ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2021, том 499, с. 63–66

ИНФОРМАТИКА

УДК 004.855

ДВУХУРОВНЕВЫЙ МЕТОД РЕГРЕССИОННОГО АНАЛИЗА, ИСПОЛЬЗУЮЩИЙ АНСАМБЛИ ДЕРЕВЬЕВ С ОПТИМАЛЬНОЙ ДИВЕРГЕНЦИЕЙ

© 2021 г. Академик РАН Ю. И. Журавлев[1,*], О. В. Сенько[1,], А. А. Докукин[1,***], Н. Н. Киселева[2,****], И. А. Саенко[3,*****]**

Поступило 17.06.2021 г.
После доработки 17.06.2021 г.
Принято к публикации 19.06.2021 г.

Рассмотрен новый двухуровневый метод регрессионного анализа, в котором корректирующая процедура применяется к оптимальным ансамблям регрессионных деревьев. При этом оптимизация производится исходя из одновременного достижения расходимости алгоритмов в пространстве прогнозов и хорошей аппроксимации данных отдельными алгоритмами ансамбля. В качестве корректирующих процедур рассматриваются простое усреднение, случайный регрессионный лес и градиентный бустинг. Приведены эксперименты по сравнению предложенного метода со стандартным решающим лесом и стандартным методом градиентного бустинга для решающих деревьев.

Ключевые слова: регрессия, коллективные методы, бэггинг, градиентный бустинг

DOI: 10.31857/S2686954321040172

ВВЕДЕНИЕ

Методы регрессионного моделирования, основанные на вычислении более точного коллективного прогноза, по прогнозам, вычисленным набором (ансамблем) менее точных и более простых исходных алгоритмов, получили самое широкое распространение в современном машинном обучении. К числу таких методов может быть отнесен регрессионный случайный лес, а также методы, основанные на использовании адаптивного или градиентного бустинга. Важную роль при построении коллективных алгоритмов играет способ получения исходного ансамбля так называемых слабых алгоритмов. Теоретический анализ показывает, что увеличение обобщающей способности может быть достигнуто за счет выбора ансамбля алгоритмов, обладающих не только высокой точностью, но и максимально расходящимися прогнозами [1]. Низкая коррелированность прогнозов потенциально позволяет также добиваться более точной аппроксимации алгоритма, объективно обеспечивающей наиболее точный прогноз, с использованием ограниченного числа алгебраических операций [2, 3]. В методе "регрессионный случайный лес" расходимость прогнозов достигается за счет обучения алгоритмов ансамбля на различных выборках, генерируемых из исходной обучающей выборки с использованием процедуры бутстрэпа [4]. В методе градиентного бустинга [5] ансамбль генерируется последовательно. При этом на каждой итерации в ансамбль добавляются деревья, аппроксимирующие первые производные функции потерь по переменным, соответствующим коллективному прогнозу.

Другой важной составляющей является способ вычисления коллективного прогноза, который может также интерпретироваться как результат взаимной коррекции прогнозов. В методе "случайный регрессионный лес" коррекция осуществляется простым вычислением средних прогнозов.

Другим возможным способом организации корректирующей процедуры является схема стэкинга, в которой выходы алгоритмов ансамбля рассматриваются как входные признаки алгоритма,

[1] *Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, Россия*
[2] *Институт металлургии и материаловедения им. А.А. Байкова Российской академии наук, Москва, Россия*
[3] *Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
**E-mail: zhur@ccas.ru*
***E-mail: senkoov@mail.ru*
****E-mail: dalex@ccas.ru*
*****E-mail: kis@imet.ac.ru*
******E-mail: i.a.saenko@mail.ru*

вычисляющего выходной скорректированный прогноз [6, 7]. Эффективность стэкинга, как правило, оказывается низкой при использовании его для вычисления коллективных решений по наборам слабых алгоритмов, генерируемых с помощью процедур, применяемых при генерации случайных лесов. Можно предположить, что причиной снижения эффективности является недостаточное расхождение слабых алгоритмов в пространстве прогнозов при стандартных способах генерации ансамблей.

Целью работы является исследование эффективности двухуровневого метода увеличения обобщающей способности, предусматривающего построение ансамбля, состоящего из алгоритмов с высокой степенью расхождения в пространстве прогнозов и хорошей аппроксимацией этими алгоритмами целевой переменной. При этом в качестве корректирующей процедуры рассматриваются простое усреднение и стэкинг.

