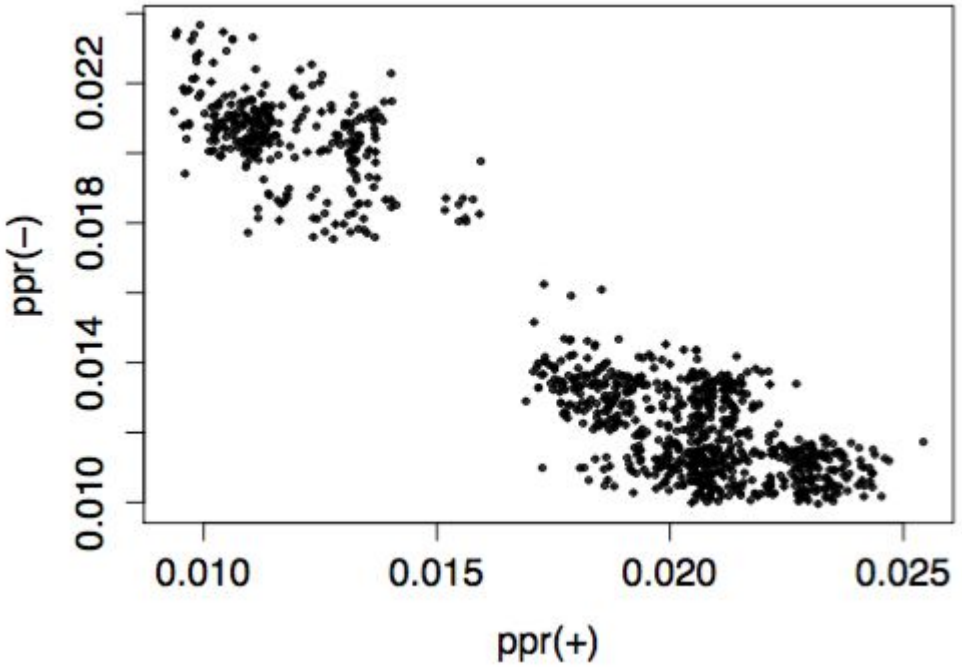
Suppose we have subgroup having attributes

$v_1, v_2, \ldots, v_n$

This subgroup can be of size 1, i.e., represent an individual

$$gds(v_1, v_2, \ldots, v_n) = \frac{ppr(\delta- \mid v_1, v_2, \ldots, v_n)}{ppr(\delta+ \mid v_1, v_2, \ldots, v_n) + ppr(\delta- \mid v_1, v_2, \ldots, v_n)}$$

Result: Individual discrimination (German credit)



Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- Bayesian networks
- Probabilistic causation
- **Privacy attack strategies**
- Predictability approach

S. Ruggieri, S. Hajian, F. Kamiran, X. Zhang (2014). Anti-discrimination analysis using privacy attack strategies. In PKDD'14.

Privacy attack strategies

Methods for direct discrimination discovery require that:

- The dataset explicitly contains an attribute denoting the PD group
  - Otherwise, we are in an indirect discrimination discovery setting
- The dataset has not been pre-processed prior to discrimination discovery
  - Otherwise, we are in a privacy-aware discrimination discovery setting

Discrimination discovery scenarios

- Indirect discrimination discovery
- Privacy-aware discrimination discovery
  (e.g., original data with all attributes is no longer available)
- Discrimination data recovery
  (e.g., original data with all attributes has been hidden from authorities)

Discrimination discovery scenarios: intuition

- Indirect discrimination discovery
- Privacy-aware discrimination discovery
- Discrimination data recovery

There is an interesting parallel between the role of the anti-discrimination authority in these three scenarios and the role of an attacker in private data publishing

The discrimination authority must infer who belongs to a protected group, based on the decisions that have been made.

Privacy attack strategies

- Combinatorial attacks based on Frèchet bounds inference
- Attribute inference attacks
- Minimality attacks

We will use risk difference (RD)

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1$$
$$p_2 = c/n_2$$
$$p = m_1/n$$
$$RD = p_1 - p_2$$

A discrimination table is α-protective (w.r.t. the *risk difference* measure RD) if $RD \leq \alpha$. Otherwise, it is α-discriminatory.

