Неявная оценка качества ранговых классификаторов с помощью когнитивной диагностики

Брескану Никита

ММП ВМК МГУ

2024

Цель исследования

предложить новый метод оценки качества ранговых классификаторов, основанный на использовании когнитивной диагностики.

Требуется предложить

метод получения новых показателей качества (уровней знаний) моделей по заранее выбранным интерпретируемым атрибутам, используя метрики на разных датасетах.

Практическая ценность

конвергенция метрик качества и датасетов; получение метрик, показывающих качество модели в сравнении с другими моделями. Применение для создания таблиц лидеров.

Когнитивная диагностика

Это раздел теории тестирования, ширико использующийся в умном образовании.

Дано

1. N студентов, M упражнений, K понятий
2. Логи решения упражнений $R = \{(x_t^s, x_t^e, r_t)\}_{t=1}^T$
3. Матрица связи упражнений и атрибутов $Q = \{Q_{jk}\}_{M \times K}$
4. Граф связи атрибутов (опционально) $G \in \{0,1\}^{K \times K}$

Найти

Скрытые уровни знаний студентов $\mathcal{M} = \{m_{ik}\}_{N \times K} \in [0,1]^{N \times K}$

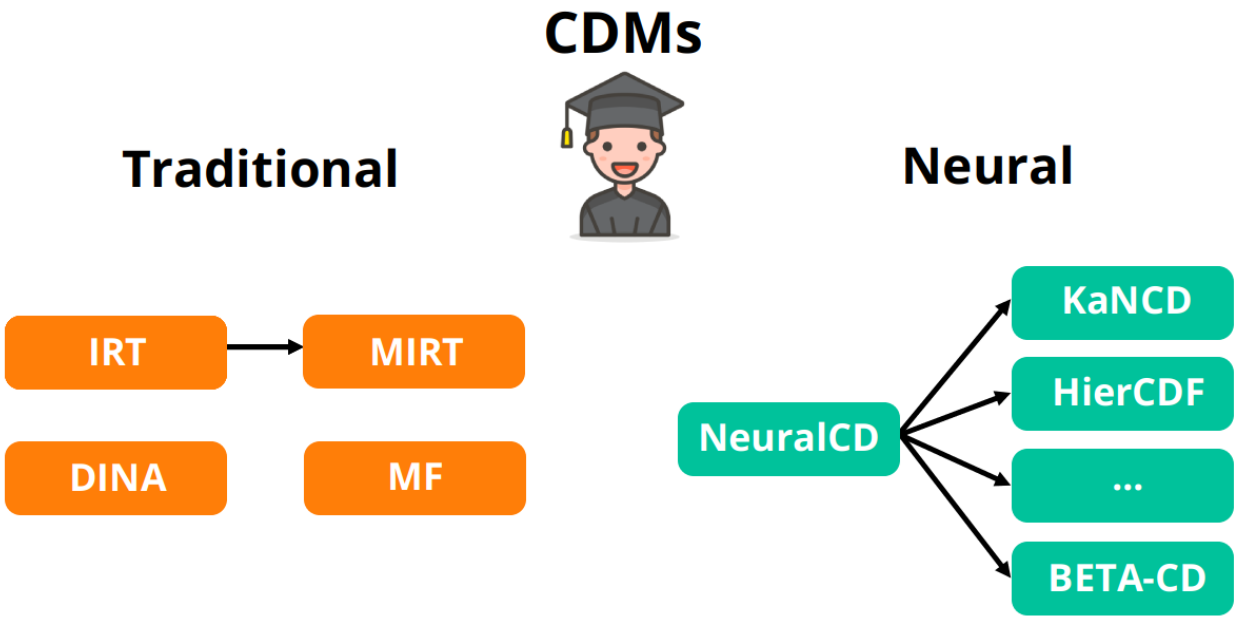
Пример когнитивной модели: MIRT

Интеракционная функция имеет вид:

$$\mathbb{P}(y = 1|m, h^{disc}, h^{diff}) = \sigma(\langle q^e \circ h^{\text{disc}}, m - h^{\text{diff}}\mathbb{I}\rangle)$$

y — ответ студента на управжнение.
q^e — строка Q-матрицы.
$h^{disc} \in [0, 1]^K$, $h^{diff} \in [0, 1]$ — обучаемые параметры для каждого упражнения.
$m \in [0, 1]^K$ — обучаемые уровни знаний для студентов.
$\mathbb{I}$ — вектор из всех единиц.
Критерий оптимизации: бинарная кросс-энтропия между ответами и предсказаниями.