ДВУХУРОВНЕВЫЙ МЕТОД

Пусть $\{A_1(X),\ldots,A_{k-1}(X)\}$ – некоторый ансамбль алгоритмов, предсказывающих значение переменной y по вектору X переменных $x_1,\ldots,x_n$. Предполагается, что алгоритмы ансамбля обучаются по выборке $S=\{(X_1,y_1),\ldots,(X_m,y_m)\}$. Предварительно выбирается базовый метод регрессионного анализа. Обычно в качестве такого метода выступает модель регрессионного дерева. Обозначим $L_k(X)=\frac{1}{k}\sum_{i=1}^{k}A_i(X)$, $Q_k(X)=\frac{1}{k}\sum_{i=1}^{k}A_i^2(X)$. В соответствии с объявленной целью ансамбль должен строиться, исходя из одновременной минимизации критерия Φ_E:

$$\Phi_E(A_1(X),\ldots,A_k(X))=\frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(y_j-A_i(X_j)\right)^2,$$

оценивающего среднюю ошибку аппроксимации y по вектору X, и Φ_V:

$$\Phi_V(A_1(X),\ldots,A_k(X))=\frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(L_k(X_j)-A_i(X_j)\right)^2,$$

представляющего собой дисперсию прогнозов вычисляемых алгоритмами ансамбля.

Задача одновременной минимизации Φ_E и максимизации Φ_E может быть сведена к минимизации

$$\Phi_G=(1-\mu)\Phi_E-\mu\Phi_V,$$

где $\mu\in[0,1]$ регулирует вклад разнородности ансамбля в смысле дисперсии прогнозов.

Обозначим через D_E^k и D_V^k изменения функционалов Φ_E и Φ_V при включении в ансамбль дополнительно алгоритма A_{k+1}.

$$D_E^k=\Phi_E(A_1(X),\ldots,A_{k+1}(X))-\Phi_E(A_1(X),\ldots,A_k(X))=$$
$$=\left(\Phi_E(A_1(X),\ldots,A_k(X))*k+\right.$$
$$\left.+\frac{1}{m}\sum_{j=1}^{m}(y_j-A_{k+1}(X_j))^2\right)\frac{1}{k+1}-$$
$$-\Phi_E(A_1(X),\ldots,A_k(X))=\frac{1}{m(k+1)}\times$$
$$\times\sum_{j=1}^{m}\left(y_j-A_{k+1}(X_j)\right)^2-\frac{1}{k+1}\Phi_E(A_1(X),\ldots,A_k(X))=$$
$$=\frac{1}{m(k+1)}\sum_{j=1}^{m}\left(y_j-A_{k+1}(X_j)\right)^2-C_E,$$

где C_E не зависит от $A_{k+1}(X)$.

Для расчета D_V потребуется использовать известное выражение для дисперсии:

$$\sum_{j=1}^{m}\sum_{i=1}^{k}\left(L_k(X_j)-A_i(X_j)\right)^2=\sum_{j=1}^{m}(Q_k(X_j)-L_k^2(X_j));$$
$$D_V^k=\Phi_V(A_1(X),\ldots,A_{k+1}(X))-\Phi_V(A_1(X),\ldots,A_k(X))=$$
$$=\frac{1}{m}\sum_{j=1}^{m}(Q_{k+1}(X_j)-L_{k+1}^2(X_j))-$$
$$-\frac{1}{m}\sum_{j=1}^{m}(Q_k(X_j)-L_k^2(X_j))=\frac{1}{m(k+1)}\times$$
$$\times\sum_{j=1}^{m}\left(-Q_k(X_j)+A_{k+1}^2(X_j)-\right.$$
$$\left.-\frac{1}{k+1}\left(kL_k(X_j)+A_{k+1}(X_j)\right)^2+(k+1)L_k^2(X_j)\right)=$$
$$=\frac{k}{m(k+1)^2}\sum_{j=1}^{m}(A_{k+1}^2(X_j)-2L_k(X_j)A_{k+1}(X_j))+C_V,$$

где C_V не зависит от $A_{k+1}(X)$.

