Управление разнообразием моделей в ансамблях регрессоров.

Managing ambiguity in regression ensembles

Ключевой проблемой при построении ансамблей в машинном обучении с учителем является одновременное достижение двух в значительной мере взаимоисключающих целей: (1) минимизация ошибок индивидуальных предикторов, (2) обеспечение их разнообразия. Данная проблема в общем виде еще не решена, хотя бы потому, что измерение разнообразия представляет трудности ввиду нечеткого определения самого этого понятия.

Наиболее популярные на сегодняшний день алгоритмы концентрируются либо на разнообразии предикторов за счет их обучения на подмножествах обучающего набора данных (Bagging) в комбинации с подмножеством признаков (Random Forest), либо на минимизации ошибки ансамбля в целом (Boosting).

Тем не менее, для задачи регрессии известно разложение ошибки ансамбля на взвешенные суммы ошибок предикторов и их несогласованностей (ambiguity). Krogh and Vedelsby (1995) показали, что в каждой точке

$$(f_E - y)^2 = \sum_{i=1}^{M} w_i (f_i - y)^2 - \sum_{i=1}^{M} w_i (f_i - f_E)^2.$$

Здесь y значение моделируемой функции в данной точке, f_E ансамбль M регрессоров f_i, $i = 1, \dots, M$, причем f_E является выпуклой комбинацией индивидуальных предикторов:

$$f_E = \sum_{i=1}^{M} w_i f_i\,, \qquad \sum_{i=1}^{M} w_i = 1\,, \qquad w_i \geq 0, \qquad i = 1, \dots, M$$

Из приведенного выражения следует, что $(f_E - y)^2 = 0$ когда, в частности, $(f_i - y)^2 - (f_i - f_E)^2 = 0$ для $i = 1, \dots, M$, что приводит к условию $(y - f_E)(y + f_E - 2f_i) = 0$.

Предположим, что мы имеем ансамбль из M-1 обученных предикторов и наша задача – добавить в него новую модель так, чтобы общая ошибка приближалась к нулю. Ансамбль на шаге M можно представить как

$$f_E = \sum_{i=1}^{M} w_i f_i = \sum_{i=1}^{M-1} w_i f_i + w_M f_M.$$

В комбинации с условием $(y - f_E) = 0$ это дает значение целевой переменной для обучения M-го предиктора:

$$f_M = \frac{1}{w_M}\left(y - \sum_{i=1}^{M-1} w_i f_i\right).$$

Однако, эту формулу нельзя использовать непосредственно, поскольку она включает неизвестные веса w_i, $i = 1, \dots, M$. Решение этой проблемы возможно путем задания априорного распределения весов модели, например, в виде $w_i = 1/M$. Тогда алгоритм построения ансамбля примет следующий вид:

Алгоритм 1. Managed Ambiguity Regressor.

Вход: обучающая выборка $(\mathbf{x}, y)$, количество регрессоров в ансамбле M.

1. Обучить предиктор f_1 на выборке $(\mathbf{x}, y)$.
2. Для всех $m = 2, \ldots, M$
   a. Найти значение $t_m = my - \sum_{i=1}^{m-1} f_i$;
   b. Обучить предиктор f_m на выборке $(\mathbf{x}, t_m)$.

Выход: ансамбль $f_E = \sum_{i=1}^{M} w_i f_i$.

Таблица 1. Среднеквадратическая ошибка различных алгоритмов на реальных наборах данных

Dataset	GB	MA	RF	BR
AEP	15223.190	33485.105	13659.290	**13601.370**
Airfoil	19.207	**18.210**	30.194	30.149
Bike	9040.414	**4535.852**	18687.546	18679.854
Boston	**21.466**	24.342	26.775	26.781
Cadata	5604.245	**4924.229**	7713.155	7712.682
CASP	24.037	**20.262**	28.884	28.883
CCPP	16.076	**14.052**	23.785	23.787
News	139.270	162.424	**136.395**	136.652
PM2.5	5247.190	**4790.049**	6616.587	6616.915
SC	223.861	**207.003**	352.993	353.010
Total wins	1	**7**	1	1

Таблица 2. Характеристики наборов данных.

Dataset	Description	Number of features	Samples	Source
AEP	Appliances Energy Prediction	26	19735	UCI[1]
Airfoil	Airfoil Self-Noise	5	1503	UCI
Bike	Bike Sharing	12	17389	UCI
Boston	Boston House Prices	13	506	Sklearn[2]
Cadata	California House Prices	8	20640	StatLib[3]
CASP	Physicochemical Properties of Protein Tertiary Structure	9	45730	UCI
CCPP	Combined Cycle Power Plant Dataset	4	9568	UCI
News	Online News Popularity	58	39644	UCI
PM2.5	Beijing PM2.5	14	41757	UCI
SC	Superconductivity	81	21263	UCI

Данный алгоритм минимизирует общую ошибку ансамбля, учитывая требование разнообразия. Результаты тестов на 10 реальных наборах данных (Таблица 1) показывают, что он значительно превосходит другие методы построения регрессионных ансамблей.

[1] UCI – UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/
[2] Data mining and data analysis library http://scikit-learn.org/stable/index.html
[3] StatLib – Carnegie-Mellon StatLib repository http://lib.stat.cmu.edu/datasets/

Характеристики наборов данных, использованных для тестирования приведены в Таблице 2.

В Таблице 1 приведены результаты сравнения следующих методов построения ансамблей: GB – Gradient Boosting, RF – Random Forest, BR – Boosting Regressor (использованы реализации из библиотеки scikit-learn) и MA – Managed Ambiguity, алгоритм, предложенный здесь. В качестве базовой модели для всех алгоритмов использовано дерево решений с максимальной глубиной 3, количество предикторов во всех моделях – 50.

Литература:

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems* (pp. 231-238).