Prediction with Model-based Neutrality

Kazuto Fukuchi[*1], Jun Sakuma[*1], Toshihiro Kamishima[*2]
*1 Dept. of Computer Science, Graduate school of SIE, University of Tsukuba
*2 National Institute of Advanced Industrial Science and Technology (AIST)

ECML PKDD 2013

Discrimination in Prediction

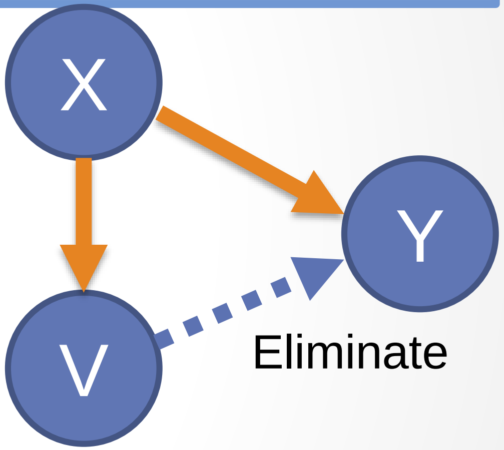
If the predictions are highly dependent on the sensitive attribute, the predictions might be discriminatory.

- Predictions have a significant impact on our lives.
  - E.g. hiring-decision, insurance rate, credit administration
- Discrimination caused by highly dependent on the sensitive attributes
  - Sensitive attributes: gender, race, ethnicity
- Discrimination must not be
  - lose your credit
  - be a violation of the law

Red-lining Effect [Calders 10]

Elimination of the sensitive attributes does not reduce discrimination.

- Indirect effects are remaining
  - If X is highly dependent on V, Y is dependent on V through X.



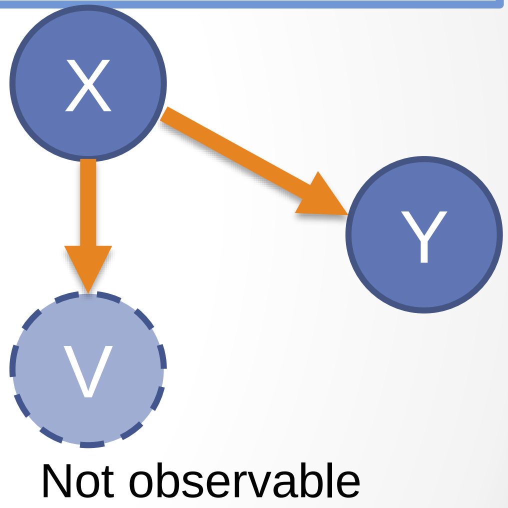
X: input variable (age, career, address)
Y: output variable (hiring-decision)
V: viewpoint variable (race, gender)

To ensure the fairness, we need aggressive way.

Effect from hidden attributes

Hidden viewpoint variable (sensitive attributes) causes discrimination if they are predictable.

- If V are predictable from X, X and V are highly correlated
- Correlationship between X and V causes discriminatory



Objective

Assume: If viewpoint variable is predictable, we could obtain the predictive model of the viewpoint variable

Ensure the neutrality of the model

Model-based neutrality could treat hidden viewpoint variable.

This presentation:

- Consider neutrality of the model
- Maximum Likelihood Estimation with neutrality
- Evaluate the performance

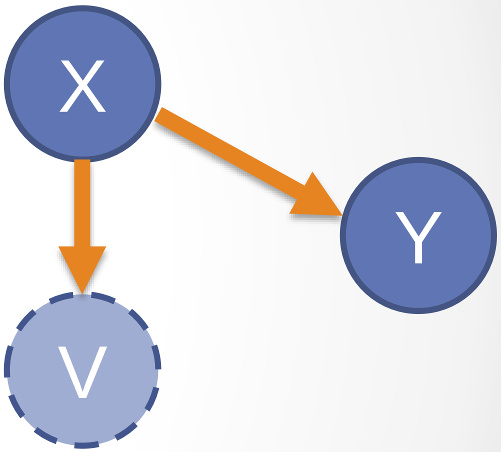
Fairness/Discrimination-aware Data Mining

- CV2NB [Calders 10]
  - Evaluate fairness with CV Score
    $$\Pr(y_+|v_+) - \Pr(y_+|v_-)$$
  - Modified parameters after learning with Naïve Bayes
- Prejudice Remover [Kamishima 12a]
  - Evaluate fairness with prejudice(mutal information)
    $$\mathrm{PI} = I(Y;V)$$
  - Reduce discrimination with regularizer
- Both of these methods assume the value of the viewpoint variable is explicitly provided.

