Разработка эффективных методов построения ансамблевых регрессионных линейных моделей, основанных на минимизации ошибки

A Preprint

Суржанский Александр
Московский государственный университет
имени М. В. Ломоносова
alexandr.surzhanskiy@yandex.ru

13 декабря 2024 г.

Abstract

В данной работе рассматриваются характерные черты и применение выпуклых корректирующих процедур (ВКП) в отношении множеств прогностических функций - предикторов. Обнаружено, что задача снижения обобщенной ошибки этих процедур может быть преобразована в задачу квадратичного программирования. Проанализированы условия, при которых невозможно сократить число предикторов без ухудшения точности оптимизированной ВКП. Эксперименты, проведенные на наборах одномерных линейных регрессий, продемонстрировали, что применение оптимизации ВКП может служить мощным методом для отбора и улучшения качеств наборов регрессоров.

1 Введение

В процессе решения задач, касающихся эмпирического прогнозирования и распознавания, исследователи часто сталкиваются с наличием множества различных прогностических моделей, или решающих правил. Каждая из этих моделей может не обеспечивать желаемую точность по отдельности. Исходные предикторы, как их принято называть, обычно создаются путем предварительного обучения на наборах данных. Повышение точности прогнозов может быть достигнуто за счет применения коллективных решающих стратегий, которые формируют окончательный прогноз путем объединения результатов от отдельных предикторов в виде выпуклой комбинации.

Рассмотрим набор из l предикторов, которые прогнозируют значения переменной Y. Обозначим прогноз, даваемый i -м предиктором для объекта ω, как $z_i(\omega)$. Пусть $c = (c_1, \ldots, c_l)$ — это вектор неотрицательных коэффициентов, сумма которых равна единице. В этом исследовании акцент сделан на выпуклых корректирующих процедурах (ВКП), которые вычисляют коллективный прогноз $Z(\omega, c)$.

Одним из элементарных примеров применения выпуклых корректирующих процедур является использование средних прогнозных значений. Эти методы нашли широкое признание в сфере теории распознавания и прогнозных вычислений на основе эмпирических данных. Например, они широко применяются в методах распознавания, которые полагаются на голосование среди входных систем закономерностей, а также в методах, которые призваны вычислять коллективные решения на основе множественных алгоритмов распознавания. ВКП, применяемые в контексте малых наборов линейных регрессий, как показано в предыдущих исследованиях, могут значительно превосходить по эффективности традиционные линейные регрессионные модели, чьи параметры часто вычисляются методом наименьших квадратов.

2 Постановка задачи

Покажем, что данное равенство верно:

$$\sum_{i=1}^{l} c_i[Y(\omega)-z_i(\omega)]^2 = \sum_{i=1}^{l} c_i[Y(\omega)-Z(\omega,c)+Z(\omega,c)-z_i(\omega)]^2 = \sum_{i=1}^{l} c_i[Y(\omega)-Z(\omega,c)]^2+\sum_{i=1}^{l} c_i[z_i(\omega)-Z(\omega,c)]^2 =$$

$$= [Y(\omega) - Z(\omega,c)]^2 + \sum_{i=1}^{l} c_i[z_i(\omega) - Z(\omega,c)]^2$$

Следовательно, при прогнозировании значения Y для объекта ω квадратичная ошибка ВКП может быть описана в форме разности:

$$[Y(\omega) - Z(\omega,c)]^2 = \sum_{i=1}^{l} c_i[Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^{l} c_i[z_i(\omega) - Z(\omega,c)]^2$$

Задачу выбора оптимальной ВКП можно рассматривать как задачу минимизации ожидаемой ошибки на пространстве Ω всех потенциально прогнозируемых объектов генеральной совокупности:

$$\Delta = \mathbb{E}_\Omega(\sum_{i=1}^{l} c_i[Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^{l} c_i[z_i(\omega) - Z(\omega,c)]^2)$$

при ограничениях

$$\sum_{i=1}^{l} c_i = 1; c_i \geq 0, i = 1, \ldots, l$$

Обозначим $\delta_i = \mathbb{E}_\Omega[Y(\omega)-z_i(\omega)]^2$ - математическое ожидание ошибки индивидуального прогностического алгоритма; $\rho_{ij} = \mathbb{E}_\Omega[z_i(\omega) - z_j(\omega)]^2$ - величина, характеризующая степень расхождения i-го и j-го прогностических алгоритмов.