Модели когнитивной диагностики



- Традиционные когнитивные модели: фиксированная интеракционная функция.
- Глубокие когнитивные модели: интеракционная функция является обучаемой.

Сведение оценки моделей к когнитивной диагностике

Сведение

- Модели $\Longrightarrow$ студенты
- Датасеты, метрики качества $\Longrightarrow$ упражнения
- Метрики качества для моделей $\Longrightarrow$ логи решения упражнений R.

Нужно заранее разметить

Атрибуты, граф G, матрицу Q

Постановка задачи

Дано

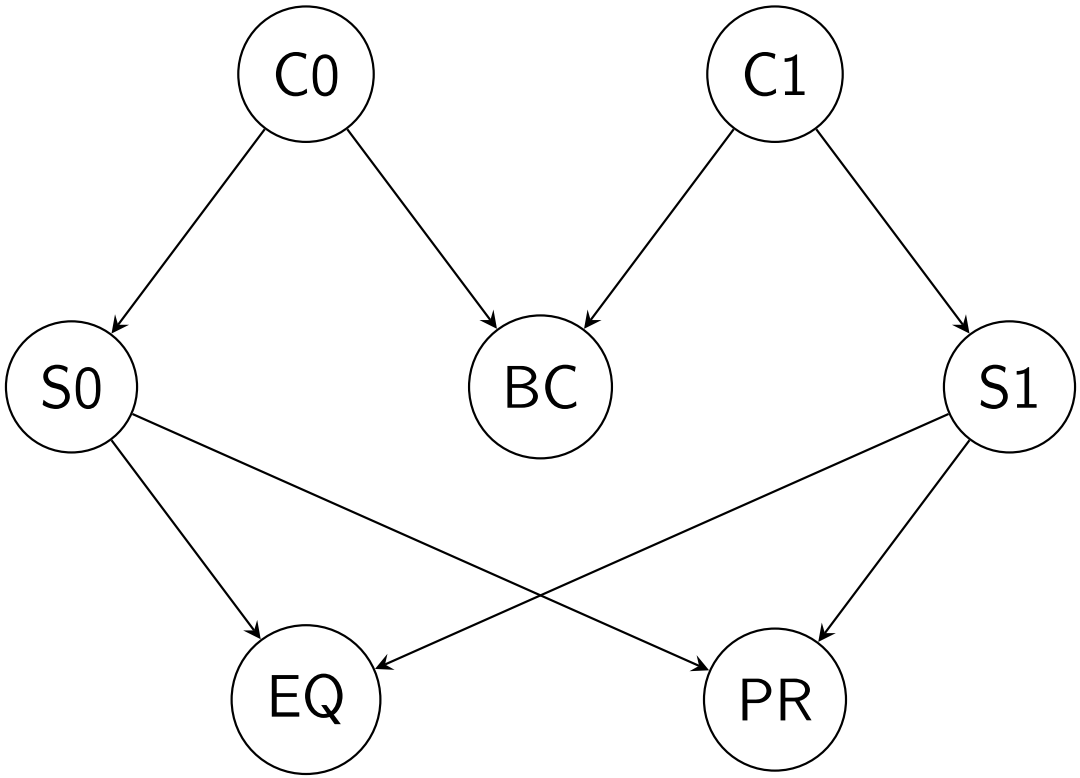
1. N моделей, D датасетов, K атрибутов, E метрик
2. Матрица связи метрик и атрибутов $Q = \{Q_{jk}\}_{M \times K}$ (размечена)
3. Граф связи атрибутов $G \in \{0, 1\}^{K \times K}$ (размечен)

Найти

Аггрегированные по датасетам метрики качества $\mathcal{M} = \{m_{ik}\}_{N \times K} \in [0, 1]^{N \times K}$

В такой постановке легко сводится к когнитивной диагностике

Выбор атрибутов



Атрибуты и их граф зависимости. Когнитивная модель будет работать из ходя из предположения, что родительские вершины необходимы для владения дочерними.

