Робастные регрессионные случайные леса при малых и зашумленных обучающих данных

Л. В. Уткин[1], М. С. Ковалев[2]
Санкт-Петербургский политехнический
университет Петра Великого
Санкт-Петербург, Россия
[1]lev.utkin@gmail.com, [2]maxkovalev03@gmail.com

Ф. Коолен
Университет г. Дарем,
Дарем, Великобритания
frank.coolen@durham.ac.uk

***Аннотация*. Предложена регрессионная модель случайного леса с учетом неточности оценок деревьев решений. Неточность исходит из условий малости и зашумленности обучающих данных, что имеет место во многих приложениях. Фактически, предлагается метамодель для обучения и вычисления оптимальных весов, назначаемых деревьям решений, которые контролируют неточность, чтобы получить робастные прогнозируемые значения для случайных лесов. Неточность оценок деревьев определяется с помощью интервальных статистических моделей, например, с использованием доверительных интервалов. Веса рассчитываются путем решения стандартной задачи квадратичной оптимизации с линейными ограничениями. Численные примеры иллюстрируют предложенную робастную модель, которая обеспечивает лучшие результаты для зашумленных и малых данных по сравнению со стандартным случайным лесом.**

Ключевые слова: случайный лес; регрессия; доверительный интервал; робастная модель; интервальная модель; квадратичная оптимизация

I. Введение

В последние годы было разработано множество ансамблевых методов композиции для решения задач машинного обучения, включая классификацию и регрессию [6]. Соответствующие методы объединяют прогнозируемые значения, полученные с использованием базовых классификаторов, чтобы получить сильный классификатор с лучшими характеристиками точности. Подробный обзор многих композиционных методов представлен в работе [11], где показано, что случайный лес (СЛ) [3] можно рассматривать как одну из наиболее эффективных и простых моделей, основанных на композиции. СЛ строится с помощью множества отдельных деревьев решений, так что каждое дерево обучается на подмножествах случайно выбранных примеров и признаков. СЛ комбинирует отдельные прогнозируемые значения деревьев решений и уменьшает возможную корреляцию между деревьями, выбирая различные подмножества пространства признаков. В зависимости от решаемой задачи машинного обучения прогнозируемые значения деревьев решений могут различаться. В частности, распределения вероятностей классов вычисляются в задачах классификации. Они оцениваются путем подсчета доли примеров различных классов, которые попали в определенный лист дерева. В задачах регрессии прогнозируемые значения деревьев решений обычно представлены в форме усредненных выходных значений примеров, которые затем еще раз усредняются по всем деревьям для получения прогнозируемого значения СЛ.

Одним из способов получения более точного решения для СЛ является введение весов деревьев или подмножеств деревьев, которые можно рассматривать как дополнительные параметры обучения, и они назначаются каждому дереву в соответствии с точностью решения дерева (смотри, например, [7], [8]). Существенным препятствием для использования весового усреднения, а также простого усреднения прогнозируемых значений деревьев является то, что они предполагаются точными, в то время как мы не можем ожидать какой-либо точности, особенно на небольшом объеме обучающих данных. Чтобы обойти это препятствие или учесть его, мы предлагаем робастную интервальную регрессионную модель СЛ. Одной из основных идей, лежащих в основе предлагаемой модели, является использование весов, приписанных деревьям решений, как функции интервальных прогнозируемых значений деревьев. Веса не используются, чтобы повысить точность прогнозирования СЛ. Веса используются, чтобы получить максиминное или робастное решение о прогнозируемых значениях СЛ при условии, что выходы деревьев решений являются интервальными.

Следует отметить, что идея учета неточности прогнозируемых значений деревьев решений была предложена в некоторых работах, например, [1], [2], где новые правила расщепления рассматривались при зашумленных обучающих данных. Однако мы предлагаем совершенно другой подход, который не изменяет деревьев решений, чтобы учесть имеющуюся неточность оценок. Мы обучаем веса деревьев, что позволяет нам контролировать неточность и реализовывать робастную стратегию принятия решений. При этом предлагаемый подход также может быть применен к модифицированным деревьям решений со специальными правилами

расщепления. Этот подход только улучшает учет неточностей, используемый в правиле расщепления.

Мы также используем следующие идеи:

1. Неточность оценок каждого дерева определяется с помощью интервальных моделей, например, с использованием обычных доверительных интервалов.
2. Предложена модификация функций потерь для вычисления оптимальных весов и решения робастной задачи оптимизации.
3. Полученные задачи оптимизации являются квадратичными с линейными ограничениями.