Тогда представим функционал обобщенной ошибки в виде

$$\Delta = \sum_{i=1}^{l} c_i\delta_i + \mathbb{E}_\Omega Z^2(\omega, c) - \sum_{i=1}^{l} c_i \mathbb{E}_\Omega z_i^2(\omega) = \sum_{i=1}^{l} c_i\delta_i + \sum_{i,j=1}^{l} c_i c_j \mathbb{E}_\Omega[z_i(\omega)z_j(\omega)] - \sum_{i=1}^{l} c_i \mathbb{E}_\Omega z_i^2(\omega)$$

Тогда

$$\sum_{i,j=1}^{l} c_i c_j \mathbb{E}_\Omega[z_i(\omega)z_j(\omega)] - \sum_{i=1}^{l} c_i \mathbb{E}_\Omega z_i^2(\omega) = \{z_i z_j = \frac{1}{2}[z_i^2 + z_j^2 - (z_i - z_j)^2]\} = -\frac{1}{2}\sum_{i,j=1}^{l} c_i c_j \mathbb{E}_\Omega[z_i(\omega) - z_j(\omega)]^2 +$$

$$+\frac{1}{2}\sum_{i=1}^{l} c_i \mathbb{E}_\Omega z_i^2(\omega) \sum_{j=1}^{l} c_j + \frac{1}{2}\sum_{i=1}^{l} c_i \mathbb{E}_\Omega z_j^2(\omega) \sum_{j=1}^{l} c_j - \sum_{i=1}^{l} c_i \mathbb{E}_\Omega z_i^2(\omega) = -\frac{1}{2}\sum_{i,j=1}^{l} c_i c_j \rho_{ij}$$

Таким образом, минимизация общей ошибки может быть выражена в форме задачи квадратичного программирования:

$$\begin{cases} \sum_{i=1}^{l} c_i\delta_i - \frac{1}{2}\sum_{i,j=1}^{l} c_i c_j \rho_{ij} \to \min \\ \sum_{i=1}^{l} c_i = 1; c_i \geq 0, i = 1, \dots l \end{cases} \tag{1}$$

3 Метод решения

Предложен метод для решения задачи квадратичного программирования (1), при котором количество предикторов постепенно увеличивается в наборах, и проверяется условие их несократимости.

Условие несократимости набора предикторов подразумевает, что удаление любого элемента из набора приведет к снижению точности соответствующей оптимальной ВКП.

Обозначим

$$\overline{D}_l = \{c | \sum_{i=1}^{l} c_i = 1; c_i \geq 0, i = 1, \dots, l\}$$

$$D_l = \{c | \sum_{i=1}^{l} c_i = 1; c_i > 0, i = 1, \dots, l\}$$

Набор предикторов считается несократимым, если найдется точка $c \in D_l$, для которой

$$\Delta(c) < \Delta(c') \quad \forall c' \in \overline{D}_l \backslash D_l$$

Если набор предикторов является несократимым, коэффициенты оптимальной ВКП определяются решением специальной системы линейных уравнений. Если же набор является сократимым, оптимальную ВКП нужно искать, рассматривая различные подмножества этого набора.

Утверждение 1. Несократимость в случае двух предикторов

Чтобы обобщенная квадратичная ошибка оптимальной ВКП на основе двух предикторов оказалась меньше, чем у каждого предиктора по отдельности, необходимо и достаточно выполнение следующего неравенства

$$|\delta_2 - \delta_1| < \rho_{12}.$$

В этой ситуации обобщенная квадратичная ошибка минимизируется, если

$$c_1 = \frac{\rho_{12} + \delta_2 - \delta_1}{2\rho_{12}} \qquad c_2 = \frac{\rho_{12} + \delta_1 - \delta_2}{2\rho_{12}}$$

Доказательство утверждения было приведено в [2].

Утверждение 2. Несократимость в общем случае

Предположим, что матрица расхождений индивидуальных предикторов $||\rho_{ij}||_{l\times l}$ является невырожденной. Тогда необходимое и достаточное условие, при котором ВКП будет корректной, заключается в одновременном выполнении неравенств

$$\sum_{i=1}^{l}[\delta_i\overline{\rho}_{ij} + \frac{\frac{1}{2} - \sum_{k,t=1}^{l}\delta_t\overline{\rho}_{kj}}{\sum_{k,t}^{l}\overline{\rho}_{kt}}\overline{\rho}_{ij}] > 0 \tag{2}$$

, где матрица $||\overline{\rho}_{ij}||_{l\times l}$ является обратной матрице $||\rho_{ij}||_{l\times l}$, и положительной определенности соответствующей квадратичной формы

$$-\frac{1}{2}\sum_{i,j=1}^{l}\rho_{ij}z_iz_j \tag{3}$$

для любого вещественного вектора $z_1,\ldots,z_l$ такого, что $\sum_{i=1}^{l}z_i = 0$.

