



ем технологий HTML5, JavaScript, дополнительных наборов библиотек (Entity Framework, Angular 2 и других) в среде программирования Visual Studio 2015. В качестве системы управления базами данных используется Microsoft SQL Server 2012 Enterprise.

Подсистема поможет пользователям получить необходимую информацию для выполнения расчета, просмотреть данные по всем системам, дозагрузить недостающие данные или перезагрузить изменившиеся данные, настроить при необходимости списки параметров, значения которых будут загружены при изменении алгоритмов расчета, выбрать алгоритмы уменьшения и восстановления данных.

Е.А. Пономарев, И.М. Куликовских

МЕТОД ГРУППОВОГО ОБУЧЕНИЯ АЛГОРИТМОВ НА ОСНОВЕ ЭФФЕКТА ЗАБЫВАНИЯ, ВЫЗВАННОГО ИЗВЛЕЧЕНИЕМ ИНФОРМАЦИИ

(Самарский национальный исследовательский университет
им. академика С.П. Королева, Самара, Россия)

Впервые идея бустинга рассмотрена в работах Майкла Кернса и Лесли Валианта [1,2]. В своих работах они ставят вопрос — может ли группа слабых алгоритмов обучения (классификаторов) сформировать один сильный?

Под *слабым* обучаемым подразумевается алгоритм обучения, дающий результат лишь несколько лучше случайного угадывания. Более строго: будем называть алгоритм бинарной классификации *слабым*, если его точность распознавания только немного больше 50%.

Сильным обучаемым будем называть алгоритм, который бы мог давать сколь угодно точные результаты. Исследования показали, что сильная обучаемость эквивалентна слабой, поскольку любую слабую модель можно усилить, построив правильную композицию.

Первым практическим применением этой идеи можно считать алгоритм AdaBoost, предложенный Йоавом Фройндом и Робертом Шапира [3]. Пусть имеется набор слабых классификаторов $h_t: X \rightarrow \{+1, -1\}$. Алгоритм AdaBoost предлагает процесс построения классификатора в виде $H(x) = \sum_{t=1}^T \alpha_t h_t(x)$, минимизирующего экспоненциальную функцию потерь $L(y, F) = \exp(-yF)$.

LogitBoost был разработан как модификация AdaBoost с учетом идей логистической регрессии. Функция потерь в LogitBoost имеет вид $L(y, F) = \log(1 + \exp(-yF))$, где $F(x) = \log\left(\frac{P(y = +1|x)}{P(y = -1|x)}\right)$.

Важным методом повышения обучающей способности разрабатываемых алгоритмов является регуляризация — введение дополнительных параметров в математические модели, добавляющие дополнительную степень свободы и, таким образом, приводящие к большей гибкости. Проблемой явной регуляризации является явное изменение структуры алгоритма и метода оптимизации. В



связи с этим большое распространение получают исследования *неявной* регуляризации.

В данной работе предлагается использование эффекта забывания, вызванного извлечением информации (Retrieval-Induced Forgetting, RIF) для повышения обучающей способности метода группового обучения алгоритмов. Данный эффект возникает как следствие сознательного запоминания тестируемой информации через ее явное извлечение (ось угадывания), что, в свою очередь, приводит к неявному забыванию (ось забывания) нетестируемой информации на бессознательном уровне.

Регрессионные модели, в том числе LogitBoost, обладают эффектами «пола» и «потолка», которые возникают из-за вида логистической функции при стремлении параметров к бесконечности. Данные эффекты могут контролироваться дополнительными параметрами, осуществляющими неявную регуляризацию. Параметры, по аналогии с теорией тестирования, можно рассматривать как параметры *забывания* и *угадывания*. Таким образом, модель регрессии модифицируется с учетом эффекта RIF.

