Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Автоматическая генерация признаков на табличных данных

Кузнецов Михаил Константинович

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
д.ф.-м.н., профессор
А. Г. Дьяконов

Постановка задачи

Введем обозначения:

- $D = (x_i, y_i)_{i=1}^n$ — выборка данных,
- $(x_i, y_i) = ((x_{i1}, x_{i2}, ..., x_{ip}), y_i)$,
- $X_j = (x_{ij})_{i=1}^n$, $X = (x_i)_{i=1}^n$, $Y = (y_i)_{i=1}^n$,
- $\boldsymbol{a}$ — модель машинного обучения,
- $\mathcal{S}$ — набор из всевозможных пар вида $(D, \boldsymbol{a})$.

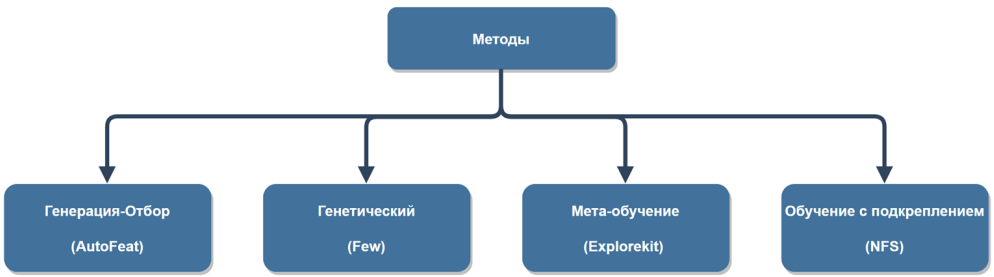
Цель

Необходимо задать функцию $\phi : \mathcal{S} \to \mathbb{R}^{n \times d}$, возвращающую трансформированные данные, которые:

- улучшают показатель качества решения итоговой задачи,
- быстро считаются,
- полезны для задач и моделей машинного обучения, отличных от пары $(D, \boldsymbol{a})$.

Обзор литературы

Классификация методов автоматической генерации признаков на табличных данных.



Нейросетевой поиск признаков (NFS) [1]

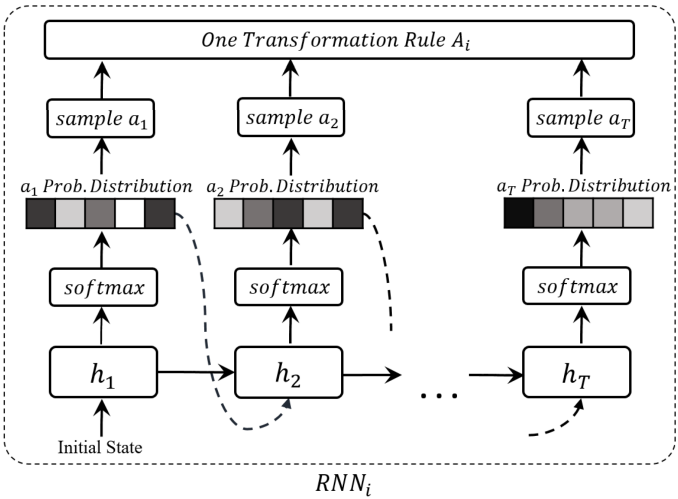


Схема работы нейронной сети для генерации трансформаций в методе NFS [1]

Нейросетевой поиск признаков (NFS) [1]

Чтобы получить награду (reward) от конкретной трансформации (action), считается разница показателей качества основной модели на датасетах с трансформированными признаками на шаге t и $t-1$:

$$R_t = Q_t - Q_{t-1}.$$

Дисконтированная кумулятивной награда считается, как

$$R_t^{(k)} = R_t + \gamma R_{t+1} + \ldots + \gamma^k R_{t+k}.$$

Пусть

$$R_t^{\lambda} = (1-\lambda) \sum_{k=1}^{p \times T} \lambda^{k-1} R_t^{(k)},$$

тогда оптимизируется функционал качества

$$L(\theta) = -\mathbb{E}_{P\left(a_{1:p\times T};\theta\right)} \left[\sum_{t=1}^{p\times T} R_t^{\lambda} \right],$$

где θ — параметры модели.

AutoLearn [2]

Autolearn в качестве отборщика признаков использует взаимную информацию (mutual Information) и корреляцию расстояния (distance correlation). Взаимная информация равна:

$$\mathrm{MI}(X_j; Y) = -\sum_{x \in \mathcal{X}_j} \sum_{y \in \mathcal{Y}} \mathbb{P}_{X_j,Y}(x,y) \log_2 \frac{\mathbb{P}_{X_j}(x)\mathbb{P}_Y(y)}{\mathbb{P}_{X_j,Y}(x,y)},$$

где $\mathbb{P}_{X_j,Y}(x,y)$ — вероятность встретить объект (x,y) в выборке, $\mathcal{Z}$ — множество значений переменной Z.

