ДОКЛАДЫ АКАДЕМИИ НАУК, 2018, том 479, № 1, с. 11–13

МАТЕМАТИКА

УДК 519.7

ДВУХУРОВНЕВЫЙ МЕТОД ПОСТРОЕНИЯ ЛИНЕЙНЫХ РЕГРЕССИЙ С ИСПОЛЬЗОВАНИЕМ НАБОРОВ ОПТИМАЛЬНЫХ ВЫПУКЛЫХ КОМБИНАЦИЙ

© 2018 г. О. В. Сенько, А. А. Докукин*, Н. Н. Киселева, Н. Ю. Хомутов

Представлено академиком РАН Ю.И. Журавлевым 27.10.2017 г.

Поступило 10.11.2017 г.

Использование многоуровневых систем обучения получает всё большее распространение при решении задач распознавания и регрессионного анализа. В сообщении рассматривается двухуровневый метод построения многомерной регрессионной модели, в котором на первом уровне генерируется семейство оптимальных выпуклых комбинаций простых одномерных МНК-регрессий. Вторым уровнем предлагаемой системы обучения является эластичная сеть. Экспериментальная проверка демонстрирует эффективность предлагаемого метода восстановления регрессии в задачах с малым количеством данных.

DOI: 10.7868/S086956521801-0016

Использование многоуровневых систем обучения, когда результаты применения узловых алгоритмов, расположенных на низком уровне, далее используются алгоритмами более высокого уровня, получает всё большее распространение при решении задач распознавания и регрессионного анализа. В этой связи следует упомянуть многослойные нейросетевые методы, а также методы алгебраической коррекции над ансамблями алгоритмов. В настоящей работе рассматривается двухуровневый метод построения многомерной регрессионной модели, в котором на первом уровне генерируется семейство оптимальных выпуклых комбинаций простых одномерных МНК-регрессий (полученных методом наименьших квадратов). Для этого используется подход [1], позволяющий генерировать семейства локально-оптимальных выпуклых комбинаций (ЛОВК) одномерных регрессий. В экспериментах на искусственно сгенерированных данных было показано, что использование взвешенных коллективных решений над наборами ЛОВК, для которых коэффициент корреляции близок к оптимальному, позволяет достигать более высокой обобщающей способности по сравнению с методом эластичной сети [2]. Вторым уровнем предлагаемой системы обучения является эластичная сеть. Экспериментальная проверка строилась на применении эластичной сети поочерёдно к исходному набору признаков некоторой задачи и к признакам, рассчитанным на основе отобранных выпуклых комбинаций.

Далее опишем сказанное более строго и приведём наиболее значимые результаты экспериментов.

Рассматривается стандартная задача многомерного регрессионного анализа. Переменная Y предсказывается по переменным $X_1,\ldots,X_n$ с помощью линейной регрессионной функции

$$\beta_0 + \sum_{i=1}^{n} \beta_i X_i.$$

Предполагается, что вектор $(\beta_0, \ldots, \beta_n)$ будем искать по обучающей выборке $(y_j, x_{1j}, \ldots, x_{nj})$, $j = 1, 2, \ldots, m$.

Предположим, что имеется набор из l предикторов, прогнозирующих значения переменной Y. Прогноз, вычисляемый i-м предиктором для некоторого объекта $x = (x_1, \ldots, x_n)$, далее обозначим через $Z_i(x)$. Пусть $c = (c_1, \ldots, c_l)$ – вектор действительных неотрицательных коэффициентов, удовлетворяющий условию $\sum_{i=1}^{l} c_i = 1$. Выпуклой комбинацией указанных предикторов будем называть процедуру, вычисляющую коллективный прогноз $Z(x, c)$ в виде

$$Z(x, c) = \sum_{i=1}^{l} c_i Z_i(X).$$

Федеральный исследовательский центр "Информатика и управление" Российской Академии наук, Москва

Институт металлургии и материаловедения им. А.А. Байкова Российской Академии наук, Москва

**E-mail: dalex@ccas.ru*

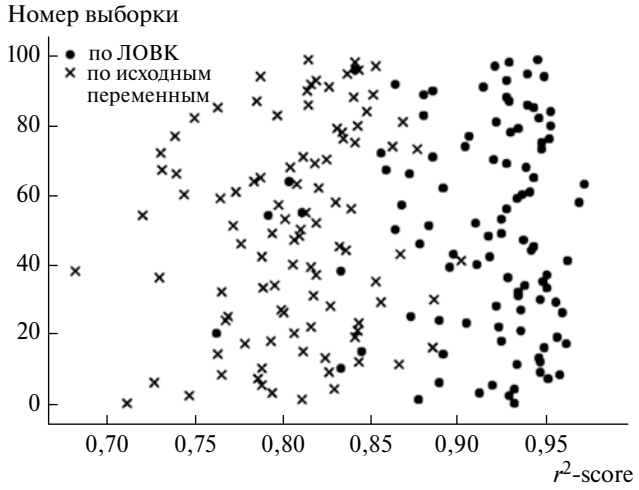

Рис. 1. Изменение качества для различных выборок при переходе в новое признаковое пространство для модельной задачи.