II. Весовое усреднение в регрессионных СЛ

Задача восстановления регрессии может быть поставлена следующим образом. Имеется обучающее множество $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ из n примеров, в котором $\mathbf{x}_i$ может принадлежать произвольному множеству $\mathbf{X} \subset \mathbf{R}^m$ и представляет собой вектор m признаков; $y_i \in \mathbf{R}$ представляет собой такое наблюдаемое выходное значение модели, что $y_i = f(\mathbf{x}_i) + \varepsilon$. Здесь ε – случайный шум с нулевым математическим ожиданием и некоторой конечной дисперсией. Задача машинного обучения заключается в построении регрессионной модели f, которая минимизирует ожидаемый риск, представляемый, например, как

$$J(S) = n^{-1} \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i) \right)^2 .$$

Одной из эффективных непараметрических моделей, позволяющей строить регрессионную зависимость, является СЛ. Перед тем, как рассматривать СЛ в целом, определим как прогнозируемые значения могут быть вычислены отдельным деревом решений. Предположим, что n_l обучающих примеров $(\mathbf{x}_i, y_i)$ с индексами из множества D_l, т.е. $i \in D_l$, попадают в l-ую вершину (лист) дерева. Тогда прогнозируемое значение $z = f(\mathbf{x})$ нового примера, который попал в эту же вершину вычисляется как

$$z = f(\mathbf{x}) = n_l^{-1} \sum_{i \in D_l} y_i .$$

Другими словами, значение z есть среднее выходных значений примеров с индексами из D_l.

Вернемся к СЛ и предположим, что он состоит из T обученных деревьев. Прогнозируемое значение СЛ вычисляется как среднее прогнозируемых значений $z^{(t)}$ всех деревьев или

$$z_{\mathrm{RF}} = T^{-1} \sum_{t=1}^{T} z^{(t)} .$$

Для улучшения СЛ можно назначить веса w_t деревьям решений и вычислить следующее весовое среднее:

$$z_{\mathrm{RF}} = \sum_{t=1}^{T} z^{(t)} w_t = \mathbf{wz}.$$

Здесь $\mathbf{w} = (w_1, ..., w_T)$ и $\mathbf{z} = (z^{(1)}, ..., z^{(T)})^{\mathrm{T}}$. Веса ограничены следующим условием:

$$\sum_{t=1}^{T} w_t = \mathbf{w} \cdot \mathbf{1}^{\mathrm{T}} = 1, \; w_t \geq 0, \; t = 1, ..., T.$$

Здесь $\mathbf{1}$ – вектор, состоящий из T единиц. Тогда задача машинного обучения заключается в построении модели f, которая минимизирует ожидаемый риск

$$J(\mathbf{w}) = n^{-1} \sum_{i=1}^{n} \left(y_i - \mathbf{wz}_i \right)^2 ,$$

где $\mathbf{z}_i$ – вектор прогнозируемых значений T деревьев для обучающего примера $\mathbf{x}_i$.

Очевидно, что оценки z для каждого листа дерева не могут быть точными при малом числе обучающих данных или при шуме. Даже при большом числе примеров, нет гарантии что многие из них попадут в определенный лист, т.е. n_l может быть малым. Отсюда следует, что интервальное значение z следует определять вместо точного. Используя стандартную технику статистического оценивания, запишем доверительный интервал в виде:

$$z \pm t_{(1-a/2)} s_z / \sqrt{n_l} ,$$

где $t_{(1-a/2)}$ определяется из таблиц t-распределения в соответствии с доверительной вероятностью $100(1-\alpha)$;

$$s_z^2 = \frac{1}{n_l - 1} \sum_{i \in D_l} (y_i - z)^2 .$$

Отсюда следует, что $z^{(t)}$ для t-го дерева имеет нижнюю $z_L^{(t)}$ и верхнюю $z_U^{(t)}$ границы, которые должны учитываться в задаче вычисления. Пусть Z_t - множество значений $z^{(t)}$, полученных из $[z_L^{(t)}, z_U^{(t)}]$. Так как каждое $z^{(t)}$ зависит от определенного примера, скажем *i*-го, и от дерева, скажем *t*-го, то будем обозначать это значение как $z^{(t)}(i)$.

