Path-Specific Causal Reasoning for Fairness-aware Cognitive Diagnosis

Dacao Zhang, Kun Zhang*, Le Wu, Mi Tian, Richang Hong, Meng Wang
School of Computer Science and Information Engineering
Hefei University of Technology

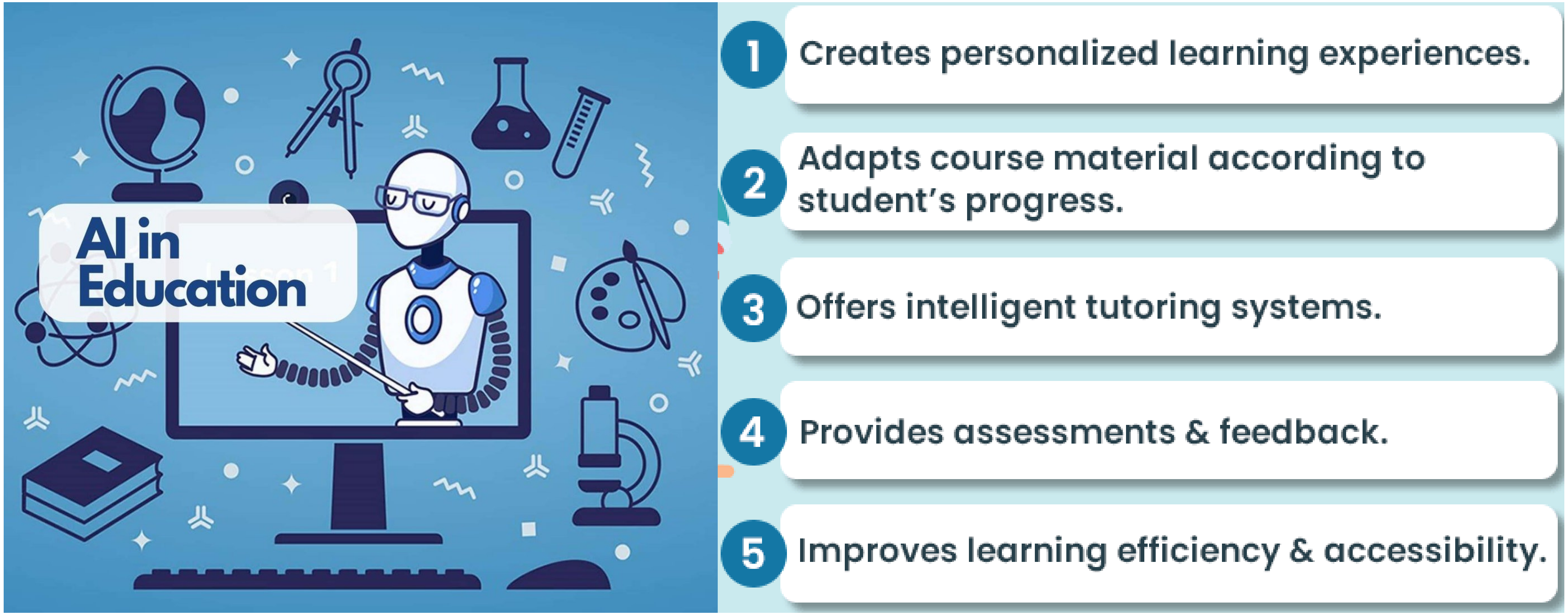
Reporter：Kun Zhang
Email: zhkun@hfut.edu.cn

Sunday, September 1, 2024

Outline

Background——AI and Education

- Artificial Intelligence (AI) enables the rapid development of personalized learning, offering significant advantages for learner from the following aspect:



Cognitive Diagnosis plays an important role in the application of intelligent education.

Background——Cognitive Diagnosis

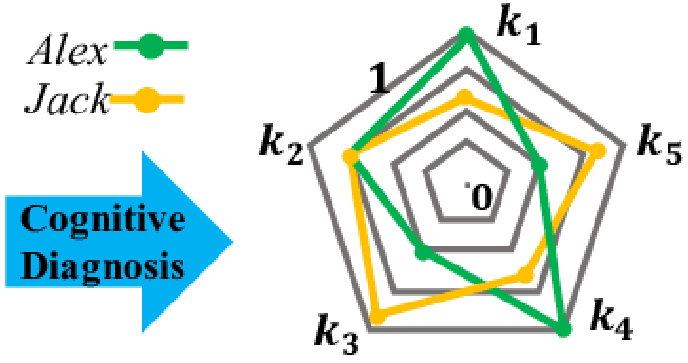
Cognitive Diagnosis (CD):

- Using a model to predict student proficiency level on knowledge concepts based on historical student-exercise logs, Q-matrix, and other collected information.
- Fundamental task in multiple Intelligent Education areas.

Exercises	e_1	e_2	e_3	e_4
Alex	✓	?	✗	✓
Jack	✗	✓	✓	?

Concept / Exercise	k_1	k_2	k_3	k_4	k_5
e_1	1	0	0	1	0
e_2	0	0	1	0	1
e_3	0	1	1	0	0
e_4	1	0	0	1	1

Cognitive Diagnosis

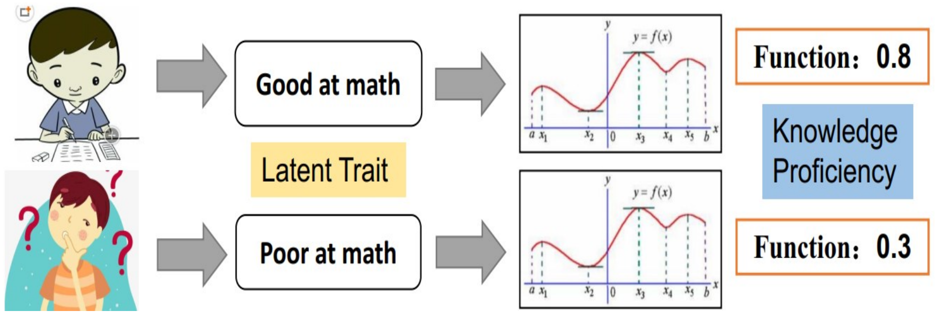


Legend

k_1: *Matrix product* k_2: *Addition*

k_3: *Matrix inversion* k_4: *Division*

k_5: *Conjugated matrices*



Student's skill proficiency Modeling

duolingo MAX

Khan Academy

学而思

Outline

Preliminary——Traditional Diagnostic Models

- **IRT, MIRT**: scalar or latent vectors for students and exercises; logistic like interaction function

$$P(R_{uv}=1|\theta_u,a_v,b_v)=\frac{1}{1+\exp(-1.7a_v(\theta_u-b_v))}$$

Skill proficiency Discrimination Difficulty

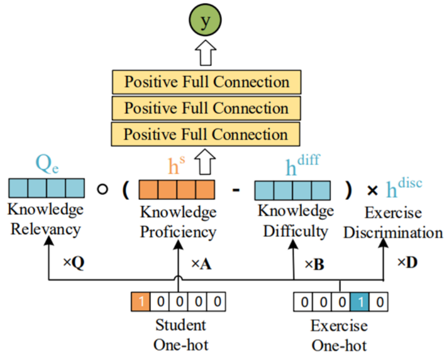
- **DINA:** binary vectors for students and exercises; conjunctive assumption in interaction function