Модифицируем алгоритм группового обучения LogitBoost, применяя неявную регуляризацию параметрами забывания и угадывания [5]. Используемая в функции потерь логистическая функция имеет вид:

$$F(x) = \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}} \quad (1)$$

Функция, использующие только параметры забывания (floor, эффект «пола») и параметр угадывания (ceil, эффект «потолка») [6]:

$$F_{floor}(x) = c + \frac{(1-c)e^{F(x)}}{e^{F(x)} + e^{-F(x)}} \quad (2)$$

$$F_{ceil}(x) = \frac{(1-c)e^{F(x)}}{e^{F(x)} + e^{-F(x)}}, \quad (3)$$

а также логистическая функция, объединяющая оба эффекта [5,6]:

$$F_{both}(x) = c_g + \frac{(1-(c_g+c_f))e^{F(x)}}{e^{F(x)} + e^{-F(x)}} \quad (4)$$

Оптимальные параметры для модели подбирались путем кросс-валидации. В качестве слабых обучающихся использовались модели линейной регрессии с взвешенным МНК.

Для проверки разработанного алгоритма использовались несколько наборов данных, в основном собранных и предобработанных вручную. В качестве таких данных выступали курс рубля к доллару, динамика стоимости золота за последние 8 лет, а также данные с криптовалютных рынков [4].

На каждом наборе данных были проведены испытания стандартного и модифицированного алгоритма. Каждый из алгоритмов запускался несколько раз, результат усреднялся и анализировался. На рисунке 1 показано сравнение полученных графиков для набора данных динамики курса золота за 4 года. Дисперсия для данного набора данных уменьшилась на 57,86% по сравнению со стандартным алгоритмом LogitBoost, среднее качество обучения незначительно ухудшилось на 0,09%. В таблице приведены результаты для нескольких тестовых наборов данных.

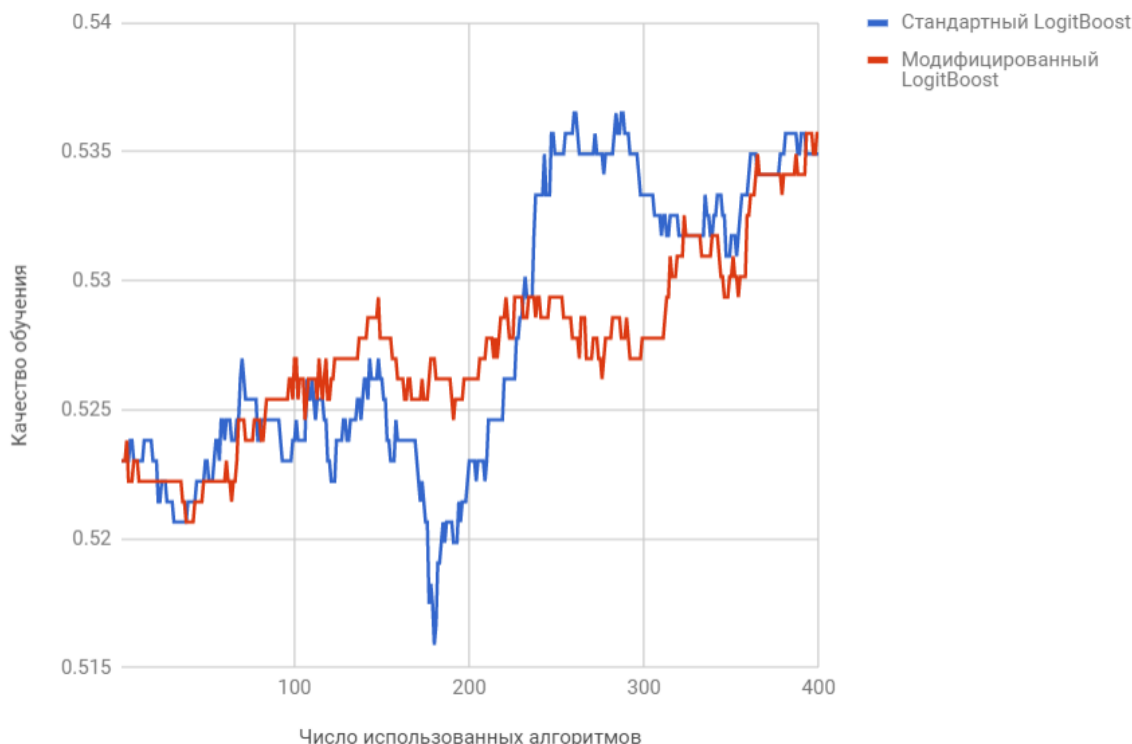


Рисунок 1 — Графики качества обучения стандартного и модифицированного алгоритма LogitBoost на наборе данных курса золота