Hidden effects are not considered in this works

Problem settings

Define two predictive models: f(Y|X;Θ), g(V|X)



- f(Y|X;θ) : the model of the output variable
- g(V|X) : the model of the viewpoint variable
  - g(V|X) is given

- Maximum likelihood estimation with neutralization

$$\max \ L(\theta)$$

subject to $f(Y|X;\theta)$ is neutral from $g(V|X)$

$L(\theta)$: log likelihood

η-Neutral

Neutrality between two models

η-Neutral

Given η ≥ 0, the probability distribution Pr(X,Y,V) is η-neutral if

$$\forall\, y \in \mathcal{Y}, v \in \mathcal{V}, \frac{\Pr(y, v)}{\Pr(y)\Pr(v)} \le 1 + \eta.$$

Defined by dependency between Y, V If Y, V is independent $\frac{\Pr(y,v)}{\Pr(y)\Pr(v)} = 1$

Evaluate most dependent pair of the Y, V

η-Neutral Model

Condition of the two models is η-neutral:

Condition of the η-Neutral

Model $M(X, Y, V) = \Pr(X)\, f(Y|X; \theta) g(V|X)$ is η-neutral if

$$\int_{x \in \mathcal{X}} \Pr(x)\, f(y|x;\theta)\big(g(v|x) - (1+\eta)\bar{g}(v)\big)dx \leq 0.$$

$\Pr(x)$ cannot be obtain $\Rightarrow$

Approximate with the frequency distribution (Empirical η-neutral)

Condition of the Empirical η-Neutral

$$N_\eta(y, v) = \sum_{x \in \mathcal{D}} f(y|x;\theta)\big(g(v|x) - (1+\eta)\bar{g}(v)\big) \leq 0$$

η-Neutral Maximum Likelihood Estimation

Maximum likelihood estimation with empirical η-neutrality constraints

$$\min_{\theta} L(\theta)$$
$$\text{s.t. } N_\eta(y, v) \leq 0 \; \forall y \in \mathcal{Y}, v \in \mathcal{V}$$

$L(\theta)$: Negative log likelihood
$N_\eta\,(y, v)$: Empirical η-neutrality

In experiments, we use following two models:

- Logistic Regression
- Linear Regression

Unfortunately, the constraints are not convex
Convexifying is future work

Any model of output variable $f(Y|X;\theta)$ can be used

Settings: Classification

	Case 1			Case 2		
	learning	neutrality	evaluate	learning	neutrality	evaluate
Existing methods	x, v	v	$\hat{y}, v$	$x, \hat{v}$	$\hat{v}$	$\hat{y}, v$
proposal	x, v	$g(v\|x)$	$\hat{y}, v$	x	$g(v\|x)$	$\hat{y}, v$

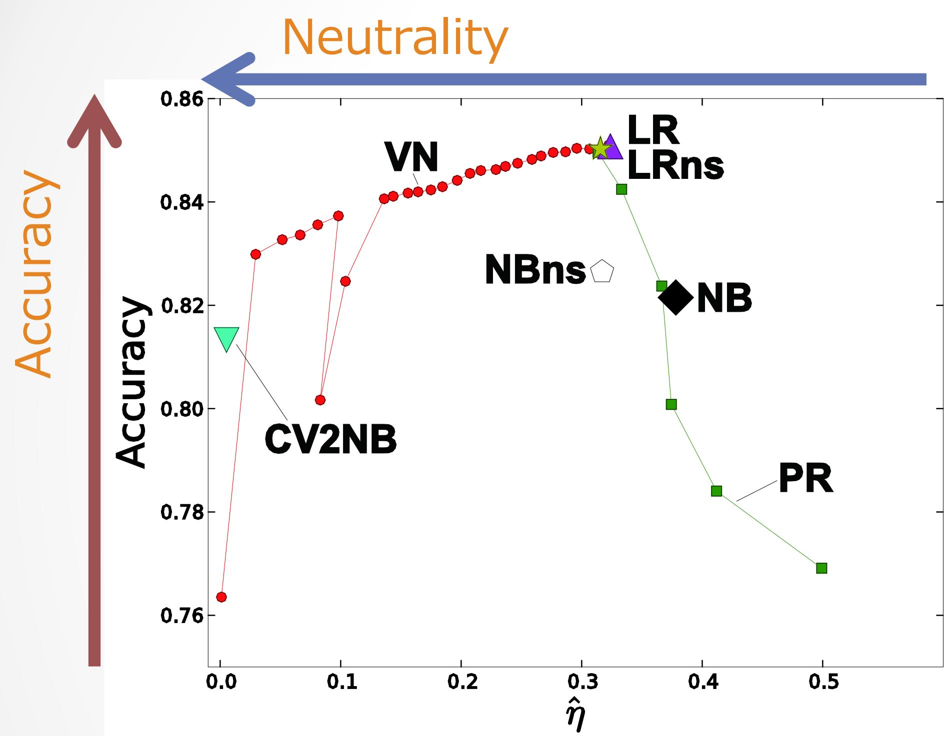
learning : training input data
neutrality : data of ensuring the neutrality
evaluate : data of calculating the indexes