Один из известных подходов для работы с интервальными данными – использование пессимистической или робастной стратегии [5]. В соответствии с ней, выбирается такое значение из интервала, которое делает функцию потерь $J(\mathbf{w})$ максимальной при фиксированном $\mathbf{w}$. Выбранное оптимальное значение может быть различным для различных $\mathbf{w}$. Эта стратегия выбирает «наихудшее» значение $z^{(t)}$ из интервала $[z_L^{(t)}, z_U^{(t)}]$, $t = 1, ..., T$, обеспечивающее максимум функции потерь $J(\mathbf{w})$.

III. Обучение весового регрессионнного СЛ

Для определения оптимальных весов $\mathbf{w}$ при известных точных значениях $z^{(t)}(i)$ необходимо решить следующую задачу оптимизации:

$$J(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i - \mathbf{w}\mathbf{z}_i \right)^2 + \lambda \left\| \mathbf{w} \right\|^2 .$$

Здесь $\left\| \mathbf{w} \right\|^2$ – регуляризационное слагаемое, λ – настраиваемый параметр.

Ограничим также множество весов некоторым выпуклым подмножеством $\mathbf{W}(u)$ единичного симплекса весов для дополнительной регуляризации. Здесь u – параметр, который определяет размер подмножества $\mathbf{W}$, например, параметр интервальной модели Дирихле или параметр ε в интервальной модели ε-засорения [9], [10], если эти модели используются для образования множества $\mathbf{W}(u)$.

Для учета того, что $z^{(t)}(i)$ является интервальным и для применения робастной стратегии, запишем задачу оптимизации как

$$J(\mathbf{w}) = \max_{z^{(t)}(i) \in [z_L^{(t)}(i), z_U^{(t)}(i)]} \min_{\mathbf{w} \in \mathbf{W}(u)} \sum_{i=1}^{n} \left(y_i - \mathbf{w}\mathbf{z}_i \right)^2 + \lambda \left\| \mathbf{w} \right\|^2 .$$

К сожалению, приведенное представление не может быть применено, так как оно приводит к квадратичной задаче с квадратичными ограничениями. Поэтому предлагается заменить первое слагаемое целевой функции

$$\sum_{i=1}^{n} \left| y_i - \sum_{t=1}^{T} z^{(t)}(i) w_t \right| = \sum_{i=1}^{n} \left| \sum_{t=1}^{T} w_t \left(y_i - z^{(t)}(i) \right) \right|,$$

на

$$\sum_{t=1}^{T} w_t \sum_{i=1}^{n} \left| y_i - z^{(t)}(i) \right|.$$

Последнее преобразование не эквивалентно. Однако мы предлагаем заменить функцию риска на другую функцию, которая ограничивает ошибку не только всего СЛ, но и каждого дерева. В итоге получаем следующую задачц квадратичной оптимизации:

$$J(\mathbf{w}) = \max_{z^{(t)}(i) \in [z_L^{(t)}(i), z_U^{(t)}(i)]} \min_{\mathbf{w} \in \mathbf{W}(u)} \sum_{t=1}^{T} w_t \sum_{i=1}^{n} \left| \left(y_i - z^{(t)}(i) \right) \right| + \lambda \left\| \mathbf{w} \right\|^2 .$$

Это – прямая форма задачи оптимизации. Зафиксируем значение $z^{(t)}(i)$ и запишем двойственную задачу оптимизации для $\mathbf{w}$. Предположим, что множество $\mathbf{W}(u)$ образовано следующими линейными ограничениями:

$$a_t \le w_t \le b_t, \; t = 1, ..., T, \; \mathbf{w} \cdot \mathbf{1}^{\mathrm{T}} = 1.$$

Эти ограничения соответствуют многим задачам интервального статистического оценивания. Запишем двойственную задачу без регуляризационного слагаемого для простоты. Она совместно с оптимизацией по $z^{(t)}(i)$ имеет вид:

$$\max_{z^{(t)}(i) \in [z_L^{(t)}(i), z_U^{(t)}(i)]} \max_{h_0, h_t, g_t} \left(h_0 + \sum_{t=1}^{T} \left(h_t b_t - h_t a_t \right) \right),$$

при ограничениях $h_t, g_t \ge 0$, $t = 1, ..., T$,

$$h_0 + \sum_{t=1}^{T} \left(h_t - g_t \right) \le \sum_{i=1}^{n} \left| \left(y_i - z^{(l)}(i) \right) \right|, \; l = 1, ..., T.$$

Можно увидеть из ограничений, что максимум по $z^{(t)}(i)$ достигается, когда в правая часть ограничений максимальна. Это имеет место, когда выполняется следующее правило: если $y_i < (z_L^{(t)}(i) + z_U^{(t)}(i))$, то $z^{(t)*}(i) = z_U^{(t)}(i)$, иначе $z^{(t)*}(i) = z_L^{(t)}(i)$.