$$P(R_{ij}=1|\alpha_i)=(1-s_j)^{\eta_{ij}}g_j^{1-\eta_{ij}}$$

Skill proficiency vector Slip Guess

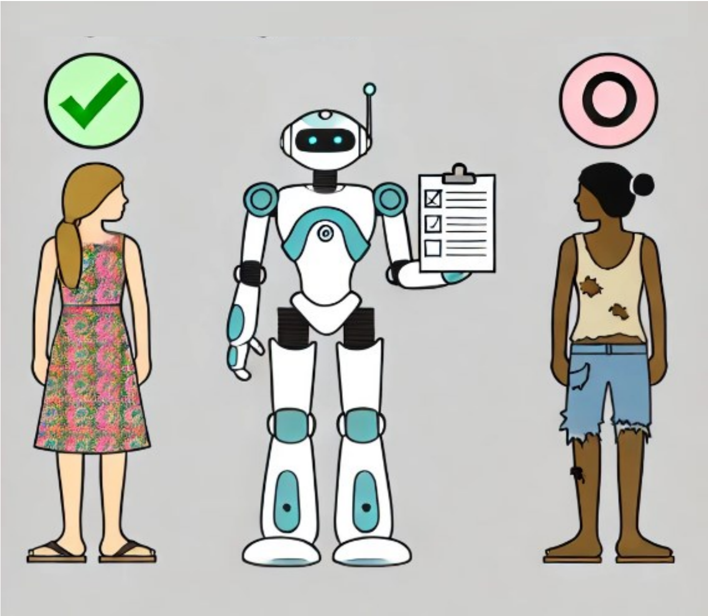
- **NCDM, KaNCD**: Use neural networks for modeling complex, nonlinear interactions

$$\boldsymbol{x}=\boldsymbol{Q}_e\circ(\boldsymbol{h}^s-\boldsymbol{h}^{diff})\times h^{disc}$$



Bias in Cognitive Diagnosis

Existing model relies on spurious associations between students' sensitive attributes and outcomes for prediction.

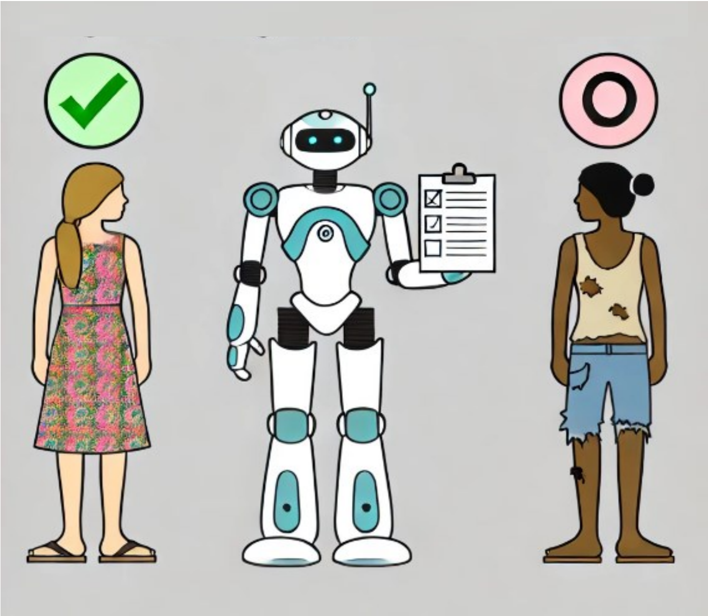


Model	Family Wealth			Country	
	Poor	Average	Wealth	Australia	Brazil
Data statistics	0.4736	0.5448	0.6434	0.5516	0.3888
NCD	0.5140	0.5861	0.6789	0.5913	0.3293
KaNCD	0.4778	0.5589	0.6643	0.5650	0.3025
NCD-*PSCRF*	0.5545	0.5798	0.6155	0.5824	0.3321
KaNCD-*PSCRF*	0.5286	0.5581	0.6271	0.5680	0.3026

Sensitive Attributes would be used for prediction, thus causing biased results

Bias in Cognitive Diagnosis

Existing model relies on spurious associations between students' sensitive attributes and outcomes for prediction.



Model	Family Wealth			Country	
	Poor	Average	Wealth	Australia	Brazil
Data statistics	0.4736	0.5448	0.6434	0.5516	0.3888
NCD	0.5140	0.5861	0.6789	0.5913	0.3293
KaNCD	0.4778	0.5589	0.6643	0.5650	0.3025
NCD-*PSCRF*	0.5545	0.5798	0.6155	0.5824	0.3321
KaNCD-*PSCRF*	0.5286	0.5581	0.6271	0.5680	0.3026

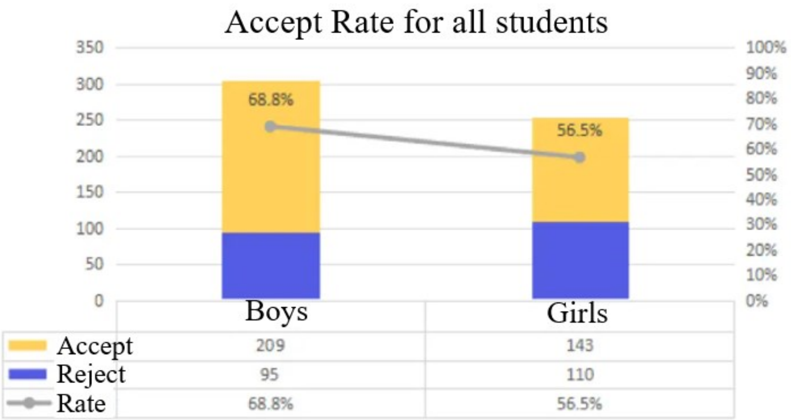
Challenge:

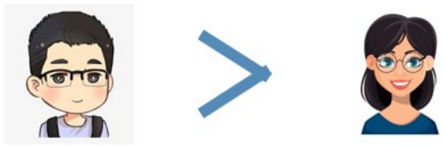
How to **exclude the abuse of student sensitive attributes** while **ensuring diagnostic performance** ?

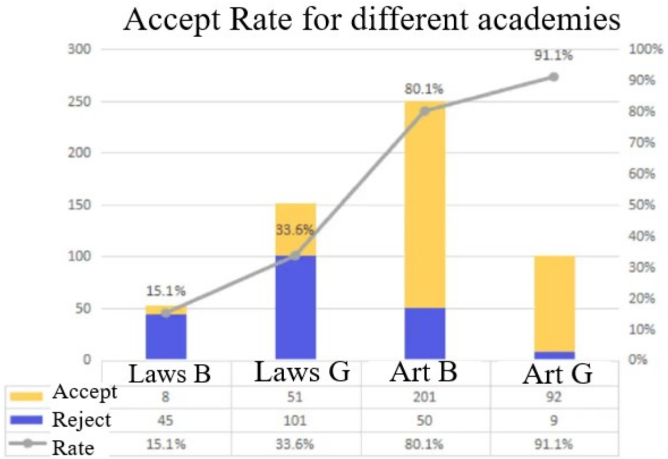
Preliminary——Counterfactual Inference

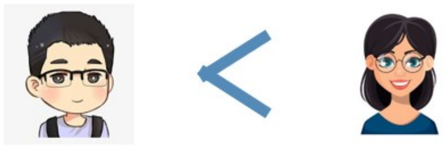
Causal Inference:

- Identify the causal relation from spurious correlation based on the observed data
- One important strategy to improve the robustness and fairness of neural models
- Conditioning vs. **Intervening**
- Modularity assumption (Intervention)