$\hat{y}, \hat{v}$ is estimated $\hat{y} = \arg\min_{y} f(y|x;\theta)$, $\hat{v} = \arg\min_{v} g(v|x)$

Case 1 : Given the viewpoint variables
Case 2 : Given only the model of the viewpoint variable

notion)In Case 2, existing methods use estimated value $\hat{v}$ in learning, but true value v in evaluation.

Result: Case 1



No neutralize

- NB — Naïve Bayes
- LR — Logistic regression

Baseline

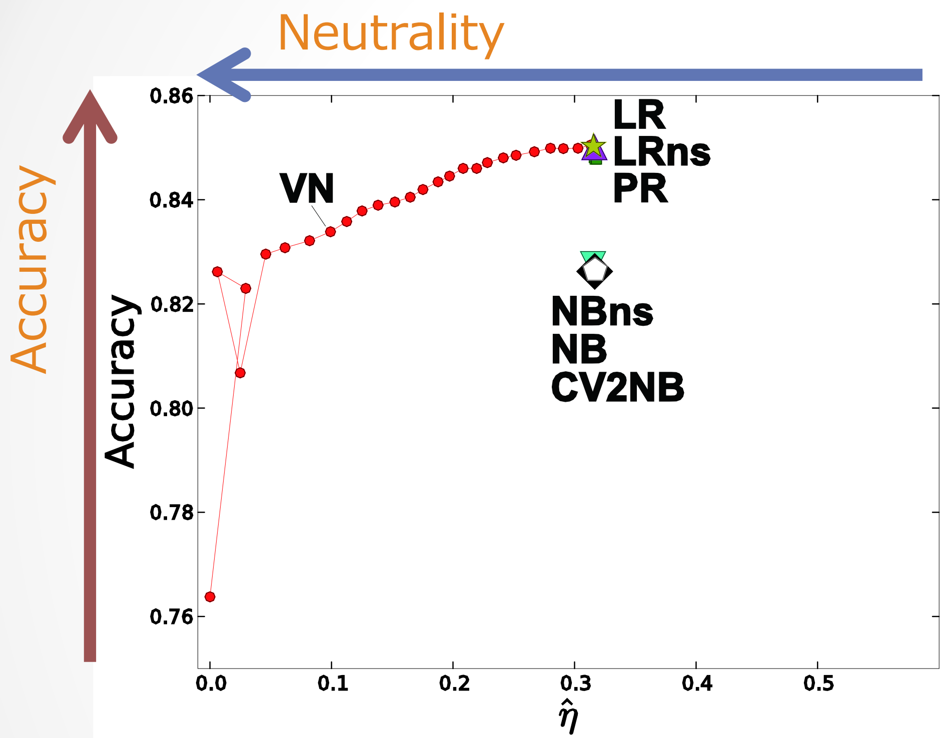
- NBns — NB without viewpoint
- LRns — LR without viewpoint

Comparer methods

- CV2NB — [Calders 10]
- PR — [Kamisima 12a]
- VN — η-neutral LR

- CV2NB achieves good performance
- PR cannot achieve lower neutrality
- VN achieves good trade off rate, though worse CV2NB
  - Enable to control trade off by parameter η