Такой же результат можно получить для задачи квадратичной оптимизации, которая учитывает регуляризационное слагаемое.

Подставляя соответствующие оптимальные значения $z^{(t)*}(i)$, зависящие от y_i, $z_L^{(t)}(i)$, $z_U^{(t)}(i)$, в прямую задачу, получаем оптимизационную задачу для вычисления оптимальных весов:

$$J(\mathbf{w}) = \min_{\mathbf{w} \in \mathbf{W}(u)} \sum_{t=1}^{T} w_t \sum_{i=1}^{n} \left| \left(y_i - z^{(t)*}(i) \right) \right| + \lambda \left\| \mathbf{w} \right\|^2 .$$

Предположим, что множество $\mathbf{W}(u)$ образовано при помощи интервальной модели ε-засорения [9] с исходными вероятностями $(T^{-1}, ..., T^{-1})$ и параметром $\varepsilon \in [0,1]$, т.е. $u = \varepsilon$. Согласно этой модели множество $\mathbf{W}(\varepsilon)$ образуется T линейными ограничениями:

$$\frac{1-\varepsilon}{T} \le w_t \le 1, \; t = 1, ..., T, \; \mathbf{w} \cdot \mathbf{1}^{\mathrm{T}} = 1.$$

Тогда задача квадратичной оптимизации имеет линейные ограничения и может быть решена с использованием стандартных процедур.

IV. Численные эксперименты

Для иллюстрации модели робастного регрессионного СЛ, сравним его с обычным СЛ, используя данные из репозитория UCI Machine Learning Repository [4] и библиотеки Python Sklearn.datasets, включая Boston house-prices ($m = 13$, $n = 506$) и Diabetes ($m = 10$, $n = 442$). Среднеквадратическая ошибка прогноза (RMSE) используется в качестве показателя качества модели, которая определяется как

$$RMSE = \sqrt{n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \left(y_i - f(\mathbf{x}_i) \right)^2}.$$

Здесь n_{test} – число тесовых примеров.

Для оценки RMSE, реализуется процедура скользящего контроля со 100 повторениями, где в каждом запуске процедуры случайным образом выбираются $n_{\text{tr}} = \gamma n$ обучающих примером и $n_{\text{test}} = (1-\gamma)n$ тестовых примеров, $\gamma \in [0,1]$ – параметр, который используется в экспериментах. Задача квадратичной оптимизации решается при выборе параметра ε множества $\mathbf{W}(\varepsilon)$, дающего наибольшую точность. Различные значения параметра λ анализируются с выбором наилучшего с точки зрения точности. СЛ состоит из 100 деревьев решений.

Для ряда экспериментов, к каждому признаку тестовых примеров добавляется шум $z_{ij} \sim N(0, \eta\sigma_j)$, имеющий нормальное распределение, для анализа робастности предлагаемого алгоритма, где η – параметр шума, σ_j – среднеквадратическое отклонение *j*-го признака.

Два графика на рис. 1 иллюстрируют зависимости RMSE обычного СЛ (штрихпунктирная линия) и предлагаемой модели (сплошная линия) от размера обучающей выборки для данных Diabetes. Параметр для построения доверительного интервала равен α=0.002. Аналогичные графики иллюстрируют такую же зависимость при условии дополнительного шума с параметром η=0.5 (рис. 2). Из рисунков можно увидеть, что предлагаемая модель более точная по отношению к обычному СЛ.

Аналогичные результаты показаны в таблице для данных Boston, где первый столбец содержит долю обучающих примеров γ, последующие три столбца показывают RMSE для обычного СЛ и предлагаемого робастного СЛ при различных α (0.2 и 0.05) без шума, следующие три столбца показывают RMSE с шумом с параметром η=0.5. Из таблицы видно, что предлагаемый робастный СЛ снова превосходит оригинальный СЛ.