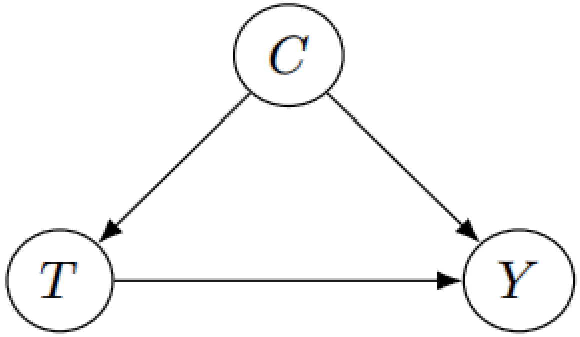








Causal Graph



https://www.bradyneal.com/causal-inference-course

Challenges

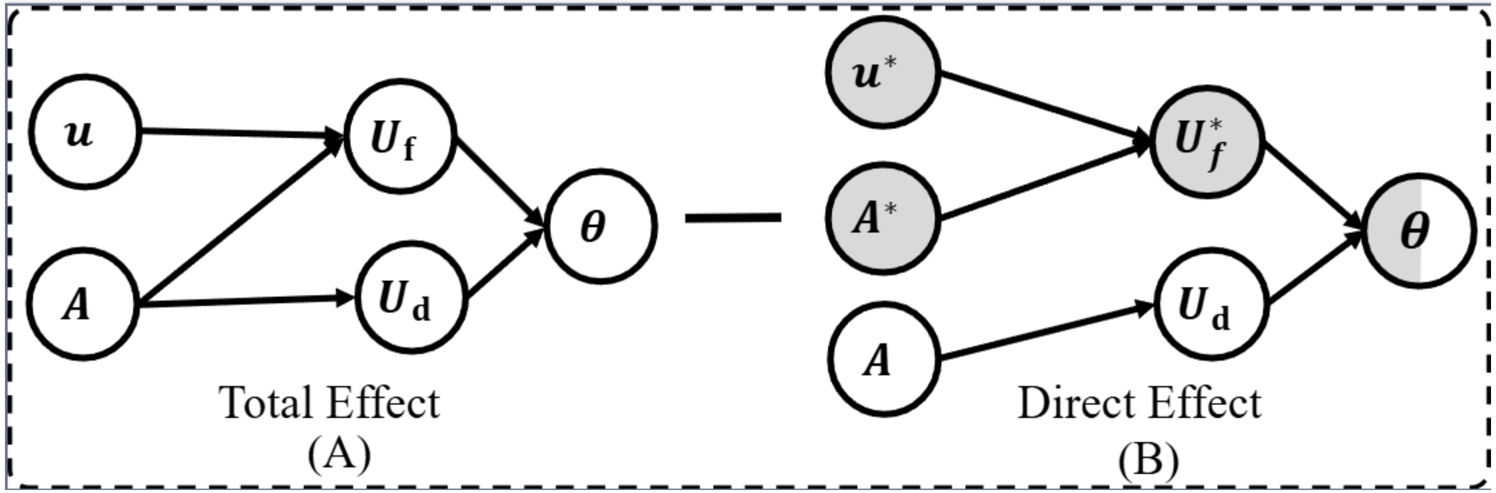
- How to describe the relations of student performance and different factors (variables)?
- How to identify the effect of sensitive attributes on student performance and realize intervention?
- How to evaluate the performance of fairness-aware Cognitive Diagnosis?

Solutions

- Conducting detailed data analysis and using a causal graph to describe the relations
- Classifying student attributes into fairness-related sensitive attributes and diagnosis-related features, and design a novel attribute-oriented predictor to realize decoupling.
- Using commonly used metrics and concentrated more on vulnerable groups

Outline

Motivation



- **Core Idea** is that sensitive attributes contain not only sensitive information that may lead to unfair predictions but also useful information that can enhance diagnostic performance.
- **Goal** is to decouple these pieces of information, so that during the diagnostic process, we can exclude features related to fairness while retaining those that are beneficial for diagnosis.

Table 8: The statistics of useful but not sensitive attributes associated with ESCS

id	name	correlation	category
ST013Q01TA	How many books are there in your home?	0.416	0: 0-100 books, 1: More than 100 books
ST012Q07NA	Tablet computers	0.402	0: Zero or one 1: More than one
ST011Q06TA	A link to the Internet	0.308	0: Yes 1: No
STo11Q04TA	A computer you can use for school work	0.301	0: Yes 1: No
ST012Q08NA	E-book readers	0.266	0: NaN 1: More than one

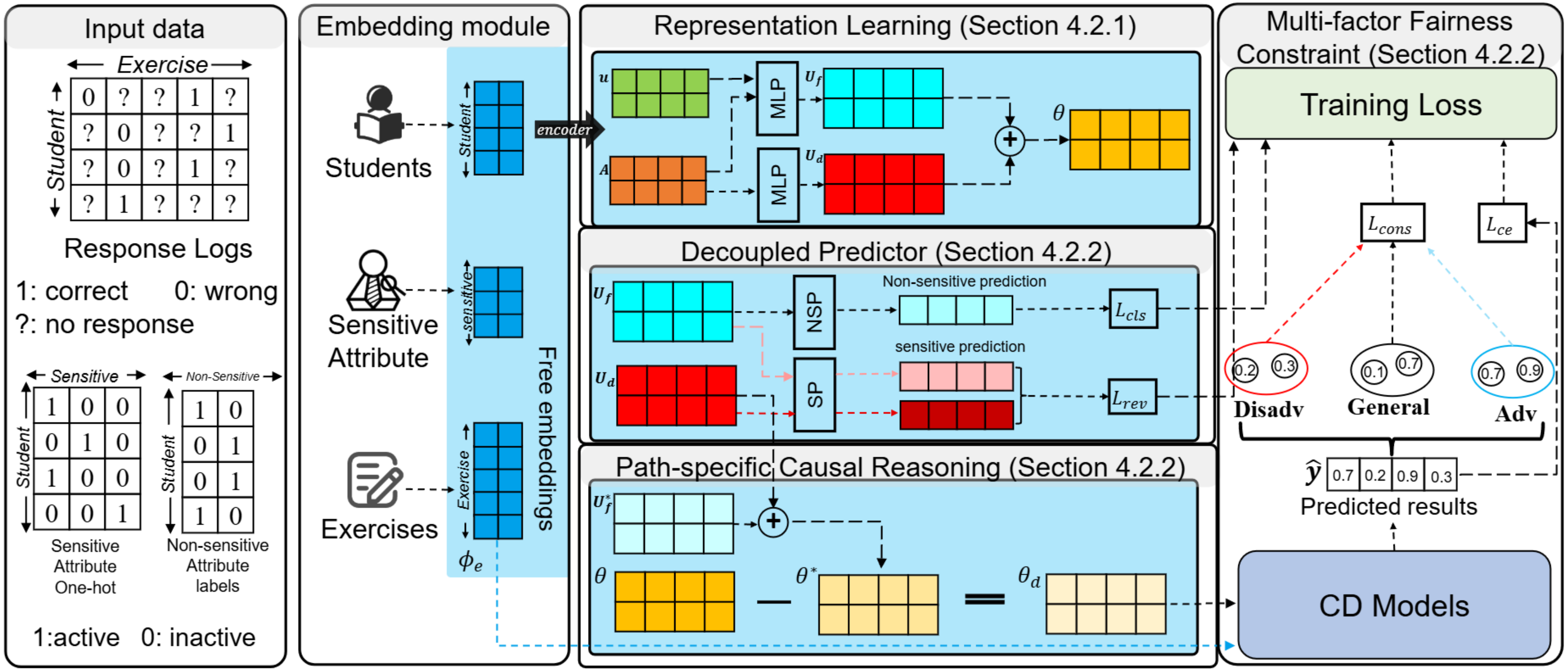
$$TE = \theta(u, A) - \theta(u^*, A^*)$$