Indirect discrimination discovery

Think of g_1 and g_2 as *redlining* attributes (i.e., correlated with being protected)

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

unknown contingency table

rel. group	decision -	decision +	
g1	$\hat{a}$	$\hat{b}$	$\hat{n}_1$
g2	$\hat{c}$	$\hat{d}$	$\hat{n}_2$
	m_1	m_2	n

known contingency table

group	rel. group g1	rel. group g2	
protected	e	f	n_1
unprotected	g	h	n_2
	$\hat{n}_1$	$\hat{n}_2$	n

background knowledge contingency table

Exploit Frèchet bounds

$n_A = supp(A)$

$$min\{n_X, n_Y\} \geq n_{XY} \geq max\{n_X + n_Y - n_{X \cap Y}, 0\}$$

Indirect discrimination discovery

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

unknown contingency table

rel. group	decision -	decision +	
g1	$\hat{a}$	$\hat{b}$	$\hat{n}_1$
g2	$\hat{c}$	$\hat{d}$	$\hat{n}_2$
	m_1	m_2	n

known contingency table

group	rel. group g1	rel. group g2	
protected	e	f	n_1
unprotected	g	h	n_2
	$\hat{n}_1$	$\hat{n}_2$	n

background knowledge contingency table

We can decompose: $a = a_1 + a_2$, where a_1 are in g1, a_2 are in g2

Let X be (rel. group="g1" and dec="-")

Let Y be (rel. group="g1" and group="protected")

$$min\{\hat{a}, e\} \geq a_1 \geq max\{\hat{a} + e - \hat{n}_1, 0\} = max\{e - \hat{b}, 0\}$$

$\min\{n_X, n_Y\} \geq n_{XY} \geq \max\{n_X + n_Y - n_{X \cap Y}, 0\}$

Indirect discrimination discovery

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

unknown contingency table

rel. group	decision -	decision +	
g1	$\hat{a}$	$\hat{b}$	$\hat{n}_1$
g2	$\hat{c}$	$\hat{d}$	$\hat{n}_2$
	m_1	m_2	n

known contingency table

group	rel. group g1	rel. group g2	
protected	e	f	n_1
unprotected	g	h	n_2
	$\hat{n}_1$	$\hat{n}_2$	n

background knowledge contingency table

We do the same with a_2, and a similar decomposition for c, obtaining:

$$min\{\hat{a}, e\} + min\{\hat{c}, f\} \geq a \geq max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\}$$
$$min\{\hat{a}, g\} + min\{\hat{c}, h\} \geq c \geq max\{g - \hat{b}, 0\} + max\{h - \hat{d}, 0\}$$

$$RD \geq RDlb = \frac{max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\}}{n_1} - \frac{min\{\hat{a}, g\} + min\{\hat{c}, h\}}{n_2}$$

Lower bound

Uses only known and background data

Privacy-aware discrimination discovery

KNOWN | Known without IDs | UNKNOWN

ID	Education	Job	Dec	GID
r_1	Bachelors	Engineer	-	1
r_2	Bachelors	Engineer	+	1
r_3	Doctorate	Engineer	+	1
r_4	Bachelors	Writer	+	1
r_5	Master	Engineer	+	2
r_6	Doctorate	Writer	+	2
r_7	Bachelors	Dancer	-	2
r_8	Master	Dancer	-	2
r_9	Master	Dancer	-	3
r_{10}	Master	Lawyer	+	3
r_{11}	Bachelors	Engineer	-	3
r_{12}	Bachelors	Dancer	-	3

GID	Religion
1	Muslim
1	Christian
1	Jewish
1	Other
2	Muslim
2	Christian
2	Jewish
2	Other
3	Muslim
3	Christian
3	Jewish
3	Other

	decision		
education=bachelors	-	+	
religion=muslim	a	b	3
religion≠muslim	c	d	3
	4	2	6

When restricting to education=bachelors, n_1 becomes muslim bachelors; n_1 is always required as input