Выбор моделей

Classifier	Implementation	Varying parameters	Number of models
Logistic regression	sklearn	C, solver	120
Decision tree	sklearn	max_depth, criterion	60
Random forest	sklearn	max_depth, n_estimators	12
Gradient boosting	sklearn	n_estimators, learning_rate	9
Gradient boosting	LGBM	n_estimators, num_leaves	9
SVM	sklearn	C, kernel	30
K nearest neighbors	sklearn	n_neighbors, weights	40
Multilayer perceptron	sklearn	hidden_layer_sizes, activation	15
Optimal classifier	<manual>	<absent>	1
Pessimal classifier	<manual>	<absent>	1
Majority classifier	<manual>	<absent>	1
Minority classifier	<manual>	<absent>	1
Mean target classifier	<manual>	<absent>	1
Uniform Random classifier	<manual>	<absent>	1
Balanced Random classifier	<manual>	<absent>	1

Искусственные классификаторы добавлены для увеличения разнообразия.

Выбор датасетов

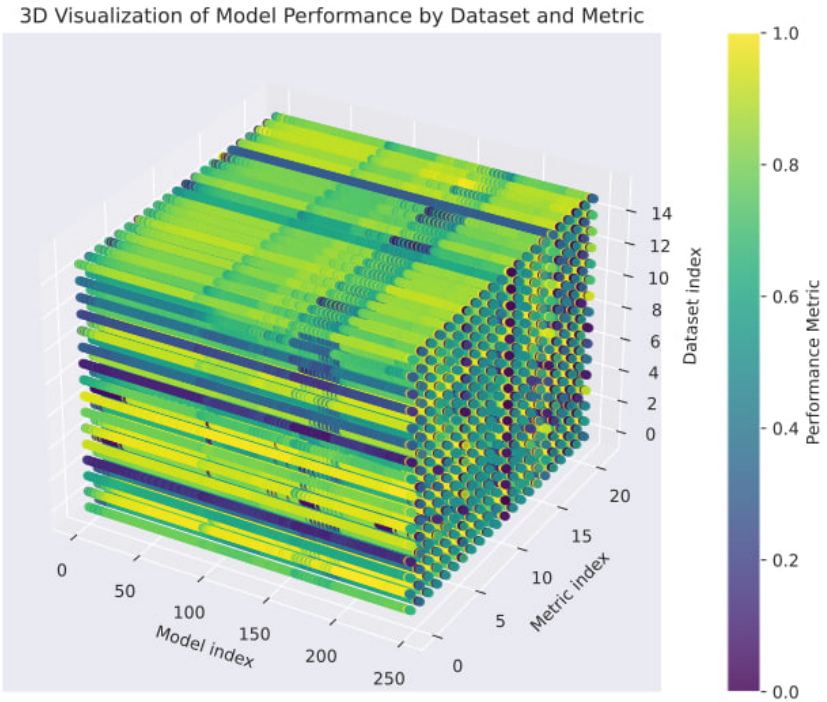
Dataset name	Samples × features	Numerical × categorical features	Class balance
Banknote-authentication	1372 × 5	5 × 0	55–45%
Blood-transfusion-service-center	748 × 5	5 × 0	76–24%
Breast-w	683 × 10	10 × 0	65–35%
Climate-model-simulation-crashes	540 × 21	21 × 0	99–1%
Cylinder-bands	277 × 40	25 × 15	64–36%
Dresses-sales	99 × 13	2 × 11	59–41%
Diabetes	768 × 9	9 × 0	65–35%
ilpd	583 × 11	10 × 1	71–29%
kc1	2109 × 22	22 × 0	84–16%
kc2	522 × 22	22 × 0	79–21%
pc1	1109 × 22	22 × 0	93–7%
pc3	1563 × 38	38 × 0	89–11%
Phoneme	5404 × 6	6 × 0	70–30%
qsar-biodeg	1055 × 42	42 × 0	66–34%
wdbc	569 × 31	31 × 0	62–38%
wilt	4839 × 6	6 × 0	94–6%

Датасеты имеют разные балансы классов, и соотношение объектов признаков.

