

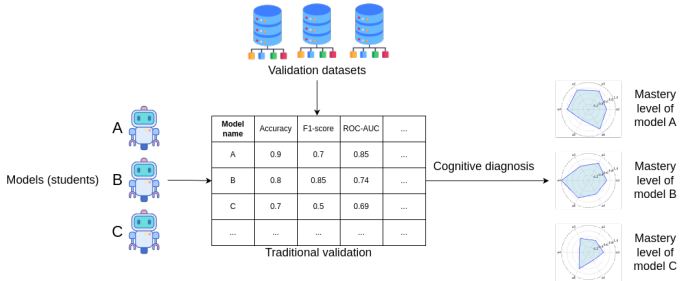
Применение теории тестирования для валидации моделей машинного обучения

Брескану Никита

ММП ВМК МГУ

2024

Введение

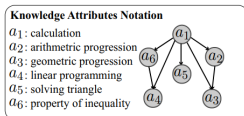


- Когнитивная диагностика используется в умном образовании
- До сих пор есть разногласия с метриками score-based (ранговых) классификаторов.
- Попытка свести задачу оценивания классификаторов к когнитивной диагностике.

Постановка задачи

Пусть имеется L датасетов $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_L\}$, N моделей (алгоритмов ML) $\mathcal{S} = \{s_1, \dots, s_N\}$, \hat{M} традиционных метрик $\hat{\mathcal{E}} = \{\hat{e}_1, \dots, \hat{e}_{\hat{M}}\}$, и K заранее выбранных атрибутов $\mathcal{K} = \{\mathcal{K}_1, \dots, \mathcal{K}_K\}$. Матрица $Q \in \mathbb{R}^{\hat{M} \times K}$ показывает соответствие между метриками и атрибутами: $Q_{jk} = 1 \iff$ атрибут \mathcal{K}_k связан с \hat{e}_j . Вычисление метрик $\hat{\mathcal{E}}$ для каждого датасета формирует множество ответов $R = \{(x^s, x^e, r)\}_{t=1}^T$, где $T = N \times M$, $M = \hat{M} \times L$ — тройки из номера модели, номера метрики (с учётом датасета), и самой метрики $r \in [0, 1]$, нормализованной до $[0, 1]$. Нужно найти $\mathcal{M} = \{m_{ik}\} \in [0, 1]^{N \times K}$, где m_{ik} — уровень знания модели s_i атрибута \mathcal{K}_k ; $m_{ik} = 1$ — идеальное владение атрибутом, и $m_{ik} = 0$ — полное невладение. Задача заключается в получении \mathcal{M} с помощью когнитивной диагностики, предсказывая r_t .

Когнитивная диагностика



Exercise	Knowledge Attributes					
	a_1	a_2	a_3	a_4	a_5	a_6
e_1	✓	✓	✗	✗	✗	✓
e_2	✗	✗	✓	✓	✗	✗
e_3	✗	✓	✓	✗	✓	✗
e_4	✓	✗	✓	✓	✗	✓

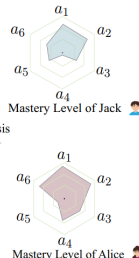
Complete Exercises



Response Logs

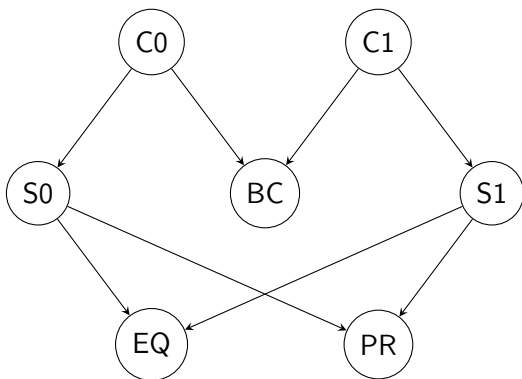
Exercise	Student Response	
	Jack	Alice
e_1	✓	✓
e_2	✗	✓
e_3	✓	✗
e_4	✗	✓

Cognitive Diagnosis



Задача когнитивной диагностики — найти скрытые уровни знаний студентов по логам решения ими упражнений.
 Есть много моделей, позволяющих решать эту задачу.

Выбор атрибутов



Атрибуты и их граф зависимости. Когнитивная модель будет работать исходя 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
Pessimist 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

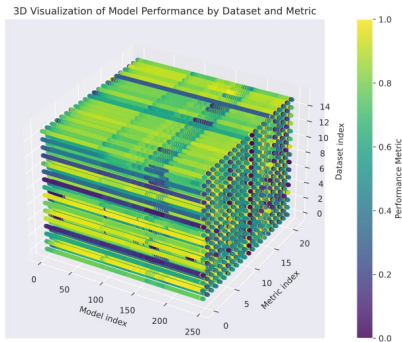
Dataset name	Samples \times features	Numerical \times categorical features	Class balance
Banknote-authentication	1372 \times 5	5 \times 0	55–45%
Blood-transfusion-service-center	748 \times 5	5 \times 0	76–24%
Breast-w	683 \times 10	10 \times 0	65–35%
Climate-model-simulation-crashes	540 \times 21	21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

Критерии оценивания:

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

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

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

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

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

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

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.

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