Frèchet bounds

n^i = *people in group* i $\quad$ n_1 = *total number of protected people*

$$\Sigma_i min\{1, n^i_-\} \geq a \geq n_1 - \Sigma_i min\{1, n^i_+\}$$

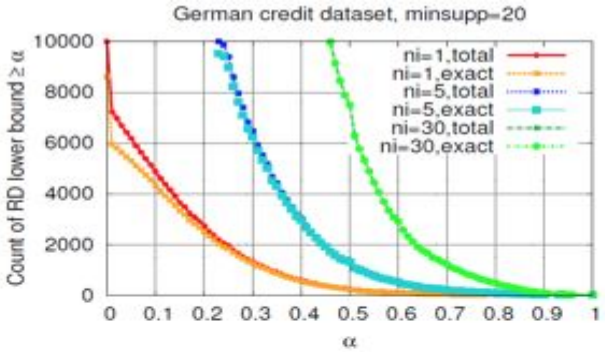
Discrimination data recovery

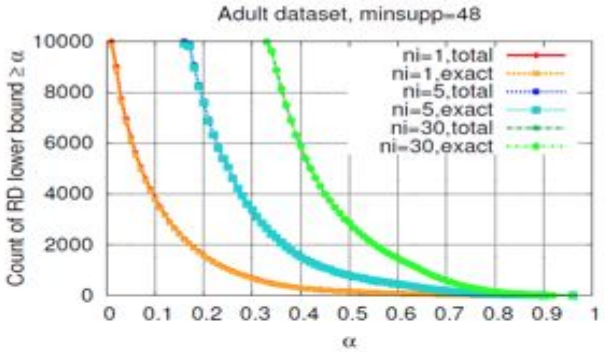
- Data owner hides discrimination by flipping a minimum number of decisions
  - Releases R' with changed decisions to have RD(R')=0, but we want R
- Suppose we know the original risk difference RD(R)
- Discrimination affects the tuples close to the decision boundary of a classifier
- To determine the decision boundary, we rank tuples of the protected and unprotected groups separately w.r.t. their positive decision probabilities accordingly to a classifier trained from the transformed data R'
- We don't fix all data at the same time, but iteratively, re-training at each iter.

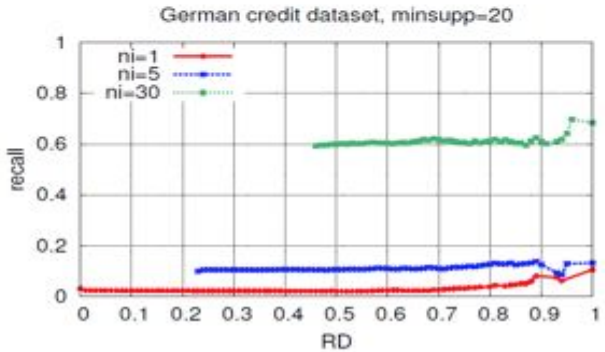
Sex	Ethnicity	Degree	Job Type	Dec	Prob
m	native	h.s.	board	+	98%
m	native	h.s.	board	+	98%
m	native	univ.	board	+	89%
m	*non-nat.*	*h.s.*	*health*	*-*	*47%*
m	non-nat.	univ.	health	-	30%

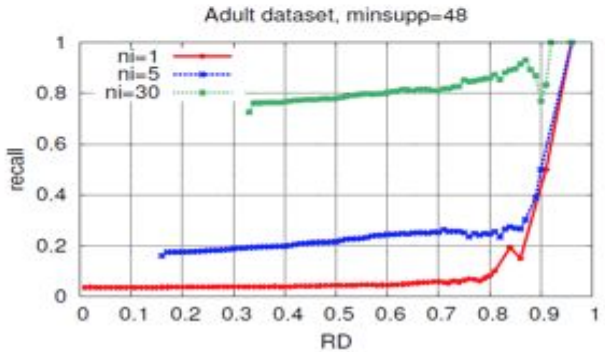
Sex	Ethnicity	Degree	Job Type	Dec	Prob
f	native	h.s.	board	+	93%
f	native	none	health	+	76%
f	*native*	*h.s.*	*edu.*	*+*	*51%*
f	non-nat.	univ.	edu.	-	2%
f	non-nat.	univ.	edu.	-	2%

Results: indirect discrimination discovery









Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- Bayesian networks
- Probabilistic causation
- Privacy attack strategies
- **Predictability approach**

M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015). Certifying and removing disparate impact. In KDD'15.