Доказательства настоящих формул и утверждения были приведены в [2].

Одним из значимых свойств выпуклых корректирующих процедур является то, что они уменьшают дисперсию коллективных прогнозов $Z(\omega)$ относительно средней дисперсии в ансамбле. Это уменьшение является непосредственным следствием следующего утверждения.

Утверждение 3. Верно следующее соотношение

$$\sqrt{\mathbb{D}(Z)} \le \sum_{i=1}^{l}c_i\sqrt{\mathbb{D}(z_i)}$$

Доказательство. Пусть $m_i = \mathbb{E}_\Omega[z_i(\omega)]$ Тогда

$$\mathbb{D}(Z) = \mathbb{E}_\Omega[c_i(z_i(\omega) - m_i)]^2 = \mathbb{E}_\Omega[\sum_{i,j=1}^{l}c_ic_j(z_i(\omega) - m_i)(z_j(\omega) - m_j)]$$

$$\mathbb{E}_\Omega[(z_i(\omega) - m_i)(z_j(\omega) - m_j)] \le \sqrt{\mathbb{D}(z_i)\mathbb{D}(z_j)}$$

Из этого следует, что

$$\mathbb{D}(Z) \le \sum_{i=1}^{l}[c_i\sqrt{\mathbb{D}(z_i)}]^2$$

□

Представим подход для решения задачи регрессии. С помощью метода наименьших квадратов строится l базовых предиктора по каждому из признаков:

$$z_i(\omega) = \alpha_i + \beta_iX_i(\omega) \quad i = 1,\ldots,l$$

После этого используется метод leave-one-out для оценки обобщенной ошибки отдельных предикторов δ_i, а также для определения параметров расхождения ρ_{ij} между ними. Затем оптимальная ВКП ищется как решение задачи квадратичного программирования (1). Пусть $c^0 = (c_1^0,\ldots,c_l^0)$ - вектор оптимальных коэффициентов ВКП. Таким образом, мы приходим к регрессионной функции

$$Z(\omega, c^0) = \sum_{i=1}^{l}c_i^0\alpha_i + \sum_{i=1}^{l}c_i^0\beta_iX_i(\omega)$$

Однако перебор всевозможных допустимых с учётом (2) и (3) пар элементарных регрессоров может требовать чрезмерное время обучения модели. В качестве альтернативной идеи будем выбирать единственную пару из допустимого набора, для которой расстояние между предикторами ρ_{ij} будет

максимальным. Так как в данной модификации при одном проходе по выборке количество возможных выпуклых комбинаций становится не более 1, будем обучать модель в течении заранее заданного числа эпох. Эпохой будем считать один проход модели по бутстрапированной выборке. Так как на разных эпохах могут строиться одинаковые наборы регрессоров, будем усреднять полученные коэффициенты для каждой выпуклой комбинации предикторов.

В задачах с высокой размерностью большинство коэффициентов c_i^0 обычно обнуляются. Благодаря этому оптимизация ВКП естественно включает в себя решение другой важной задачи регрессионного анализа — отбора информативных предикторов. В этой связи стоит упомянуть метод Лассо (см. [7]), который ищет оптимальные значения регрессионных коэффициентов при условии существования верхней границы на сумму этих коэффициентов. Подобно ВКП, метод Лассо формулируется как задача квадратичного программирования и позволяет отобрать релевантные регрессоры. Однако Лассо является, по существу, эвристическим методом, зависящим от внешнего параметра, который необходимо подбирать исследователю.

Прогноз $Z(\omega)$, получаемый с помощью ВКП, может демонстрировать высокую корреляцию с Y. Однако из-за того, что ВКП понижают общую дисперсию прогнозов относительно дисперсии отдельных предикторов, квадратичная ошибка прогнозирования может быть значительной. Поэтому требуются дополнительные линейные преобразования для $Z(\omega)$:

$$Y_{\text{pred}} = \alpha + \beta Z,$$

где коэффициенты оцениваются по обучающей выборке с помощью МНК.

4 Эксперименты

Исследования в работе проводились с помощью языка программирования Python в среде Google Colab на вычислительном устройстве Intel(R) Xeon(R) CPU @ 2.20GHz. Для воспроизводимости экспериментов был зафиксирован $random\_seed = 1$.

