С. В. Новиков, Т. С. Полонская, М. В. Гресик
Орловский государственный университет, г. Орел
Пензенский государственный университет, г. Пенза
УДК 004.85

МЕТОДЫ ОТБОРА ПРИЗНАКОВ И ИХ ВЛИЯНИЕ НА КОНЕЧНЫЙ РЕЗУЛЬТАТ В МОДЕЛЯХ ОБУЧЕНИЯ С УЧИТЕЛЕМ

Аннотация. *Статья посвящена анализу выбора метода отбора признаков для обучения с учителем. Выделяются и описываются характерные особенности существующих методов. Проанализированы результаты и полученные коэффициенты моделей, определен наилучший метод для двух наборов данных с различным количеством строк. Обобщен итог и подведено наиболее лучшее решение в отборе.*

Ключевые слова: *машинное обучение, обучение с учителем, методы отбора признаков, алгоритмы регуляризации.*

Введение. В настоящее время сложно представить прикладные задачи, с которыми бы не справились модели машинного обучения (МО). Процесс обучения моделей для несложных задач занимает малое количество времени и результат, как правило, превосходит ожидания. Так, например, уже полностью можно переложить прогнозирования банковских операций на модели [1]. Конечно, только правильно подобранные параметры модели и учет всех факторов действительны способны предоставить работоспособную модель. Поэтому возникает очевидная проблема правильного отбора признаков, от которого всецело зависит успех любой модели машинного обучения.

Проблема исследования. Отбор признаков позволяет повысить качество обучения любых моделей МО: с учителем и без, а также уменьшает время на обучение и снижает требования к вычислительным мощностям аппаратуры.

Существуют некоторые алгоритмы отбора, с помощью которых становится возможным определение подходящих признаков для обеспечения более высокого качества работы задач обучения с учи-

телем (например, в задачах классификации и регрессии). Для корректной работы алгоритмов необходимо предоставить бесперебойный доступ к исходным размеченным данным. С не размеченными данными происходит похожая работа. Для них также имеется конкретный перечень методов отбора, которые сравнивают и принимают решение на основе различных критериев: дисперсии, энтропии, наличие способности сохранять локальную схожесть, и другое [2]. Самые оптимальные и, показавшие лучшие результаты признаки, обнаруженные с помощью эвристических методов без учителя, имеют место и в применении их в моделях с учителем, потому что позволяют отследить в исходных данных иные паттерны, помимо корреляции признаков с целевой переменной.

Задачу можно выделить как определение наиболее лучшего метода отбора признаков для моделей машинного обучения с учителем на примере задач регрессии.

Материалы и методы. Обучение моделей будет происходить на примере нескольких подходов с использованием методов отбора: рекурсивное удаление признаков RFE, алгоритмы регуляризации (разреженная мультиноминальная логистическая регрессия SMLR, регрессия автоматического определения релевантности ARD, Lasso, гребневая регрессия, Elastic Net) [3]. Для проведения исследования были выбраны следующие данные: набор для прогнозирования поступления в аспирантуру с точки зрения Индии на 500 строк и 8 атрибутов и набор для прогнозирования конечной стоимости дома на 1480 строк и 81 атрибут.

Результаты. Оба набора обучались на разных моделях с применением перечисленных методов. Так как они значительно различаются в количестве атрибутов, то результаты получились также различными. Перед обучением каждый массив данных был обработан только первичной обработкой данных, как например, замена пустых значений и перевод категориальных признаков в числовые.

Так как все методы проводят преобразования над признаками, вес параметров будет различен для каждой модели. Чтобы определить, в какой мере применение методов изменило старые зависимости, представим график весов параметров [4]. На рис. 1 представлен

график весов параметров перечисленных моделей для первого набора данных.

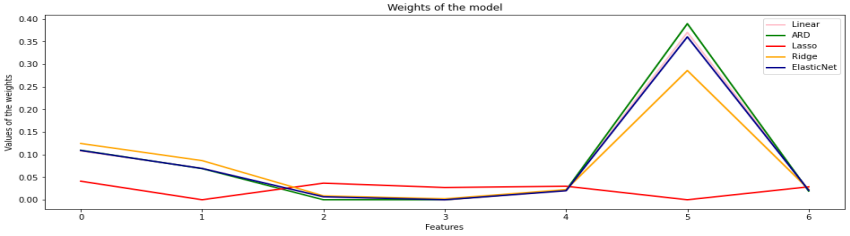

Рис. 1. График параметров первого набора данных

На рис. 2 представлен график веса параметров перечисленных моделей для второго набора данных.