Result: Case 2



No neutralize

◆	NB	Naïve Bayes
▲	LR	Logistic regression

Baseline

⬠	NBns	NB without viewpoint
★	LRns	LR without viewpoint

Comparer methods

▼	CV2NB	[Calders 10]
■	PR	[Kamisima 12a]
●	VN	η-neutral LR

- CV2NB, PR did not work well
- VN achieves good performance
  - Enable to control trade off by parameter η

Settings: Regression

- Dataset: Housing dataset (UCI Repository)
  - Input: 12 attributes
  - Output: MEDV (median value of owner-occupied homes, in $1000s)
  - Viewpoint: LSTAT (% lower status of the population)

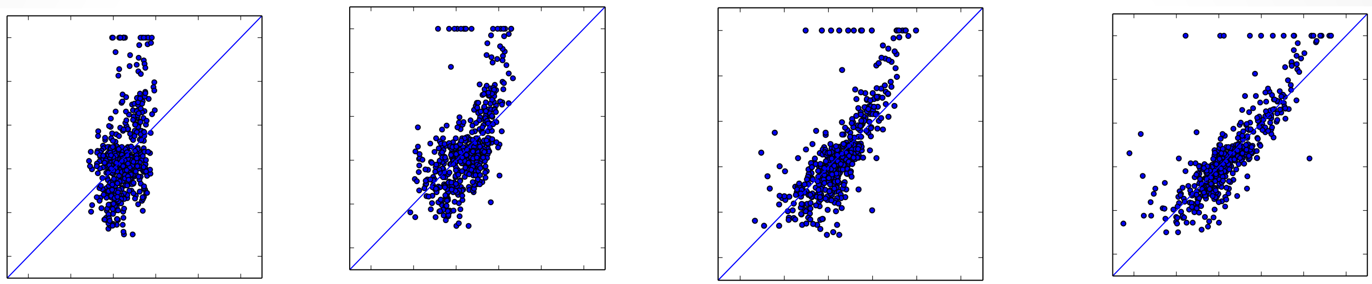
- Evaluation

  Accuracy: root-mean-square error (RMSE)
  Neutrality: η

Result: Regression

Good accuracy, if plots arrange on the diagonal line

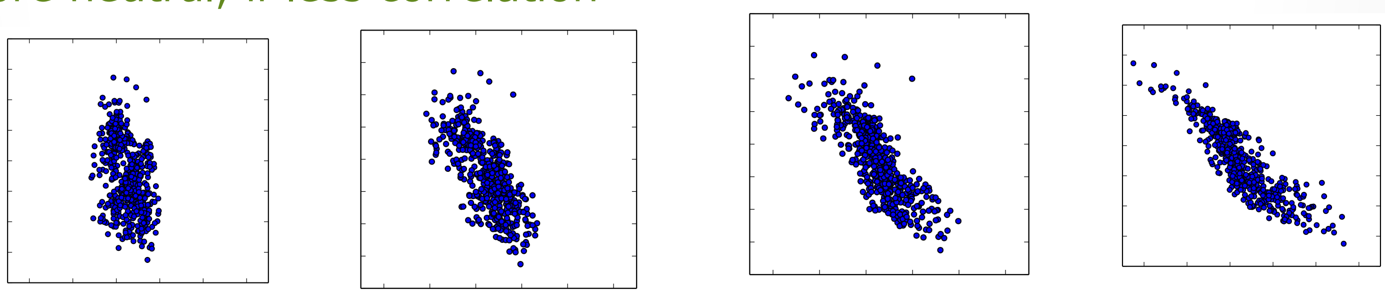
y-$\hat{y}$



RMSE=8.49 RMSE=7.54 RMSE=6.26 RMSE=5.25

More neutral, if less correlation

$\hat{y}$-$\hat{v}$



η = 1.0 η = 3.0 η = 10.0 no neutralization

$\hat{y}, \hat{v}$ is estimated $\hat{y} = w^T x, \hat{v} = w_v^T x$

- To ensure high neutrality, the output is a constant value
  - Enable to control trade off by parameter η

Conclusion & Future Works

We propose a framework for learning probabilistic model with model-based neutralization.

Contribution

- Neutrality of the probabilistic model
- Maximum likelihood estimate with η-neutral constraint
- Experimental results show our method achieves neutralization even when only a model is provided

Future Works

- To convexify η-neutral constraint