**УДК 004.852 + 577.112**

# ИСПОЛЬЗОВАНИЕ ГРАДИЕНТНОГО БУСТИНГА ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ ПРЕДСКАЗАНИЯ СТАБИЛЬНОСТИ ВОДОРОДНОЙ СВЯЗИ В БЕЛКЕ

**П.Н. Дружков, Н.Ю. Золотых**

Метод градиентного бустинга деревьев решений используется для предсказания стабильности водородной связи в молекуле белка. Данный подход позволяет улучшить качество предсказания по сравнению с методом [1], использующим одиночные деревья решений.

**Ключевые слова:** деревья решений, градиентный бустинг, водородная связь, белок.

Наиболее распространенным в настоящее время методом определения стабильности водородной связи молекул белка является энергетический подход. К сожалению, он не позволяет с большой точностью прогнозировать время существования связи. Для более точного предсказания можно использовать свойства локального окружения связанных атомов. Например, в [1] стабильность водородной связи предсказывается на основе 32 показателей, что позволяет существенно повысить качество по сравнению с моделью, использующей лишь энергию связи. В качестве модели в [1] используется дерево решений с алгоритмом CART [2], модифицированным с учетом специфики задачи. В настоящей работе вместо одиночного дерева используется градиентный бустинг деревьев решений (gradient boosting trees) [3, 4]. Эксперименты показали, что качество предсказания при этом улучшается.

В качестве меры стабильности водородной связи $H$ возьмем функцию $\bar{\sigma}(H, c, \Delta)$, где $c$ – конфигурация белка в нулевой момент времени; $\Delta$ – промежуток времени, для которого рассчитывается стабильность (подробности см. в [1]). Задача заключается в построении аппроксимации $\sigma$ для этой функции. В качестве предикативных переменных рассматриваются 32 показателя, среди которых есть как постоянные во времени характеристики (атомы, между которыми существует связь, аминокислотные остатки, содержащие эти атомы и т.д.), так и изменчивые кинематические показатели (расстояния, углы и т.д.). Чтобы снизить влияние температурного шума на значения кинематических характеристик, выполняется их усреднение по 50 предыдущим по времени конфигурациям белка.

Обучение модели и оценка ее качества проводилась следующим образом. Поочередно данные о водородных связях траектории каждого белка из 6 рассматриваемых [1] объявлялись тестовыми. Из показателей существующих связей оставшихся 5 траекторий формировалась выборка, на которой производилась настройка модели, причем из данных для каждого белка случайным образом выбиралось лишь 10% от общего числа присутствующих водородных связей. Эксперимент повторялся 10 раз для каждого тестового белка. Таким образом, всего было обучено 60 моделей. Такой подход позволяет оценить не только качество предсказания стабильности связи, но и независимость данного подхода от исследуемого белка.

Использовались следующие параметры алгоритма градиентного бустинга деревьев решений: функция штрафа – квадратичная; число деревьев – 1000; максимальная глубина деревьев – 3; параметр регуляризации (shrinkage) – 0,05; доля обучающей выборки, используемая на каждой итерации алгоритма – 0,6.

На тестовой выборке вычислялись следующие показатели качества модели: корень квадратичной ошибки RMSE($\sigma$) (Root Mean Square Error) и уменьшение ошибки относительно оптимальной константной модели $\sigma_0$: RBED($\sigma$, $\sigma_0$) (Relative Base Error Decrease). В таблице приведены средние значения RBED для каждого тестового набора данных для построенной модели и для метода [1], использующего одиночное дерево глубины 5. Применение градиентного бустинга позволило добиться лучших результатов по сравнению с одиночным деревом решений, а, следовательно, и с энергетической моделью. Значительное улучшение получено на данных complex.

| Белок | 1c9oA | 1e85A | 1eia | 1g9oA_1 | 1g9oA_2 | complex |
|---|---|---|---|---|---|---|
| Дерево решений [1], % | 46,92 | 59,37 | 42,6 | 50,93 | 45,29 | 37,9 |
| Градиентный бустинг, % | 50,09 | 60,7 | 44,93 | 52,84 | 47,78 | 47,65 |

Таблица. Средние значения RBED для каждого тестового набора (белка)

Работа выполнена в рамках программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы», государственный контракт № 11.519.11.4015.

1. Chikalov I., Yao P., Moshkov M., Latombe J.C. Learning probabilistic models of hydrogen bond stability from molecular dynamics simulation trajectories // Journal of Intelligent Learning Systems and Applications. – 2011. – № 3. – P. 155–170.
2. Breiman L., Friedman J., Olshen R., Stone C. Classification and Regression Trees. – Wadsworth, 1983. – 232 P.
3. Friedman J.H. Greedy function approximation: a gradient boosting machine. Technical report. – Stanford University, Dept. of Statistics. – 1999. – 22 p.
4. Дружков П.Н., Золотых Н.Ю., Половинкин А.Н. Программная реализация алгоритма градиентного бустинга деревьев решений // Вестник Нижегородского государственного университета им. Н.И. Лобачевского. – 2011. – № 1. – С. 193–200.

***Дружков Павел Николаевич*** – Нижегородский государственный университет им. Н. И. Лобачевского, студент, druzhkov.paul@gmail.com
***Золотых Николай Юрьевич*** – Нижегородский государственный университет им. Н. И. Лобачевского, кандидат физ.-мат. наук, доцент, Nikolai.Zolotykh@gmail.com