AutoLearn [2]

Пусть

$$a_{i,j} = \|x_i - x_j\|_2, b_{i,j} = \|y_i - y_j\|_2, i,j = 1,2,\ldots,n,$$
$$A_{i,j} := a_{i,j} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}, \quad B_{i,j} := b_{i,j} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..},$$

где $\bar{a}_{i.}$ — среднее значение i-ой строчки, $\bar{a}_{.j}$ — среднее значение j-го столбца.

$$\mathrm{Cov}_n^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}B_{i,j}$$

Тогда квадрат корреляции расстояния равен

$$\mathrm{Cor}_n^2(X,Y) = \begin{cases} \frac{\mathrm{Cov}_n^2(X,Y)}{\sqrt{\mathrm{Cov}_n^2(X,X)\,\mathrm{Cov}_n^2(Y,Y)}}, & \mathrm{Cov}_n^2(X,X)\,\mathrm{Cov}_n^2(Y,Y) > 0 \\ 0, & \mathrm{Cov}_n^2(X,X)\,\mathrm{Cov}_n^2(Y,Y) = 0 \end{cases}$$

Данные

id	name	abbr	nrow	ncol	ncat	nrel	source
971	mfeat-fourier	MF	2000	76	0	76	autolearn
44	spambase	SB	4601	57	0	57	safe
979	waveform	WV	5000	40	0	40	autolearn
41146	sylvine	SL	5124	20	0	20	openml
1471	eeg-eye-state	ES	14980	14	0	14	safe
42477	credit-default	DF	30000	23	0	23	nfs
4135	amazon	AZ	32769	9	0	9	nfs
1461	bank-marketing	BM	45211	16	9	7	openml
41150	miniboone	MB	130064	50	0	50	openml
1169	airlines	AL	539383	7	3	4	openml

Датасеты для тестирования методов автоматической генерации признаков.

Модели

Во всех экспериментах участвуют следующие модели:

- **бустинг (LightGBM)** с параметрами по умолчанию, за исключением параметров: bagging_freq=1, metric=auc roc, num_boost_round=10000, early_stopping_rounds=10,
- **линейная модель (Logistic Regression)** с параметрами по умолчанию, за исключением параметра max_iter=1000.

Оптимизируемые гиперпараметры бустинга, нижняя и верхняя границы.

name	low value	high value
num_leaves	2	256
feature_fraction	0.4	1.0
bagging_fraction	0.4	1.0

Оптимизируемый гиперпараметр логистической регрессии по логарифмической шкале, нижняя и верхняя границы.

name	low value	high value
C	1e-5	20.0

Ранги | Линейная модель

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
nfs	99.99	96.64	94.15	96.16	–	73.65	53.70	**91.04**	–	–	5.42
base	**100**	96.53	92.65	96.17	67.50	71.43	53.73	90.90	95.45	68.82	5.20
tfc	**100**	96.66	95.09	96.15	67.76	71.47	54.30	90.62	–	–	4.75
few	**100**	96.53	92.93	96.58	66.27	75.00	64.57	90.91	95.41	–	4.27
autolearn	**100**	**97.43**	95.12	96.08	67.50	71.43	53.73	90.90	95.45	68.82	4.25
lama	**100**	96.74	92.66	96.18	**75.47**	71.41	**85.83**	90.18	95.57	**71.44**	4.00
safe	**100**	96.70	94.92	97.29	60.22	74.50	60.45	90.70	96.34	68.40	4.00
autofeat	**100**	97.24	**95.28**	**97.44**	60.41	**75.70**	64.92	90.90	**97.47**	–	2.44

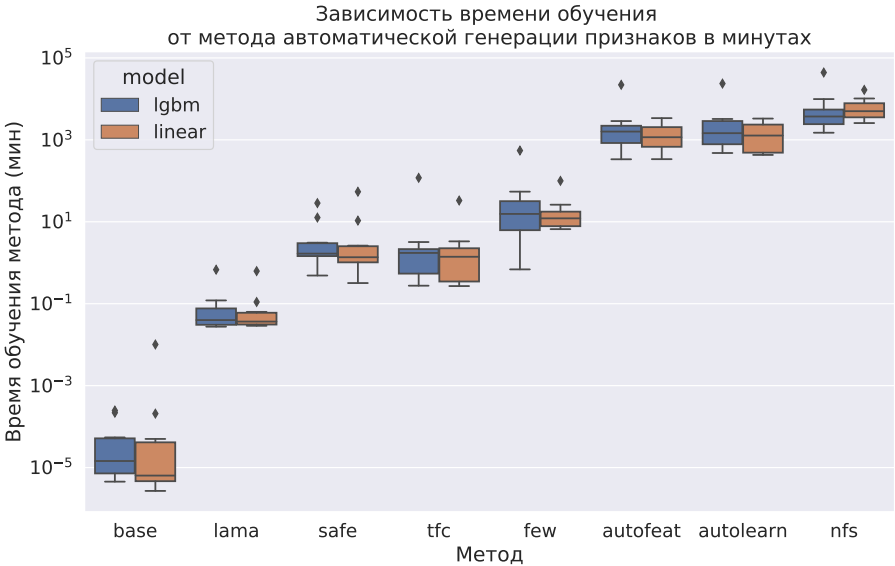
ROC AUC для модели linear на тестовых данных в зависимости от датасета и метода автоматической генерации признаков. Результаты представлены после подбора гиперпараметров моделей.