Оптимальной выпуклой комбинацией считается комбинация, максимально коррелирующая с целевой переменной Y [1]. В [3] описан алгоритм поиска оптимальной выпуклой комбинации, основанный на концепции несократимого нерасширяемого ансамбля предикторов.

Несократимым относительно коэффициента корреляции называется ансамбль, никакая выпуклая комбинация подмножества которого не позволяет получить большей корреляции с целевой переменной, чем некоторая выпуклая комбинация исходного набора.

Несократимый ансамбль называется нерасширяемым, если не существует другого несократимого набора, содержащего в себе все предикторы исходного.

Процедуры проверки несократимости ансамбля и вычисления коэффициентов оптимальной выпуклой комбинации для заданного набора предикторов из [1, 3] позволяют находить оптимальный ансамбль путём последовательного увеличения числа предикторов в нём. Одномерные регрессии R_i, $i = 1, 2, \ldots, n$, строятся стандартным методом наименьших квадратов на выборках отдельных признаков, т.е. множеств (Y, X_i). Проверка несократимости в этом случае означает проверку положительности коэффициента корреляции R_i и Y. Затем рассматриваются все пары предикторов, из которых отсеиваются неудовлетворяющие условию несократимости выпуклой комбинации. Далее оставшиеся пары достраиваются до троек и т.д. до тех пор, пока будут получаться несократимые наборы. Для всех несократимых наборов вычисляются коэффициенты корреляции, что позволяет найти оптимальный набор и получить множество наборов, близких к нему по качеству.

Отобранные регрессии рассматривались как новые признаки для исходной задачи. Экспериментальная проверка производилась путём обучения эластичной сети [2] на исходном наборе признаков, новом наборе и их объединении. Проверка качества осуществлялась с помощью скользящего контроля в режиме исключения одного объекта. В качестве метрик качества использовались коэффициент корреляции с переменной отклика и коэффициент детерминации r^2-score.

Эксперименты проводились с модельными данными, а также с набором задач прогнозирования переменных из областей медицины и неорганической химии. В большинстве случаев модель эластичной сети в новом признаковом пространстве имела лучшую обобщающую способность, а в некоторых случаях значительно лучшую. Однако были и отрицательные результаты, которые свидетельствуют о том, что метод всё же не универсален.

Наиболее показательны результаты модельной задачи. В данной задаче генерировалось 550 признаков, только 5% из которых релевантны. Переменная отклика линейно зависела от релевантных признаков. Было сгенерировано 100 пар выборок (обучающих и тестовых), в каждой выборке по 40 объектов.

В рамках эксперимента для каждой пары выборок выполнялась следующая процедура, результаты которой представлены на рис. 1. Сначала эластичная сеть настраивалась по обучающей выборке и полученное качество оценивалось на тестовой выборке (крестики). Затем на обучающей выборке запускалась процедура построения выпуклых комбинаций, и отбирались комбинации, коэффициент корреляции с переменной отклика которых был не меньше, чем 95% от величины корреляции для самой лучшей комбинации. Отобранные комбинации использовались в качестве нового признакового пространства, в котором уже производилось измерение качества эластичной сети при настройке на обучающей выборке и оценивались на тестовой выборке (кружки). Как видно из рис. 1, в большинстве случаев наблюдалось заметное улучшение качества.

В качестве примера прикладной задачи, для которой использование ЛОВК привело к значительному увеличению точности прогноза, можно привести задачу прогнозирования температуры плавления соединений вида $AHal_3$, где A – разные металлы, Hal – F, Cl, Br или I. В признаковое описание была отобрана информация о 55 параметрах химических элементов A и Hal. Обучающая выборка включала данные о 155 соединениях с известной температурой плавления. Исследования

с использованием режима скользящего контроля с одним направляемым на контроль объектом продемонстрировали существенное увеличение точности прогноза при использовании выпуклых комбинаций. Величина r^2-score возросла с 0,548 при использовании эластичной сети только на исходных переменных до 0,775 при использовании смеси из исходных переменных и 120 ЛОВК.

Таким образом, представленные результаты демонстрируют эффективность предлагаемого метода в задачах с малым количеством данных.

Работа выполнена при финансовой поддержке РФФИ, проекты 17–07–01362, 18–07–00080.

СПИСОК ЛИТЕРАТУРЫ

1. *Сенько О.В., Докукин А.А.* Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // ЖВМиМФ. 2011. Т. 51. № 9. С. 1751–1760.
2. *Zou H., Hastie T., Efron B., Hastie T.* Regularization and variable selection via the elastic net // J. Roy. Stat. Soc. 2005. V. 67. № 2. P. 301–320.
3. *Сенько О.В., Докукин А.А.* Регрессионная модель, основанная на выпуклых комбинациях, максимально коррелирующих с откликом // ЖВМиМФ. 2015. Т. 55. № 3. С. 530–544.