Таблица 1. Анализ качества обучения

Набор данных	Изменение среднего качества обучения, %	Изменение в дисперсии качества обучения, %
Курс золота за 4 года	-0,09	-57,86
Курс золота за 1 год	-2,11	-48,31
Курс рубля к доллару за 5 лет	0,29	-11,25

Качество полученного алгоритма зависит от набора данных и изменения ряда параметров (размер окна прогнозирования, коэффициент скорости обучения, количество итераций). Тем не менее, сравнение стандартного и улучшенного алгоритмов проводилось при фиксированных параметрах. Единственным изменением было введение дополнительных коэффициентов c_g и c_f .

Таким образом, улучшенный алгоритм демонстрирует большую устойчивость и более стабильный процесс обучения при практически неизменном качестве обучения.

Благодарности

Работа выполнена при поддержке гранта Президента РФ № МК-6218.2018.9.

Литература

1. Kearns M. (1988). Thoughts on Hypothesis Boosting, Unpublished manuscript (Machine Learning class project, December 1988).



2. Kearns M.; Valiant L. (1989). Cryptographic limitations on learning Boolean formulae and finite automata. Symposium on Theory of computing. ACM. – 21. – pp. 433–444.
3. Freund Y., Shapire R. Experiments with a New Boosting Algorithm. – 1996.
4. Investing.com [Электронный ресурс] — <https://www.investing.com>, дата обращения 14.01.2018.
5. Kulikovskikh I.M. (2017). Cognitive validation map for early occupancy detection in environmental sensing // Engineering Applications in Artificial Intelligence – 65. – pp. 330-335.
6. Kulikovskikh I.M., Prokhorov S.A. (2017). Minimizing the effects of floor and ceiling to improve the convergence of log-likelihood // Procedia Engineering. – 201. – pp. 779-788.

А.В. Пулькин

ОЦЕНКА НАПРАВЛЕНИЯ РОСТА ТРЕЩИНЫ В УСЛОВИЯХ
СМЕШАННОГО НАГРУЖЕНИЯ: КРИТЕРИИ РАЗРУШЕНИЯ,
ОСНОВАННЫЕ НА МНОГОПАРАМЕТРИЧЕСКОМ ОПИСАНИИ ПОЛЯ
НАПРЯЖЕНИЙ (КОМПЛЕКС ПРОГРАММ ДЛЯ ВЫЧИСЛЕНИЯ
НАПРАВЛЕНИЯ РОСТА) И НАТУРНЫЙ ЭКСПЕРИМЕНТ

(Самарский национальный исследовательский университет
имени академика С.П. Королева)

Одним из важных вопросов современной вычислительной механики разрушения является аккуратная оценка направления распространения трещины в условиях смешанного деформирования, когда в образце одновременно реализуются и нормальный отрыв, и поперечный сдвиг [1-9]. В настоящее время сформулировано три критерия разрушения: критерий максимального тангенциального напряжения, критерий минимума плотности энергии упругой деформации и деформационный критерий разрушений, основанный на максимальном значении окружной деформации [1-9]. Математическая формулировка критерия максимального тангенциального напряжения гласит, что трещина будет расти в направлении, в котором тангенциальное напряжение достигает своего максимального значения:

$$\frac{\partial \sigma_{\theta\theta}}{\partial \theta} = 0, \quad \frac{\partial^2 \sigma_{\theta\theta}}{\partial \theta^2} < 0. \quad (1)$$

В классической механике хрупкого разрушения в критерий (1) подставляют асимптотическое разложение М. Уильямса поля напряжений в окрестности вершины трещины:

$$\sigma_{ij}(r, \theta) = \sum_{m=1}^{m=2} \sum_{k=1}^{\infty} a_k^m r^{k/2-1} f_{m,ij}^{(k)}(\theta), \quad (2)$$