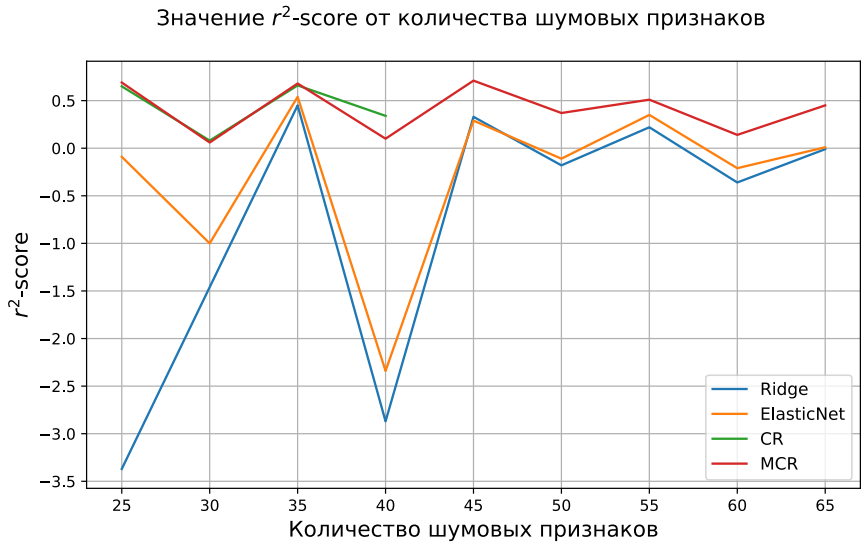

Рис. 1: Сравнение моделей по прогностическим свойствам

Решалась задача регрессии для зашумлённых данных при малом количестве объектов, поэтому датасет генерировался синтетически по следующему правилу:

- Задаются параметры генерации: n_good и n_bad, где n_good - количество признаков, от которых будет зависеть целевая переменная, n_bad - количество шумовых признаков, $n\_features = n\_good + n\_bad$ - общее число признаков
- $x_i^j \sim U(-10, 10), i = 1, \ldots, n\_features$
- $\alpha_i^j \sim N(0, 3), i = 0, \ldots, n\_good$

- $y^j = \alpha_0^j + \sum_{i \in good} \alpha_i^j x_i^j + \varepsilon^j$, где $\varepsilon^j \sim N(0, 30)$ - шумовая компонента
- Общее число объектов равняется 90.
  Выборка делится на train:validation:test в соотношении 1:1:1.

Проводилось сравнение MCR (ВКП, предложенная в данной статье) с моделями Ridge, ElasticNet, CR (стандартная ВКП) по прогностическим свойствам. В качестве основной метрики использовался r^2-score. Параметр n_good фиксировался равным 5. На графике (Рис. 1) видно, что при увеличении числа шумовых признаков модифицированный метод не имеет существенных потерь в качестве, для всех экспериментов значения r^2-score у MCR превышали значения у Ridge и ElasticNet, различия по метрике с Выпуклой регрессией (на графике продемонстрированы результаты, если время работы алгоритма менее 1000 секунд) оказались незначительными.

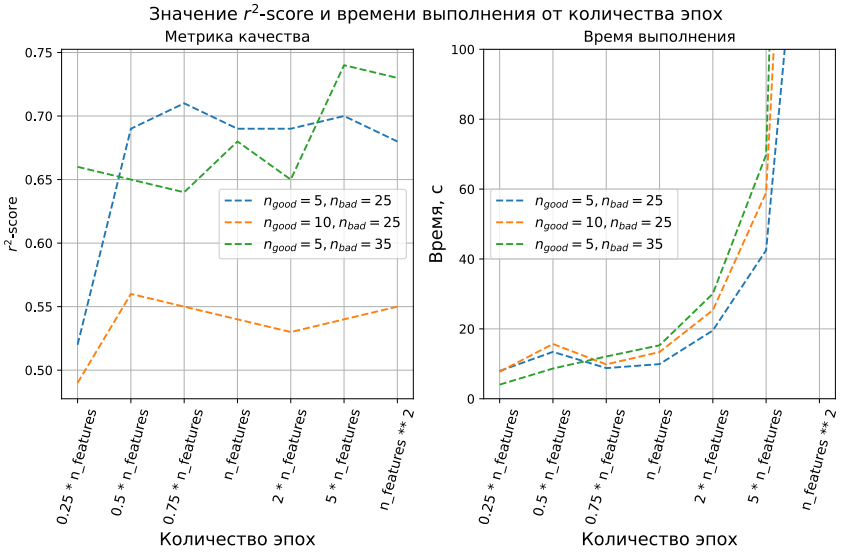

Рис. 2: Влияние количества эпох на качество модели и время обучения

В следующем эксперименте исследовалось влияние гиперпараметра модели - количества эпох, на прогностические свойства алгоритма и время его выполнения (Рис. 2). В результате выявлено, что увеличение числа эпох не влечёт за собой обязательное улучшение качества предсказаний метода. Оптимальным значением параметра по r^2-score и времени выполнения является общее количество признаков в выборке.

