ОЦЕНКА ВЛИЯНИЯ ПРЕДОБРАБОТКИ ДАННЫХ НА РАБОТУ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ

Е.И. Губин, к.ф.-м.н., доцент ОИТ ИШИТР
В.А. Галлингер, студент гр. 8К71
Томский политехнический университет
E-mail: vag40@tpu.ru

Введение

Одним из важнейших направлений стратегии развития любого ВУЗа является повышение качества образования, так как оно оказывает большое влияние на компетентность будущих специалистов-выпускников ВУЗа и, следовательно, на рейтинги ВУЗа.

В данной статье сравнивается работа предсказательных моделей как на основе предобработанных и проанализированных данных, так и на непредобработанных данных (их также называют «сырыми» данными, raw data).

Работа с непредобработанными данными

В первую очередь на вход предсказательных моделей были поданы необработанные данные. Под необработанными данными понимается исходная таблица, содержащая миссинги. Алгоритмы построены так, что NaN значения являются недопустимыми, и по этой причине работать они с такими данными не будут. Самым простым решением в таком случае – убрать те наблюдения, в которых есть хотя бы одно пропущенное значение. В итоге, после выполнения данной операции, количество наблюдений (строк) сократилось почти в два раза, но и этого количества вполне достаточно, чтобы описать данные, т.к. число различных значений в каждом столбце уменьшилось всего на 30%. После некоторых преобразований данных к виду, принимаемому алгоритмами, они начали работать, но их прогнозная точность была близка к случайному угадыванию (около 33%). Это было связано с тем, что необходимо было задать всем признакам категориальный тип данных, т.к. 1 или 4 курс не является чем-то подобным температуре, и к тому же данные, более чем на половину состоящие из категориальных данных (в переводе на читаемый вид для нейронных сетей это чаще всего 0 и 1), будут плохо обрабатываться алгоритмами, если вместе с ними будут числовые данные (к примеру, признаки со значениями [0,1] и признаки со значениями от 0 до 800). После приведения всех признаков к категориальному типу, алгоритмы стали показывать точность, близкую к 99%, что объясняется наличием среди признаков таких, которые линейно зависели от целевой переменной, и такая модель, как нейронная сеть, быстро определила формулу, заданную исследователями. Для поиска линейно зависимых признаков была построена матрица корреляции, по которой были определены признаки, наиболее связанные с целевой, и были удалены из набора данных. В итоге, из 8551 строки и 16024 столбцов осталось всего 4225 строк и 8124 столбца. Получив на вход такие данные, алгоритмы работают достаточно долго, а их точность колеблется около 70%. Отдельно стоит отметить работу модели нейронной сети. В 100 эпох модель сумела обучиться на данных и достичь почти 100% точности, но как выяснилось на тесте (точность на тесте меньше 70%), модель просто сумела выставить веса таким образом, что данные, проходящие по нейронам, преобразовывались и давали верный результат. Иначе говоря, нейронной сети хватило 100 эпох, чтобы выучить данные, и в итоге она переобучилась на данных. Уменьшая число эпох до 50 и далее до 30, удалось повысить точность до 70,3%.

Работа с предобработанными данными

В случае предобработанных данных, которые были предварительно предобработаны и проанализированы А.В. Семенютой, работа алгоритмов производится заметно быстрее, что позволило поэкспериментировать с параметрами алгоритмов и построить графики зависимости точности модели от этих параметров. График зависимости точности модели от максимальной глубины дерева представлен на рисунке 1.

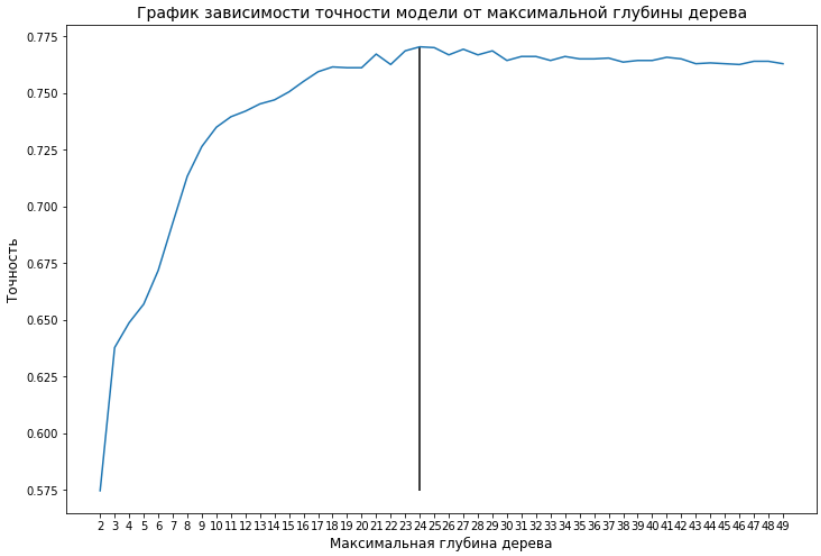

Рис. 1. Зависимость точности модели алгоритма Random Forest Classifier от его параметра

В среднем прогнозная точность на тестовой выборке на 5.5% лучше, а медианное время выполнения в 11.5 раз лучше.

Таблица 1. Показатели работы алгоритмов на предобработанных / непредобработанных данных

Название алгоритма	Прогнозная точность, %	Время обучения, с
Линейная регрессия	0.615 0.575	0.212 20.582
Логистическая регрессия	0.747 0.705	0.874 10.237
К-ближайших соседей	0.724 0.660	0.492 5.924
Случайный лес	0.770 0.692	1.418 6.680
Метод опорных векторов	0.774 0.718	6.843 87.592
Модель нейронной сети	0.754 0.703	~120 за 100 эпох ~150 за 100 эпох

Заключение

В результате проведенной работы можно заключить, что предобработка данных является важным этапом, который нельзя пропустить. В процессе работы с сырыми данными так или иначе приходилось производить операции над данными, чтобы алгоритмы смогли работать с ними. Для получения хоть какого-то адекватного результата пришлось очищать данные от пропусков, извлекать линейно зависимые признаки, а это и есть предобработка данных. Поэтому перед тем, как начинать выполнять интеллектуальный анализ данных, применять алгоритмы Data Mining, необходимо полностью убедиться, что данные подготовлены и приведены в нужный вид.

Список использованных источников

1. С. Хайкин, «Нейронные сети: полный курс, 2-е издание»: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.
2. Спицын В.Г., Цой Ю.Р. Интеллектуальные системы: Учебное пособие – Томск: Издательство ТПУ, 2012. – 72 с.
3. Губин Е.И. Методология подготовки больших данных для прогнозного анализа. / Современные технологии, экономика и образование: Сборник трудов Всероссийской научно-методической конференции. / Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2019. – 139с. – [С. 25-28].