Одним из способов решения задачи минимизации Φ_G является сведение ее к поиску и включению в ансамбль алгоритма A_{k+1}, для которого оказывается минимальным функционал D_G:

$$D_G^k=(1-\mu)D_E^k-\mu D_V^k=$$
$$=\frac{1-\mu}{m(k+1)}\sum_{j=1}^{m}\left(y_j-A_{k+1}(X_j)\right)^2-$$
$$-\frac{\mu k}{m(k+1)^2}\sum_{j=1}^{m}(A_{k+1}^2(X_j)-2L_k(X_j)A_{k+1}(X_j))+C_G,$$

Таблица 1. Результаты экспериментов

	reference	forest	boosting	average
$I2/m$, a	0.943 ± 0.0012	0.948 ± 0.0031	0.941 ± 0.005	0.94 ± 0.0025
$I2/m$, b	0.692 ± 0.0087	0.78 ± 0.0038	0.746 ± 0.012	0.758 ± 0.0088
$I2/m$, c	0.847 ± 0.004	0.875 ± 0.0052	0.856 ± 0.0051	0.87 ± 0.0063
$I4/m$, c	0.99 ± 0.0015	0.993 ± 0.0006	0.993 ± 0.0007	0.994 ± 0.0003
$Fm3(-)m$, a	0.636 ± 0.0019	0.639 ± 0.0026	0.615 ± 0.0034	0.631 ± 0.0041
$P2_1/n$, a	0.593 ± 0.0075	0.575 ± 0.005	0.575 ± 0.0052	0.572 ± 0.0041
$P2_1/n$, b	0.928 ± 0.0004	0.903 ± 0.0019	0.895 ± 0.0025	0.906 ± 0.0009
$P2_1/n$, c	0.441 ± 0.0047	0.516 ± 0.0103	0.439 ± 0.013	0.54 ± 0.008
$R3(-)$, c	0.319 ± 0.0146	0.343 ± 0.013	0.364 ± 0.0178	0.346 ± 0.0134
A_3BHal_6, Tm	0.903 ± 0.0001	0.893 ± 0.0015	0.89 ± 0.0014	0.895 ± 0.0009
$ABHal_3$, Tm	0.861 ± 0.02	0.874 ± 0.022	0.881 ± 0.021	0.871 ± 0.021

где C_G не зависит от $A_{k+1}(X)$. Построение алгоритма, включаемого в ансамбль на шаге $k+1$ производится с помощью комбинации бэггинга и варианта градиентного бустинга, реализуемой в два этапа:

1. На первом этапе из исходной выборки S с использованием датчика случайных чисел генерируется выборка с возвращениями S^0_{k+1}. Алгоритм A^0_{k+1} обучается далее по выборке S^0_{k+1}.

2. На втором этапе строится алгоритм A_{k+1}, который собственно должен войти в ансамбль. Прогноз значения y в точке X вычисляется алгоритмом A_{k+1} по формуле

$$A_{k+1}(X) = A^0_{k+1}(X) - \varepsilon G_{k+1}(X),$$

где $G_{k+1}(X)$ – алгоритм, вычисляющий в точке $(A^0_{k+1}(X_1),\ldots,A^0_k(X_m))$ прогноз градиента функционала $D^k_G(A_k(X_1),\ldots,A_{k+1}(X_m))$.

Алгоритм G_{k+1} обучается по выборке $\left\{\left(X_1,\left.\frac{\partial D^k_G}{\partial A_{k+1}(X_1)}\right|_{A^0_{k+1}(X_1)}\right),\ldots,\left(X_m,\left.\frac{\partial D^k_G}{\partial A_{k+1}(X_m)}\right|_{A^0_{k+1}(X_m)}\right)\right\}$.

Нетрудно показать, что

$$\frac{\partial D^k_G}{\partial A_{k+1}(X_i)} = \frac{-2(1-\mu)}{m(k+1)}\left(y_i - A_{k+1}(X_i)\right) - \frac{\mu 2k}{m(k+1)^2}\left(A_{k+1}(X_i) - L_k(X_i)\right).$$

ЭКСПЕРИМЕНТЫ

Метод был реализован на языке python с помощью библиотеки scikit-learn [8]. Так, базовые деревья строятся с помощью метода BaggingRegressor. На втором уровне используется либо GradientBoostingRegressor (далее этот вариант называется boosting), либо RandomForestRegressor (forest), либо усредняются результаты базовых методов (average). GradientBoostingRegressor также используется в качестве референсного метода для оценки эффективности предлагаемого метода.