Ранги | Бустинг

	MF	SB	WV	SL	ES	DF	AZ	BM	MB	AL	rank
safe	99.91	98.50	95.67	**99.16**	98.84	76.65	81.83	92.55	98.50	71.60	6.00
tfc	99.98	98.46	95.14	98.07	98.94	77.26	84.18	92.79	98.55	71.89	5.25
few	99.96	**98.85**	95.46	98.26	98.91	77.22	83.54	93.38	98.53	–	4.72
base	99.95	**98.85**	95.46	98.26	98.91	77.22	85.76	93.24	98.53	**72.22**	4.45
autolearn	98.31	98.71	95.09	98.59	99.04	77.47	85.76	93.24	98.53	**72.22**	4.34
autofeat	**99.99**	98.72	95.66	98.92	98.62	77.58	85.27	93.27	–	71.82	3.77
nfs	99.96	98.70	95.49	98.45	**99.13**	77.48	85.96	**93.60**	–	72.16	3.27
lama	99.98	98.80	**95.68**	98.32	98.91	**77.64**	**87.05**	92.48	**98.58**	72.19	3.05

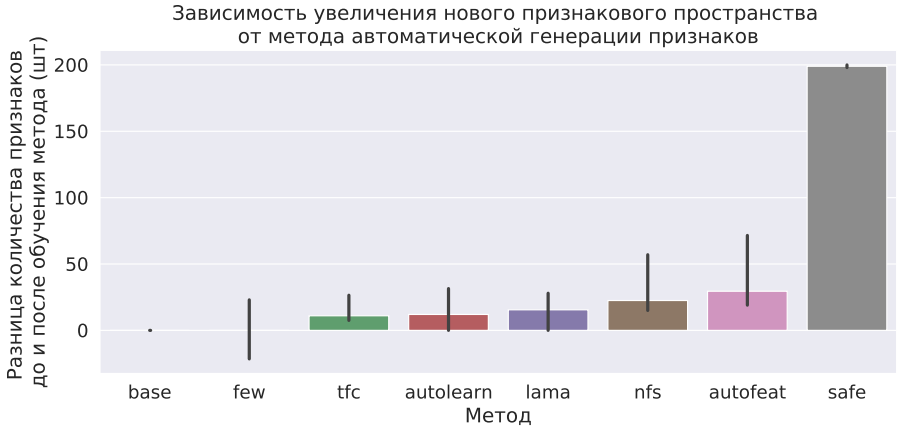
ROC AUC для модели lgbm на тестовых данных в зависимости от датасета и метода автоматической генерации признаков. Результаты представлены после подбора гиперпараметров моделей.

Время работы



Среднее время работы методов автоматической генерации признаков на 10 датасетах.

Изменение количества признаков



Среднее увеличение количества признаков после применения метода автоматической генерации признаков на 10 датасетах. Высота столбца считается, как медиана. Отрезок возле вершины столбца — 95% доверительный интервал.

Заключение

Основное улучшение качества удается добиться для логистической регрессии с помощью AutoFeat. На защиту выносится:

1. Реализация таких методов автоматической генерации признаков, как AutoLearn, TFC. Доработка NFS, SAFE.
2. Добавление интерфейса на языке python ко всем использованным методам, а также возможности запуска эксперимента в docker окружении. Код выложен и доступен в [3], [4].
3. Проведение экспериментов на искусственных и реальных данных.

[1] X. Chen.
Neural feature search: A neural architecture for automated feature engineering.
IEEE International Conference on Data Mining (ICDM). – IEEE, pages 71–80, 2019.

[2] A. Kaul, S. Maheshwary, and V. Pudi.
Autolearn—automated feature generation and selection.
IEEE International Conference on data mining (ICDM), pages 217–226, 2017.

[3] M. Kuznetsov.
Accompanying repository: Automatic feature generation for tabular data.
URL: https://hub.docker.com/u/mikkuz, 2022.

[4] M. Kuznetsov.
Accompanying repository: Automatic feature generation for tabular data.
URL: https://github.com/MikhailKuz/afg, 2022.