Метрики и Q-матрица

Exercise (performance metric)	C0	C1	BC	S0	S1	EQ	PR
ROC-AUC	0	0	1	0	0	1	0
PR-AUC for class 0	0	0	0	1	0	0	0
PR-AUC for class 1	0	0	0	0	1	0	0
Gain chart AUC for class 0	0	0	0	1	0	0	0
Gain chart AUC for class 1	0	0	0	0	1	0	0
KS statistic	0	0	0	0	0	1	0
Kendall's tau	0	0	1	0	0	1	0
Accuracy (EER)	0	0	0	0	0	0	1
Precision for class 0 (EER)	0	0	0	1	0	0	0
Recall for class 0 (EER)	1	0	0	1	0	0	0
Precision for class 1 (EER)	0	0	0	0	1	0	0
Recall for class 1 (EER)	0	1	0	0	1	0	0
Balanced accuracy (EER)	0	0	1	0	0	1	0
F1-score for class 0 (EER)	0	0	0	1	0	0	0
F1-score for class 1 (EER)	0	0	0	0	1	0	0
Average F1-score (EER)	0	0	0	0	0	1	0
FM-score for class 0 (EER)	0	0	0	1	0	0	0
FM-score for class 1 (EER)	0	0	0	0	1	0	0
Markedness (EER)	0	0	0	0	0	1	0
Matthews coefficient (EER)	0	0	0	0	0	1	0
Jaccard index (EER)	0	0	0	0	0	0	1
Cohen's kappa (EER)	0	0	0	0	0	1	0

Получение датасета для когнитивной диагностики



- 14 датасетов с разным балансов классов, вплоть до 99:1
- 22 метрики
- 248 моделей

Проверка качества выделенных знаний по атрибутам

$$R^2 = 1 - \frac{\text{MSE(CDM)}}{\text{MSE}(\bar{x})}$$

$$DOA_k = \frac{\sum_{a,b \in S} [m_{ak} > m_{bk}] \frac{\sum_{j=1}^{M} Q_{jk}[x_{aj} > x_{bj}]}{\sum_{j=1}^{M} Q_{jk}[x_{aj} \neq x_{bj}]}}{\sum_{a,b \in S} [m_{ak} > m_{bk}]}$$

$$DOA = \frac{1}{K} \sum_{k=1}^{K} DOA_k$$

$$DOC_j = \frac{\sum_{a,b \in S} [x_{aj} > x_{bj}] \frac{\sum_{k=1}^{K} Q_{jk}[m_{ak} > m_{bk}]}{\sum_{k=1}^{K} Q_{jk}[m_{ak} \neq m_{bk}]}}{\sum_{a,b \in S} [x_{aj} > x_{bj}]}$$

$$DOC = \frac{1}{M} \sum_{j=1}^{M} DOC_j$$

Критерии оценивания:

- R2-score на предсказании метрик
- Mastery1 > Mastery2 $\implies$ 1 имеет лучше метрики чем 2
- 1 имеет лучше метрики чем 2 $\implies$ Mastery1 > Mastery2

Результаты работы когнитивных моделей

Model	# parameters	R2	DOA	DOC
Random mastery	0	-	0.499 ± 0.009	0.498 ± 0.016
MIRT	4552	-0.011 ± 0.000	0.619 ± 0.001	0.619 ± 0.001
NeuralCD	7177	0.887 ± 0.003	0.586 ± 0.009	0.586 ± 0.009
KaNCD	22467	0.885 ± 0.001	0.584 ± 0.006	0.584 ± 0.006
HierMIRT	9544	0.848 ± 0.028	0.577 ± 0.007	0.577 ± 0.007
HierNCD	11657	0.892 ± 0.001	0.600 ± 0.006	0.600 ± 0.006
QCCDM (small)	9882	0.940 ± 0.038	0.539 ± 0.005	0.539 ± 0.005
QCCDM	144282	0.955 ± 0.054	0.542 ± 0.018	0.542 ± 0.018

Приведены доверительные интервалы по 3 запускам.
QCCDM имеет лучший R2, но худшие DOA и DOC, и слишком много параметров.
MIRT имеет лучшие DOA и DOC, но очень плохой R2
У всех моделей маленький DOA и DOC, близкий к случайному.

Итоги

- В работе предложена схема сведения оценки классификаторов по разным датасетам к когнитивной диагностике.
- Проведён эксперимент с таким сведением, и оказалось, что значения мало интерпретируемы.
- Вероятная причина низкого качества — выбранные атрибуты плохо описывают ранговые классификаторы.
- Общая причина: слишком много упражнений, и слишком мало атрибутов.

Направление дальнейшего исследования:

- Разметить более подходяшие атрибуты, их должно быть больше. Возможно, атрибуты должны быть как-то связаны с балансом классов.