$$NDE = \theta(u^*, A) - \theta(u^*, A^*)$$

$$TIE = TE - NDE = \theta(u, A) - \theta(u^*, A)$$

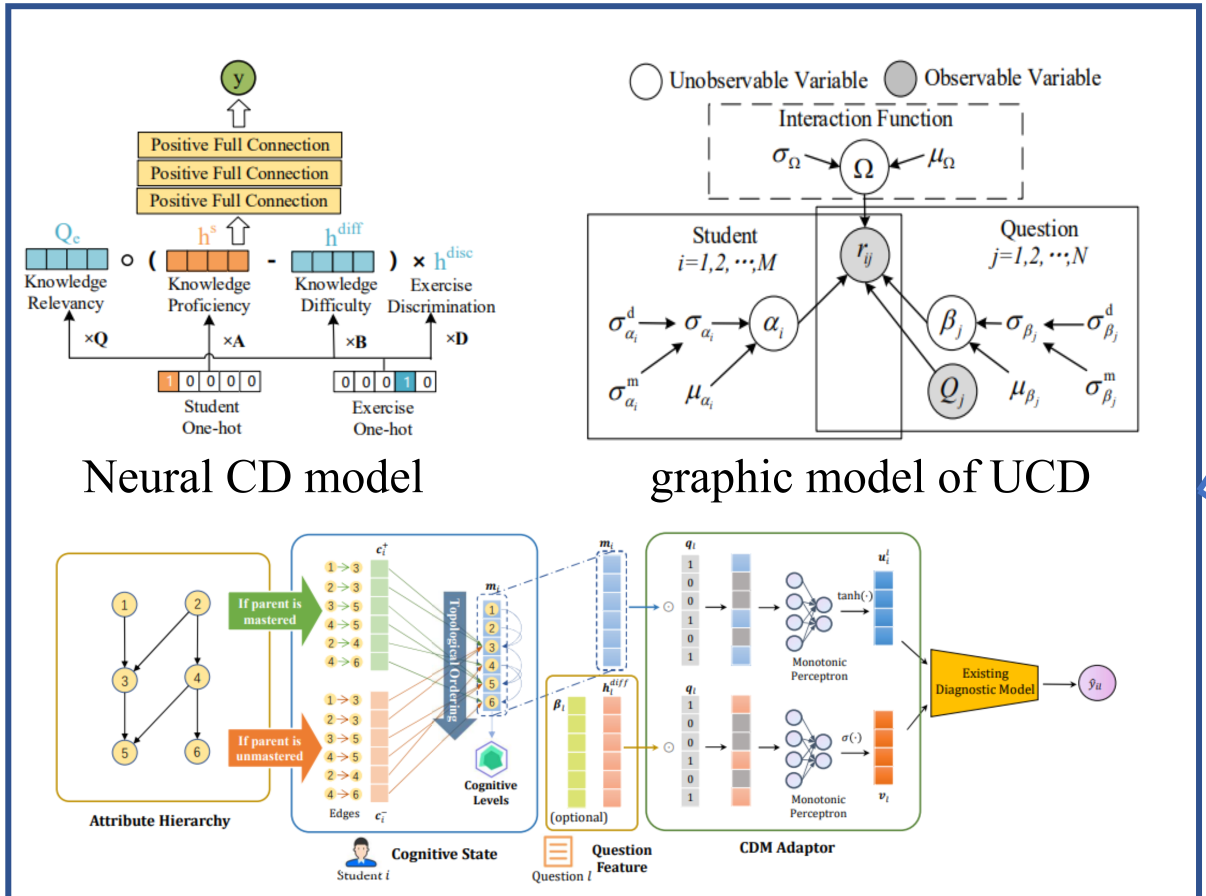
PSCRF Framework

- **Main Contribution**
  - **Decoupled Predictor**
  - **Causal Reasoning Prediction**
  - **Multi-factor Fairness Constraint**



PSCRF Framework

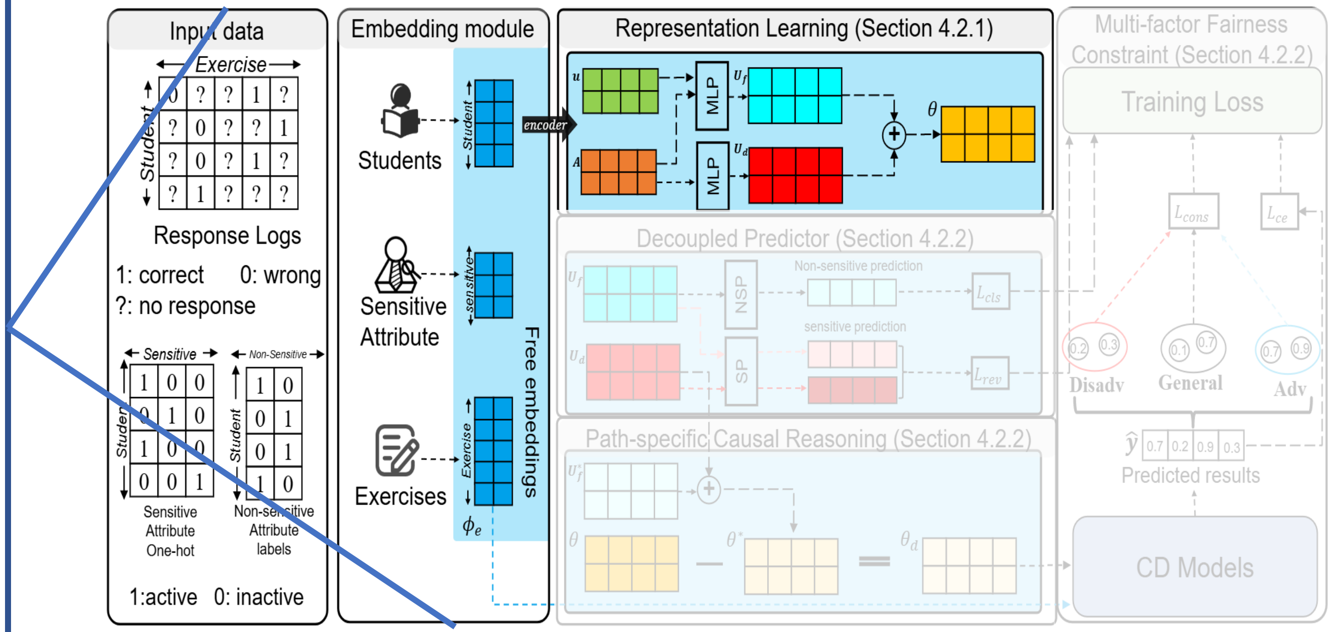
- **Main Contribution**
  - **Model-agnostic representation learning**



Neural CD model

graphic model of UCD

Hierarchical Cognitive Diagnosis framework



PSCRF Framework

- **Main Contribution**
  - **Model-agnostic representation learning**

sensitive attributes contain fairness-related sensitive features and diagnosis-related features.