Разработанный метод использовался для прогнозирования различных параметров кристаллической решетки сложных неорганических соединений: параметров кристаллической решетки соединений состава $A_2^{2+}B^{+3}C^{+5}O_6$, и температуры плавления галогенидов состава A_3BHal_6 и $ABHal_3$. Для различных пространственных групп симметрии приведены результаты для части параметров, допускавших достаточно надежный прогноз, а именно: для моноклинных пространственных групп (пр. гр. $I2/m$ и $P2_1/n$) предсказывались параметры a, b и c, для тетрагональных (пр. гр. $I4/m$) и гексагональных (пр. гр. $R3(-)$) – параметр c, для кубических (пр. гр. $Fm3(-)m$) – параметр a. Для оценки точности использовался стандартный показатель r^2, или коэффициент детерминации, который вычислялся с использованием метода кросс-валидации. В связи с тем, что генерация алгоритмов в процедуре бэггинга происходит в значительной мере случайно, результаты решения одной и той же задачи изменяются от эксперимента к эксперименту. Поэтому в таблице для каждой задачи приведено значения показателя r^2, усредненное по 10 экспериментам.

ЗАКЛЮЧЕНИЕ

Из результатов, приведенных в табл. 1, видно, что в большинстве случаев предлагаемый двухуровневый метод позволяет добиваться более высоких результатов по сравнению со стандартным

алгоритмом градиентного бустинга, что указывает на целесообразность дальнейших исследований в данном направлении. Предполагается исследование возможности подбора оптимальной в смысле ряда критериев длины шага градиентного спуска при коррекции алгоритмов, сгенерированных с помощью бэггинга.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 18-29-03151, 20-01-00609, 21-51-53019.

СПИСОК ЛИТЕРАТУРЫ

1. *Докукин А.А., Сенько О.В.* Регрессионная модель, основанная на выпуклых комбинациях, максимально коррелирующих с откликом // ЖВМиМФ. 2015. Т. 55. № 3. С. 530–544.
2. *Журавлёв Ю.И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов I // Кибернетика. 1977. 4. С. 14–21.
3. *Журавлёв Ю.И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов II // Кибернетика. 1977. 6. С. 21–27.
4. *Breiman L.* Bagging predictors // Machine Learning. 1996. № 24. P. 123–140.
5. *Hastie T., Tibshirani R., Friedman J.H.* 10. Boosting and Additive Trees. The Elements of Statistical Learning (2nd ed.). New York: Springer, 2009. P. 337–384. ISBN 978-0-387-84857-0.
6. *Wolpert D.H.* Stacked Generalization // Neural Networks. 1992. V. 5. № 2. P. 241–259.
7. *Breiman L.* Stacked Regressions // Machine Learning. 1996. V. 24. P. 49–64.
8. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.* Scikit-learn: Machine Learning in Python // J. Machine Learning Research. 2011. № 12. P. 2825–2830.

TWO-LEVEL REGRESSION METHOD USING ENSEMBLES OF TREES WITH OPTIMAL DIVERGENCE

Academician of the RAS Yu. I. Zhuravlev[a], O. V. Senko[a], A. A. Dokukin[a], N. N. Kiselyova[b], and I. A. Saenko[c]

[a] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russian Federation*
[b] *Institute of Metalluegy and Materials Science of the Russian Academy of Sciences, Moscow, Russian Federation*
[c] *Lomonosov Moscow State University, Moscow, Russian Federation*

The article discusses a new two-level regression analysis method, in which a corrective procedure is applied to optimal ensembles of regression trees. Optimization is carried out based on the simultaneous achievement of the divergence of the algorithms in the forecast space and a good approximation of the data by individual algorithms of the ensemble. Simple averaging, random regression forest, and gradient boosting are considered as corrective procedures. Experiments are presented comparing the proposed method with the standard decision forest and the standard gradient boosting method for decision trees.

Keywords: regression, ensembles, bagging, gradient boosting