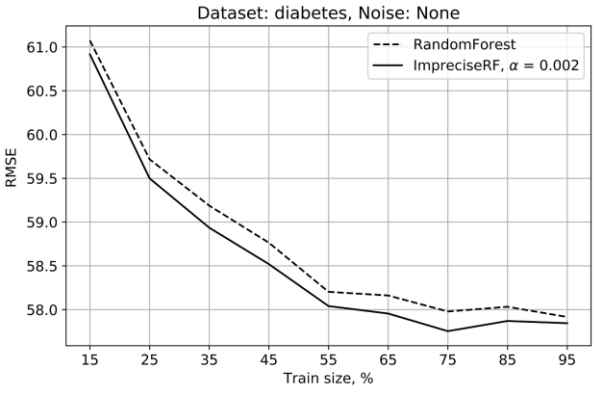

Рис. 1. RMSE для Diabetes без шума

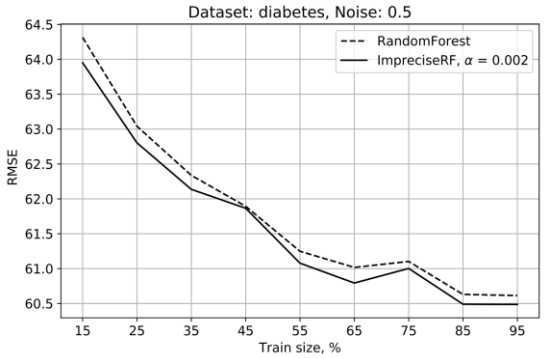

Рис. 2. RMSE для Diabetes с шумом, имеющим параметр η =0.5

ТАБЛИЦА I RMSE ДЛЯ ДАННЫХ BOSTON

γ	СЛ	Робастный СЛ		СЛ	Робастный СЛ	
	без шума			*с шумом (η =0.5)*		
		α =0.2	*α =0.05*		*α =0.2*	*α =0.05*
15	4.787	4.742	4.742	6.025	6.015	6.020
25	4.408	4.380	4.380	5.746	5.727	5.725
35	4.188	4.165	4.167	5.639	5.615	5.612
45	4.036	4.010	4.011	5.549	5.532	5.53
55	3.910	3.891	3.891	5.450	5.428	5.430
65	3.825	3.793	3.789	5.435	5.430	5.426
75	3.764	3.734	3.736	5.320	5.312	5.314
85	3.690	3.660	3.660	5.311	5.305	5.308
95	3.609	3.588	3.592	5.323	5.308	5.305

V. ЗАКЛЮЧЕНИЕ

Одна из реализаций СДО была представлена в работе. Главная особенность системы заключается в том, что она использует методы «неглубокого» обучения, которые значительно упрощают процедуру обучения. Важно отметить, что метод хорд в сочетании с сегментацией изображений легких на основе пороговых значений плотностей позволяет получить интересные результаты, которые сравнимы с другими существующими современными подходами, применяемыми для построения СДО.

Предложенную реализацию СДО можно рассматривать как первую попытку использования метода хорд. Более того, мы рассмотрели только использование случайных лесов для классификации новообразований. Однако мы предполагаем, что использование более сложных классификаторов, например глубокого леса [15], может значительно повысить эффективность системы. Это является интересным направлением для дальнейших исследований. Другим направлением для дальнейших исследований является совместное использование предлагаемого представления признаков со стандартным представлением изображения, где классификация изображений выполняется путем применения СНС или снова глубокого леса с элементами обработки исходных изображений большой размерности.

Список литературы

[1] J. Abellan, C.J. Mantas, and J.G. Castellano. A random forest approach using imprecise probabilities. Knowledge-Based Systems, 134:72–84, 2017.

[2] J. Abellan, C.J. Mantas, and S. Moral-Garcia, J.G. Castellano. Increasing diversity in random forest learning algorithm via imprecise probabilities. Expert Systems With Applications, 97:228–243, 2018.

[3] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[4] M. Lichman. UCI machine learning repository, 2013.

[5] C.P. Robert. The Bayesian Choice. Springer, New York, 1994.

[6] L. Rokach. Ensemble-based classifiers. Artificial Intelligence Review, 33(1-2):1–39, 2010.

[7] L.V. Utkin, M.S. Kovalev, and A.A. Meldo. A deep forest classifier with weights of class probability distribution subsets. Knowledge-Based Systems, 173:15–27, 2019.

[8] L.V. Utkin and M.A. Ryabinin. A Siamese deep forest. Knowledge-Based Systems, 139:13–22, 2018.

[9] P. Walley. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London, 1991.

[10] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. Journal of the Royal Statistical Society, Series B, 58:3–57, 1996.

[11] Z.-H. Zhou. Ensemble Methods: Foundations and Algorithms. CRC Press, Boca Raton, 2012.