✓ **fairness-related sensitive feature generator**

$$U_d^i = \sigma(MLP_1(\boxed{A_{[i]}})),$$

Sensitive attribute embeddings

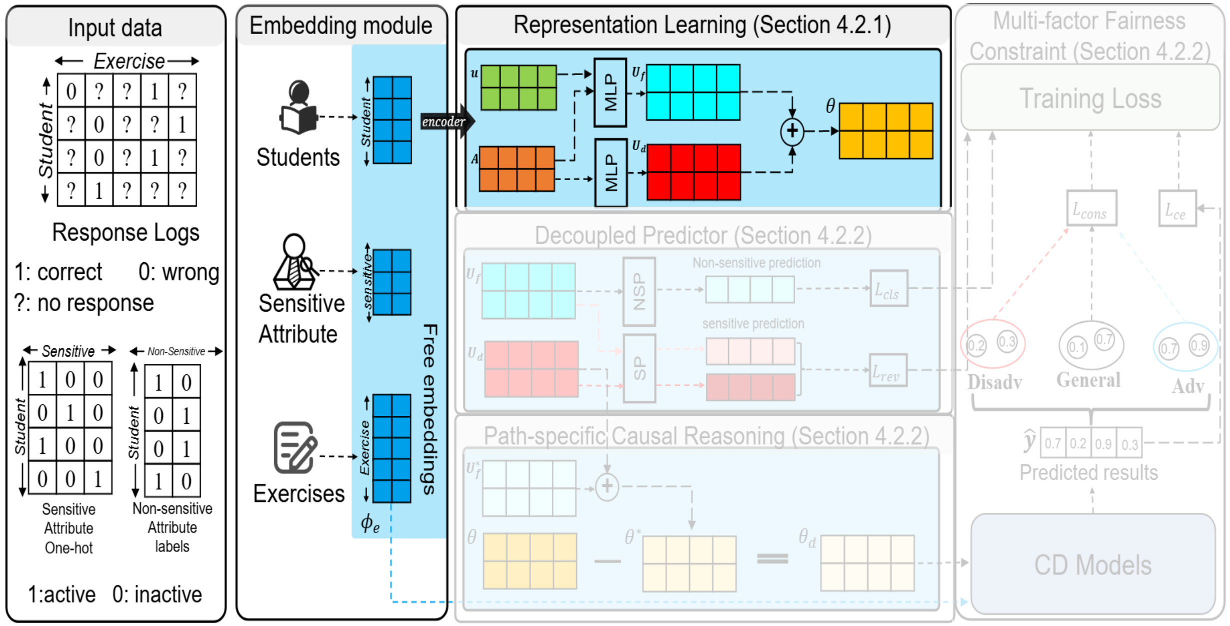
✓ **diagnosis-related feature extractor**

$$U_f^i = \sigma(MLP_2(concat(\boxed{u_i}, A_{[i]}))),$$

Student embedding

✓ **student proficiency level Modeling**

$$\theta_i = \sigma((1-\alpha)U_f^i + \alpha U_d^i).$$

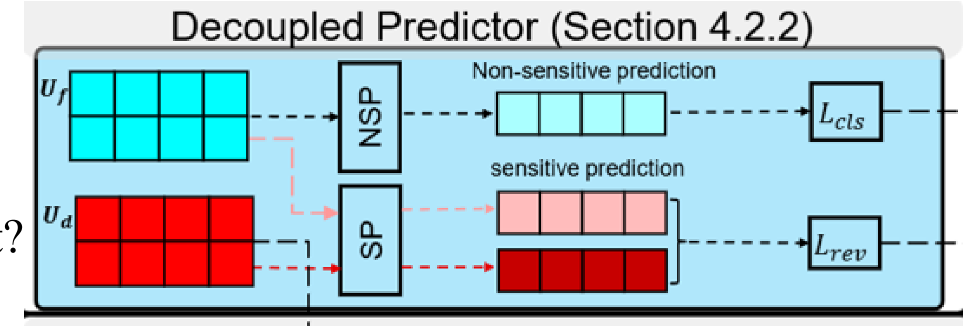


PSCRF Framework

- **Main Contribution**
  - **Decoupled Predictor**

    Sensitive Attributes: e.g., Family Wealth

    Non-sensitive attributes: e.g., whether you have the link to the internet?



- **Non-Sensitive Features Prediction:** We use U_f to predict useful features associated with sensitive attributes in order to preserve the useful information in sensitive attributes.

The k-th non-sensitive attribute

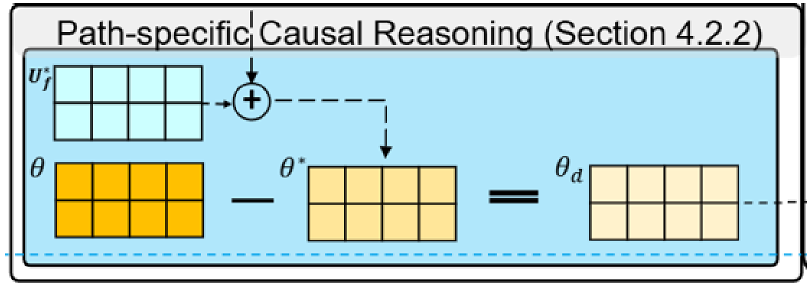
$$L_{cls} = \frac{1}{K}\sum_{k=1}^{K} CE\left(MLP(U_f), \boxed{Label_k}\right)$$

- **Sensitive Attributes Prediction**: We use U_d to directly predict sensitive attributes to improve the accuracy of its modeling of sensitive attributes, and U_f to predict counterfactual sensitive attributes to avoid unfairness.

$$L_{rev} = L(SMLP(U_d), A) + L(SMLP(U_f), \boxed{A^*})$$

PSCRF Framework

- **Main Contribution**
  - **Causal Reasoning Prediction**



- **Causal Reasoning Prediction:** We perform counterfactual debias inference based on the previous causal graph using learnable parameters β control the degree of debiasing.

- Factual Reasoning

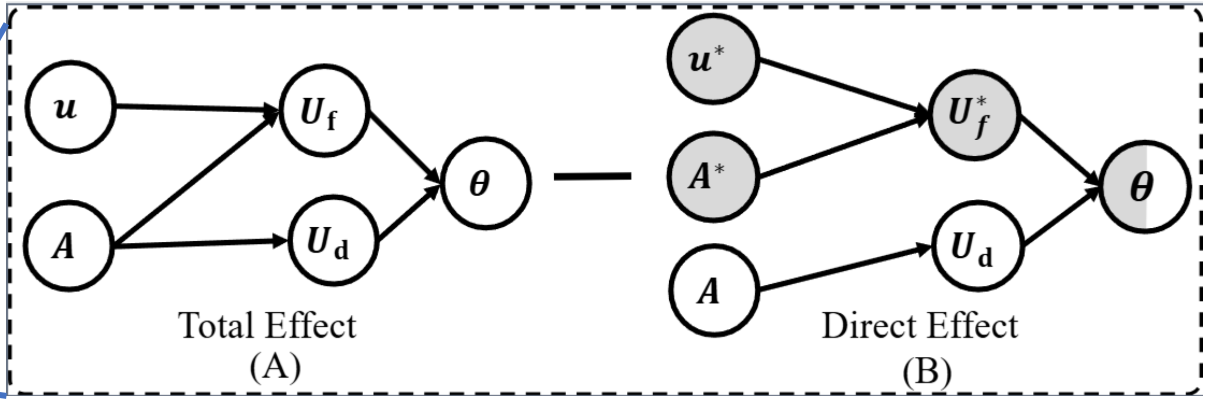
$$U_f^i = \sigma(MLP_2(concat(\boldsymbol{u}_i, \boldsymbol{A}_{[i]}))),$$

$$\boldsymbol{\theta}_i = \sigma((1-\alpha)U_f^i + \alpha U_d^i).$$

- Counterfactual Reasoning

$$U_f^* = \sigma(MLP(concat(\boldsymbol{u}^*, \boldsymbol{A}^*))),$$

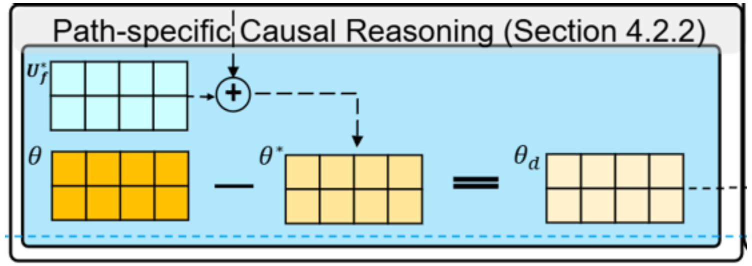
$$\boldsymbol{\theta}^* = \sigma((1-\alpha)U_f^* + \alpha U_d),$$



$$\theta_d = \sigma(\theta - \beta\theta^*)$$

PSCRF Framework

- **Main Contribution**
  - **Multi-factor Fairness Constraint**



- **Causal Reasoning Prediction:** We perform counterfactual debias inference based on the previous causal graph using learnable parameters β control the degree of debiasing.