Disparate impact

Measured using risk ratio p_1/p_2

$$\frac{\Pr(\text{decision=deny\_benefit} \mid \text{group=protected})}{\Pr(\text{decision=deny\_benefit} \mid \text{group=unprotected})}$$

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

Feldman et al. note this is:

1/LR+ = (1-specificity)/sensitivity

LR+ is the likelihood ratio of the positive class

US Equal Employment Opportunity Commision (EEOC) says $p_1/p_2 \leq 0.8$ means disparate impact

M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015). Certifying and removing disparate impact. In KDD'15.

Disparate impact as predictability

Alice runs an algorithm on D = (X, Y), with X protected, Y unprotected

Bob receives (D, C), with C the outcomes

If Bob can predict X based on D\X, C,

then the algorithm was discriminatory

Essentially, we want to know if the decisions are leaking information about the protected attributes

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

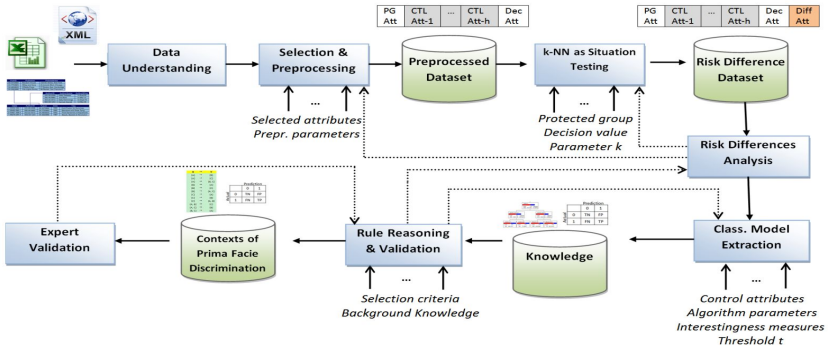
Crime suspect dataset

- **Dataset**: a real world dataset of Statistics Netherlands, which is a census body in the Netherlands.
- **Application**: the use of classifiers for predicting whether an individual is a crime suspect, or not, to support law enforcement and security agencies.
- **Results**:
  - The results show that **discrimination does exist** in real world datasets and blind use of classifiers learned over such datasets can **exacerbate** the discrimination problem.
  - It demonstrates that discrimination-aware classification methods can **mitigate the discriminatory effects** and that they lead to rational and legally acceptable decisions.

F. Kamiran, A. Karim, S. Verwer and H. Goudriaan (2012). Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. In ICDMW, pp. 370-377.

Scientific project evaluation

- **Dataset**: scientific research proposals
  - In 2008, the Italian Ministry of University and Research published a call for scientific research projects under the Basic Research Investment Fund (FIRB) reserved to young scientists.
- **Method**: classification rule mining and k-NN classification approaches to discover gender bias in scientific research funding



A. Romei, S. Ruggieri and F. Turini (2013). Discrimination discovery in scientific project evaluation: A case study. In Expert Systems with Applications, 40(15).

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

Discrimination in online services is not rare

- Non-black hosts can charge ~12% more than black hosts in Airbnb
  - Edelman, Benjamin G. and Luca, Michael, Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054.
- Price steering and discrimination in many online retailers
  - Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson: Measuring Price Discrimination and Steering on E-commerce Web Sites. Proc. of IMC. Vancouver, Canada, November 2014.
- China is about 21% larger by pixels when shown in Google Maps for China
  - Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson: MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. Proc. of WWW. Montreal, Quebec, Canada, April 2016

Examples from WWW'16 tutorial by Strohmeier, Aiello, Wagner, and Weber: https://sites.google.com/site/csswwwtutorial/slides

Discrimination discovery on the web

How do we measure the “fairness of the web”?