Модель	**Датасет 1**		**Датасет 2**	
	r^2-score	**Ненулевые признаки**	**r^2-score**	**Ненулевые признаки**
ВКП	0.9367	19	0.9521	31
CMC	0.9527	42	0.9613	68
Ridge	0.8903	83	0.9511	87
Lasso	0.8427	39	0.9492	71
ElasticNet	0.8849	62	0.9517	80

Рис. 3: Сравнение качества и отбора признаков на датасете с химическими элементами

Также были проведены эксперименты на основе двух датасетов размерами 176×94 и 92×100, содержащие различные химические элементы (CaAuBi, CdAgSb, ZrRhSb, ZrRuSb и другие). Предлагается

спрогнозировать определённый параметр (в данном случае - «a, A», «c, A») химического элемента на основании его набора характеристик. Сравнение проводилось с моделями ВКП (предложенная в данной статье), СМС (ВКП, основанная на максимизации корреляции с откликом), Ridge, Lasso и ElasticNet. По таблице (Рис. 3) можно видеть, что предложенный алгоритм имеет лучшее качество (r^2-score) по сравнению со стандартными регрессионными моделями, однако процедура, максимизирующая корреляцию с откликом, в обоих случаях имела наилучший результат. Главным преимуществом предложенного алгоритма является способность существенного отбора признаков без значительных потерь в качестве.

5 Заключение

Было установлено, что коэффициенты оптимальной выпуклой корректирующей процедуры (ВКП) определяются исключительно двумя факторами: обобщенной ошибкой каждого из компонентов и матрицей квадратов отклонений прогнозов между парами предикторов. Определение данных коэффициентов сводится к решению задачи квадратичного программирования, включающей избыточность набора предикторов. Изучена задача поиска неизбыточного множества предикторов при условии фиксированных прогнозов, получаемых от отдельных предикторов. Была исследована линейная регрессия, основанная на оптимизации ВКП и включающая интуитивный отбор значимых переменных. Эксперименты демонстрируют значительное преимущество ВКП перед пошаговой регрессией в задачах с высокой размерностью. Метод успешно сохраняет эффективность в отборое переменных и прогоностическую способность в ситуациях, когда количество переменных значительно превышает число объектов. В будущем алгоритм можно усовершенствовать, используя концепцию дивергентного леса.

Список литературы

[1] Zhuravlev Yu.I., Kuznetsova A.V., Ryazanov V.V et al. The use of pattern recognition methods in tasks of biomedical diagnostics and forecasting // Pattern Recognition and Image Analysis. 2008. V. 18. № 2. P. 195–200

[2] А. А. Докукин, О. В. Сенько, Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности, Ж. вычисл. матем. и матем. физ., 2011, том 51, номер 9, 1751–1760

[3] Kuznetsov V.A., Senko O.V. et al. Recognition of fuzzy systems by method of statistically weighed syndromes and its using for immunological and hematological norm and chronic pathology // Chem. Phys. 1996. V. 15. № 1. P. 81–100

[4] Kuncheva L.I. Combining pattern classifiers. Methods and algorithms. New Jersey: Wiley Intersci. 2004

[5] Senko O.V. The use of collective method for improvement of regression modeling stability // Statistics on Internet. 2004. http://statjournals.net/

[6] Senko O.V. An optimal ensemble of predictors in convex correcting procedures // Pattern Recognition and Image Analysis, MAIK “Nauka/Interperiodica”. 2009. V. 19. № 3. P. 465–468

[7] Tibshirani R. Regression shrinkage and selection via the lasso // J. Roy. Statist. Soc. 1996. V. 58. P. 267–288

[8] 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

[9] Efron B., Hastie T., Jonnstone I., Tibshirani R. Least angle regression // Annals of Statistics. 2004. V. 32. № 2. P. 407–499

[10] Brown G., Wyatt J.L., Tino P. Managing diversity in regression ensembles // J. Machine Learning Research. 2005. V. 6. P. 1621–1650