$$\theta_d = \sigma(\theta - \beta\theta^*)$$

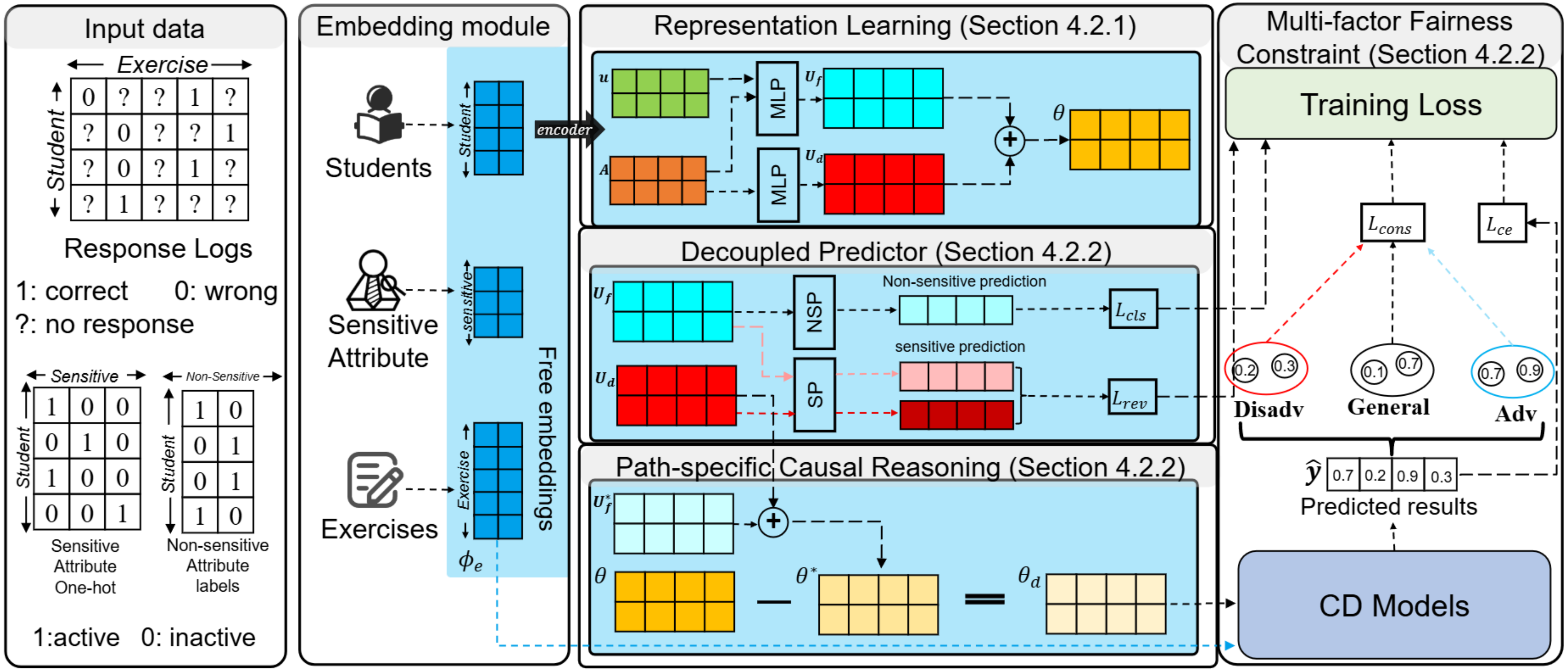
- **Multi-factor Fairness Constraint:** We minimize the variance of the predicted mean between different groups of debiased features to achieve fairness constraints, while maximizing the variance of the predicted mean U_d sensitive representation.

$$L_{cons} = std(\bar{y}_{dis}, \bar{y}_{gene}, \bar{y}_{adv})_{\theta_d} - std(\bar{y}_{dis}, \bar{y}_{gene}, \bar{y}_{adv})_{U_d}$$

$$\mathcal{L}_{\text{total}} = w_1\mathcal{L}_{\text{ce}} + w_2\mathcal{L}_{\text{cls}} + w_3\mathcal{L}_{\text{rev}} + w_4\mathcal{L}_{\text{cons}},$$

PSCRF Framework

- **Main Contribution**
  - **Decoupled Predictor**
  - **Causal Reasoning Prediction**
  - **Multi-factor Fairness Constraint**



Outline

Experiments

◆ **Dataset descriptions, evaluation metrics and the sensitive attribute selection.**

Dataset	Students	Exercises	Exercise Records
Australia	8,485	184	249,727
Brazil	5,777	183	143,314

Evaluation Metrics

Diagnosis Performance Metrics: ACC, AUC, IR, DOA

Fairness Performance Metrics: EO, D_{disadv}^{under}

Identified Rate: $$IR = \frac{5 \times precision_{disadv} \times recall_{disadv}}{(4 \times precision_{disadv}) + recall_{disadv}}$$

Equal opportunity: $$EO = Std(TPR_{disadv}, TPR_{gene}, TPR_{adv},)$$

Disadv group underestimation rate: $$D_{disadv}^{under} = FNR_{disadv} - FNR_{adv}$$

We use two common sensitive attribute:

1.ESCS: Index of Economic, Social, and Cultural Status

2.Father's education level

id	name	correlation	category
ST013Q01TA	How many books are there in your home?	0.416	0: 0-100 books, 1: More than 100 books
ST012Q07NA	Tablet computers	0.402	0: Zero or one 1: More than one
ST011Q06TA	A link to the Internet	0.308	0: Yes 1: No
STo11Q04TA	A computer you can use for school work	0.301	0: Yes 1: No
ST012Q08NA	E-book readers	0.266	0: NaN 1: More than one

The statistics of useful but not sensitive attributes associated with ESCS

Pearson correlation coefficient

Overall Experiments

- PSCRF outperforms baseline methods and other approaches in terms of the fairness metric.
- PSCRF performs excellently on diagnostic performance metrics.