- Need to model/understand user browsing behavior
- Evaluate how web sites respond to different behavior/attributes
- Cope with noisy measurements

Moritz Hardt (2013). Fairness through Awareness Presentation (slides).

Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. (2012). Detecting price and search discrimination on the internet. In Proceedings of the 11th ACM Workshop on Hot Topics in Networks, pages 79–84. ACM

The importance of being Latanya

Names used predominantly by black men and women are much more likely to generate ads related to arrest records, than names used predominantly by white men and women.

Discovered by researcher Latanya Sweeney.

L. Sweeney (2013). Discrimination in online ad delivery. Queue, 11(3). See also N. Newman (2011) in Huffington Post.

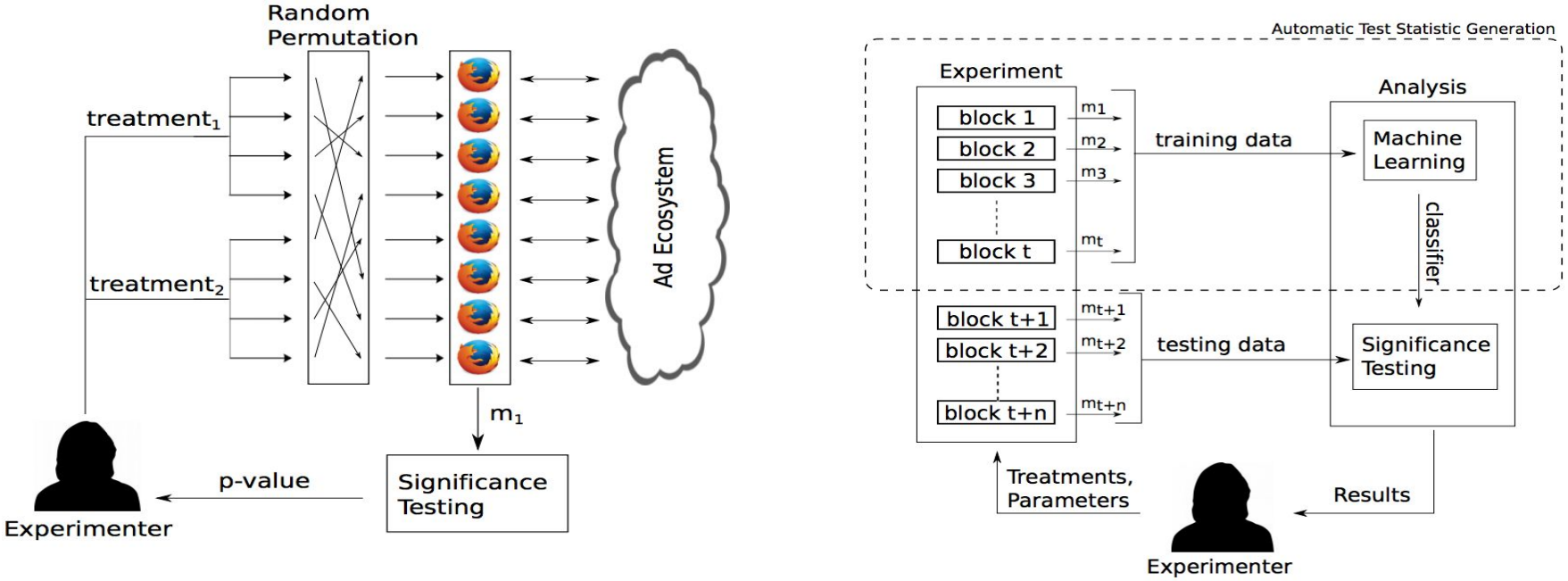
The dark side of Google Ads

AdFisher: tool to automate the creation of behavioral and demographic profiles.

Used to demonstrate that setting gender = female results in less ads for high-paying jobs.

A. Datta, M. C. Tschantz, and A. Datta (2015). Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015(1):92–112.

Adfisher tool to uncover discrimination in ads



A. Datta, M. C. Tschantz, and A. Datta (2015). Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015(1):92–112.