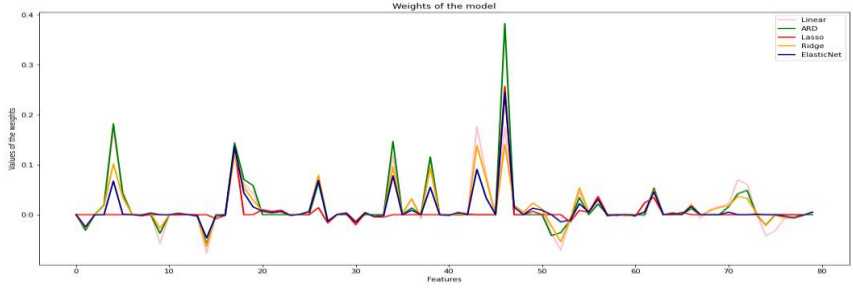

Рис. 2. График параметров второго набора данных

В результате обучения моделей по перечисленным методам получились следующие характеристики качества, представленные в табл. 1. Для каждого метода приведено значение коэффициента детерминации на тестовой выборке.

Вторая выборка дала отличные результаты от первой после применения методов. Полученные значения говорят о том, что при большой размерности данных в наборе данных качество от использования методов повышается только в том случае, если параметры имеют достаточную зависимость от целевой переменной. На основе приведенной таблице получилось, что для небольшого набора данных лучшее качество показывает, как и обычная линейная регрессия без регуляризаторов, так и с l2-регуляризатором, показав

наиболее больший коэффициент детерминации. Очевидно, что признаки, которые дают хороший результат на обычной линейной модели, при уменьшении этих признаков в гребневой регрессии, также будут возвращать относительно неплохой результат.

Таблица 1

Качество моделей
на основе метрики коэффициента детерминации

Название метода	*1 набор данных*	*2 набор данных*
LinearRegression	0.79	0.65
RFE	0.73	0.67
SMLR	0.75	0.68
ARD	0.77	0.68
Lasso	0.64	0.79
Ridge	0.79	0.70
ElasticNet	0.79	0.75

Для второго набора данных, который имеет большее число признаков, результаты показали, что лучшим методом является Lasso, что характеризуется тем, что сильный разброс параметров не дает моделям выявить закономерность и для этих целей требуется значительная предобработка. L1-регуляризация позволила занулить незначащие признаки, выявить основную закономерность в данных. Можем утверждать, что хорошо подобранные параметры для модели ElasticNet также будут давать лучший результат, так как использует тот же регуляризатор в обучении.

Заключение. Подводя итог, можно заключить, что для быстрого результата без длительной обработки данных в массивах с небольшим числом признаков подходят методы понижения размерности (гребневая регрессия). Хотя использование данного приема дает неустойчивость оценок коэффициентов регрессии, стоит это учитывать. С большим количеством признаков хорошо справляется уменьшение ограничение модели, то есть применение методов модели Lasso для зануления некоторых коэффициентов, которые повышают устойчивость модели.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян С. А. Прикладная статистика: основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — Москва : Финансы и статистика, 1983. — 471 с. — Текст : непосредственный.
2. Флах П. Машинное обучение / П. Флах. — Москва : ДМК Пресс, 2015. — 400 с. — Текст : непосредственный.
3. Рашка С. Python и машинное обучение / С. Рашка. — Москва : ДМК Пресс, 2017. — 418 с. — Текст : непосредственный.
4. Элбон К. Машинное обучение с использованием Python. Сборник рецептов / К. Элбон. — Санкт-Петербург : БХВ-Петербург, 2019. — 384 с. — Текст : непосредственный.

Г. А. Нестеров, А. Н. Абдуллин, А. С. Мингалева, И. Ю. Карякин

Тюменский государственный университет, г. Тюмень

УДК 004.45

РАЗРАБОТКА СИСТЕМЫ ПО ЧАСТИЧНОЙ АВТОМАТИЗАЦИИ РАБОТЫ КОНДИТЕРА

Аннотация. *В статье представлен процесс разработки информационной системы. Система позволяет: представить и настроить текущий ассортимент, сформировать базу клиентов и зафиксировать их заказы с учетом ограничений кондитера.*

Ключевые слова: *кондитер, интернет-кондитерская, Web-приложение, информационная система.*

Введение. Информационные системы служат для оптимизации работы человека с информацией. Поэтому такие системы распространены во всех областях деятельности человека. Так, в своей статье исследователь Екатерина Викторовна отмечает, что вовлеченный в общественное производство человек с трудом представляет себя без информационных систем [1]. Для оптимальной и быстрой разработки информационной системы необходимо использовать актуальные методы проектирования. Например, в своей статье исследователь Константин Константинович отмечает такие инструменты, как: IDEF1X, IDEF0, DFD и ARIS [2].