Evaluating accuracy and fairness performance associated with sensitive attribute ESCS

Model		Australia						Brazil					
		EO↓	D^{under}_{disadv}	IR↑	AUC↑	ACC↑	DOA↑	EO↓	D^{under}_{disadv}	IR↑	AUC↑	ACC↑	DOA↑
IRT	Base	0.0338	0.0826	0.7353	0.7979	0.7266	-	0.0582	0.1407	0.5018	0.7794	0.7269	-
	Base†	0.0604	0.1473	0.7025	**0.8080**	**0.7322**	-	0.1025	0.2510	0.4700	**0.7958**	**0.7324**	-
	Reg	0.0110	0.0270	**0.7544**	0.7961	0.7249	-	0.0277	0.0665	0.5301	0.7769	0.7250	-
	Adv	0.0286	0.0697	0.7449	0.7969	0.7264	-	0.0669	0.1609	0.4935	0.7797	0.7268	-
	PSCRF	**0.0051**	**0.0002**	0.7339	0.8022	0.7249	-	**0.0162**	**0.0357**	**0.5760**	0.7893	0.7255	-
MIRT	Base	0.0575	0.1408	0.7013	0.8027	0.7299	-	0.0913	0.2227	0.5109	0.7836	0.7280	-
	Base†	0.0645	0.1523	0.6973	**0.8088**	**0.7339**	-	0.1251	0.3053	0.4663	**0.7950**	**0.7316**	-
	Reg	0.0284	0.0694	0.7279	0.8010	0.7278	-	0.0512	0.1246	**0.5539**	0.7813	0.7258	-
	Adv	0.0554	0.1357	0.7009	0.8030	0.7288	-	0.0956	0.2335	0.5036	0.7840	0.7283	-
	PSCRF	**0.0098**	**0.0227**	**0.7520**	0.7983	0.7237	-	**0.0279**	**0.0403**	0.5248	0.7804	0.7205	-
NCD	Base	0.0425	0.1040	0.7183	0.7868	0.7170	0.6248	0.0669	0.1588	0.5220	0.7675	0.7140	0.5972
	Base†	0.0857	0.2039	0.6615	0.7911	0.7199	0.6384	0.1274	0.3108	0.4491	0.7718	0.7166	0.6394
	Reg	0.0331	0.0811	0.7277	0.7863	0.7172	0.6245	0.0522	0.1229	0.5370	0.7669	0.7131	0.5965
	Adv	0.0528	0.1292	0.6644	0.7801	0.7111	0.5715	0.0506	0.1234	0.5388	0.7601	0.7112	0.5648
	PSCRF	**0.0029**	**0.0010**	**0.7538**	**0.7997**	**0.7234**	**0.7040**	**0.0030**	**0.0028**	**0.5599**	**0.7788**	**0.7209**	**0.6806**
KaNCD	Base	0.0464	0.1133	0.7113	0.8017	0.7273	0.6584	0.0742	0.1792	0.4877	0.7793	0.7221	0.6046
	Base†	0.0770	0.1878	0.6957	**0.8076**	**0.7310**	0.6917	0.1210	0.2963	0.5103	**0.7910**	**0.7284**	**0.6848**
	Reg	0.0255	0.0622	0.7299	0.8004	0.7260	0.6552	0.0464	0.1115	0.5138	0.7775	0.7207	0.6015
	Adv	0.0532	0.1303	0.7075	0.8009	0.7282	0.6615	0.0686	0.1664	0.5388	0.7802	0.7244	0.6357
	PSCRF	**0.0110**	**0.0252**	**0.7484**	0.8045	0.7299	**0.7013**	**0.0363**	**0.0888**	**0.5145**	0.7892	0.7267	0.6840

Father's Education Level Results

◆ **We obtain similar results as before, demonstrating the generalizability of our method.**

Evaluating accuracy and fairness performance associated with sensitive attribute Father's education level

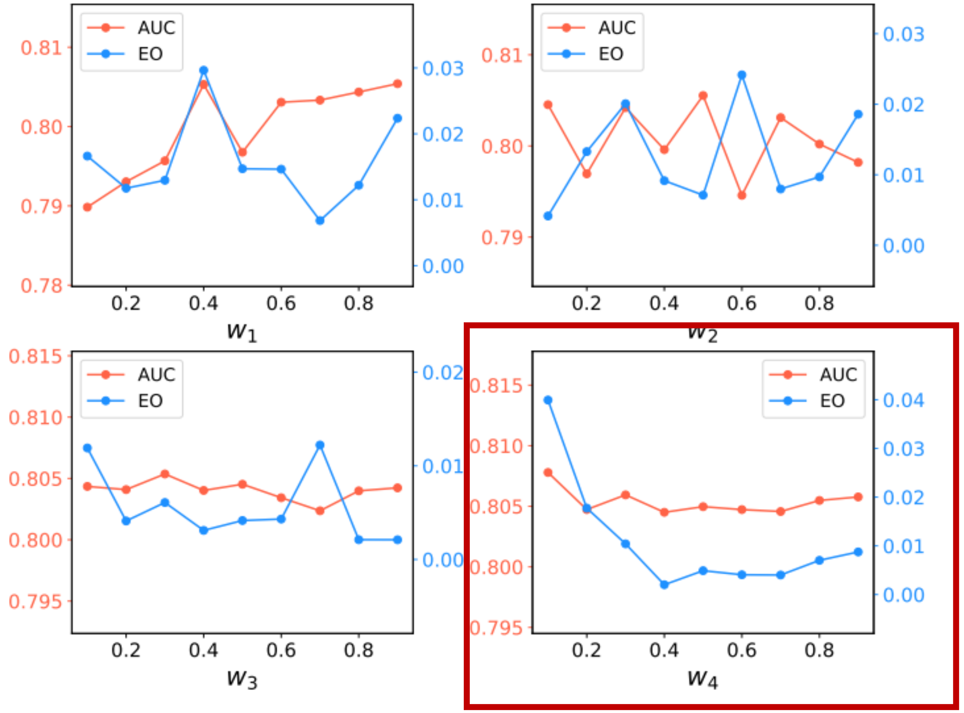
Model		Australia						Brazil					
		EO↓	D^{under}_{disadv}	IR↑	AUC↑	ACC↑	DOA↑	EO↓	D^{under}_{disadv}	IR↑	AUC↑	ACC↑	DOA↑
IRT	Base	0.0293	0.0705	0.7346	0.7979	0.7266	-	0.0366	0.0896	0.5209	0.7794	0.7269	-
	Base†	0.0440	0.1069	0.6980	**0.8087**	**0.7312**	-	0.0720	0.1723	0.5379	**0.7959**	**0.7314**	-
	Reg	**0.0132**	0.0303	**0.7515**	0.7969	0.7255	-	0.0190	0.0465	0.5468	0.7781	0.7262	-
	Adv	0.0231	0.0546	0.7319	0.7919	0.7200	-	0.0314	0.0716	0.5404	0.7752	0.7242	-
	PSCRF	0.0162	**0.0015**	0.7342	0.8034	0.7277	-	**0.0021**	**-0.0049**	**0.5640**	0.7911	0.7274	-
MIRT	Base	0.0437	0.1051	0.7084	0.8027	0.7299	-	0.0572	0.1398	0.5472	0.7836	0.7280	-
	Base†	0.0554	0.1326	0.7195	**0.8106**	**0.7347**	-	0.0849	0.1820	0.4931	**0.7896**	0.7279	-
	Reg	**0.0194**	**0.0449**	0.7383	0.8025	0.7297	-	0.0300	0.0735	0.5787	0.7812	0.7272	-
	Adv	0.0389	0.0947	0.7110	0.8043	0.7316	-	0.0528	0.1292	0.5548	0.7821	**0.7282**	-
	PSCRF	**0.0194**	0.0474	0.7095	0.8073	0.7285	-	**0.0247**	**0.0422**	**0.5879**	0.7827	0.7214	-
NCD	Base	0.0313	0.0747	0.7265	0.7868	0.7170	0.6248	0.0428	0.1042	0.5409	0.7675	0.7140	0.5972
	Base†	0.0477	0.1147	0.6981	**0.8021**	0.7263	0.6478	0.0855	0.1745	**0.6095**	0.7785	0.7026	0.6518
	Reg	0.0293	0.0679	0.6940	0.7834	0.7119	0.6167	0.0324	0.0794	0.5396	0.7689	0.7145	0.6010
	Adv	0.0323	0.0789	0.6791	0.7825	0.7130	0.5918	0.0484	0.1177	0.5470	0.7635	0.7150	0.5722
	PSCRF	**0.0227**	**-0.0116**	**0.7362**	0.8003	**0.7276**	**0.7096**	**0.0280**	**0.0465**	0.5745	**0.7879**	**0.7293**	**0.6889**
KaNCD	Base	0.0370	0.0887	0.7146	0.8017	0.7273	0.6584	0.0433	0.1058	0.5189	0.7793	0.7221	0.6046
	Base†	0.051	0.1207	0.6938	**0.8084**	**0.7310**	**0.7183**	0.0555	0.1302	0.5434	0.7862	0.7256	0.6731
	Reg	0.0251	0.0578	0.7364	0.8010	0.7274	0.6642	**0.0288**	**0.0699**	**0.5826**	0.7799	0.7242	0.6347
	Adv	0.0405	0.0972	0.7144	0.8006	0.7278	0.6618	0.0419	0.1020	0.5609	0.7802	0.7239	0.6352
	PSCRF	**0.0114**	**0.0275**	**0.7768**	0.8066	0.7269	0.7097	0.0340	0.0746	0.5130	**0.7930**	**0.7278**	**0.6847**

Ablation Studies and Parameter Sensitivity Analysis

- When the Multi-Factor Fairness **Constraint was removed**, the model's fairness performance **significantly declined**
- Removing the **Decouple Module** caused the model to fail in decoupling sensitive information, **impacting fairness**.
- Increasing the **weight** of the **fairness constraint** improved the model's fairness, but slightly reduced diagnostic performance.

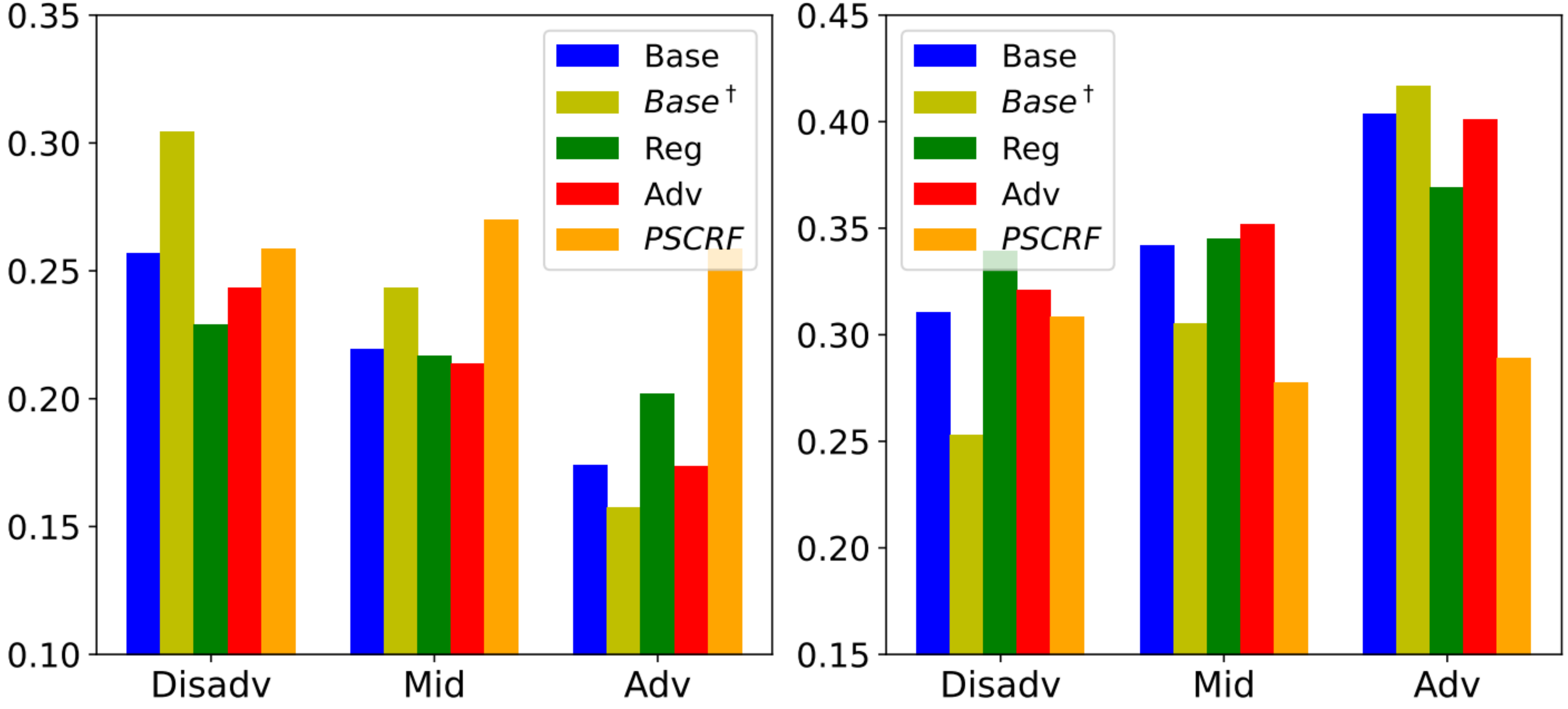
Conditions	EO↓	D^{under}_{disadv}	IR↑	AUC↑	ACC↑
Base	0.0338	0.0826	0.7353	0.7979	0.7266
PSCRF	**0.0051**	**0.0002**	0.7339	0.8022	0.7249
w $\mathcal{L}_{ce}$	0.0545	0.1335	0.6885	0.8089	0.7329
w $\mathcal{L}_{cls}$	0.0503	0.1231	0.6982	0.8088	0.7328
w $\mathcal{L}_{rev}$	0.0525	0.1287	0.6919	0.8090	0.7332
w $\mathcal{L}_{cons}$	0.0088	0.0069	0.7392	0.8016	0.7250
w/o $\mathcal{L}_{cls}$	0.0137	0.0318	0.7206	0.8057	0.7279
w/o $\mathcal{L}_{rev}$	0.0112	-0.0057	0.7277	0.8022	0.7258
w/o $\mathcal{L}_{cons}$	0.0609	0.1493	0.7127	**0.8092**	**0.7337**
w/o $\mathcal{L}^{*}_{cons}$	0.0132	-0.0322	**0.7565**	0.8021	0.7257

PSCRF* based on IRT with varying *w



Case Study

- The baseline model exhibits clear biases. Introducing sensitive attributes further exacerbates this phenomenon.
- Resampling and Adversarial methods can reduce underestimation rates, but they achieve this by sacrificing overestimation rates. And they do not effectively balance the gaps between different groups.
- PSCRF significantly reduces the overestimation rate for advantaged students and decreases the gap in prediction distributions between different groups.



(a)FNR(underestimation rate)
(b)FPR(overestimation rate)

Outline

Conclusions and Future Work

Contributions

- Introduce a novel PSCRF for fairness-aware cognitive diagnosis.
- Design an attribute-oriented predictor to decouple sensitive attributes into fairness-related and diagnosis-related features.
- Extensive experiments on real-world datasets demonstrate the effectiveness of PSCRF.

Future Directions

- Explore the application of PSCRF with various types of sensitive attributes in diverse educational scenarios.
- Taking into account multiple sensitive attributes as well as considering invisible sensitive attributes.

Thanks!

Check out paper and opensource project at
https://github.com/NLPfreshman0/PSCRF
https://zhangkunzk.github.io/