Метод обучения деревьев решений с нелинейными разделителями *

Д. А. Девяткин, О. Г. Григорьев

Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия

Аннотация. Деревья решений с одномерными разделителями, применяемые при обработке разреженных данных большой размерности, характеризуются низкой вычислительной эффективностью. Деревья решений с многомерными разделителями обладают большей выразительной способностью при классификации данных, но переобучаются на небольших выборках. В статье предложен метод обучения деревьев с многомерными нелинейными разделителями, который повышает точность классификации на наборах изображений и текстов. Это достигается за счёт совместной оптимизации расстояния от объектов обучающей выборки до разделяющей поверхности и критерия неоднородности данных при построении каждого узла дерева. Эффективность метода подтверждается результатами тестов.

Ключевые слова: дерево решений, нелинейный разделитель, метод опорных векторов, масштабирование переменных невязки, случайные леса деревьев решений.

DOI 10.14357/20718594220308

Введение

Бинарные деревья решений и их композиции, в том числе случайные леса, применяются во многих системах интеллектуальной обработки данных [1-3]. Деревья решений состоят из узлов, каждый из которых делит анализируемые данные на два подмножества с помощью некоторого разделителя. Разделитель, как правило, представляет собой простое решающее правило, в котором значение определенного признака сравнивается с заданным порогом. Для обучения разделителя оптимизируется некоторый критерий неоднородности данных, такой как Джини (Gini impurity), прирост информации или информационная энтропия. Для поиска параметров, минимизирующих этот критерий, решается задача дискретной оптимизации. В деревьях решений с одномерными разделителями эта оптимизация тривиальна, так как имеется возможность протестировать все возможные пороговые значения для каждого признака обрабатываемых данных. Однако подобный тип деревьев решений имеет ограниченную применимость при анализе многомерных разреженных данных, которые часто используются в задачах обработки текста и изображений.

Одним из подходов к решению этой проблемы является обучение деревьев с более сложными, многомерными разделителями, например, на основе гиперплоскостей. К сожалению, подобные деревья решений имеют низкую обобщающую способность, поэтому они применимы только в составе случайных линейных композиций, построенных методами бэггинга, со-бэггинга или случайного леса. Кроме того, известные подходы к построению таких

* Исследование выполнено в рамках научной программы Национального центра физики и математики (направление №9 «Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах»).

✉ *Девяткин Дмитрий Алексеевич. E-mail: devyatkin@isa.ru*

деревьев имеют низкую вычислительную эффективность и большое количество гиперпараметров. Тем не менее, даже при использовании подобных композиций могут наблюдаться эффекты, связанные с переобучением.

В настоящей работе предложен алгоритм для построения деревьев решений с нелинейными (полиномиальными и Гауссовскими) разделителями. Построение этих разделителей осуществляется путем совместной оптимизации расстояния от объектов обучающей выборки до разделяющей поверхности в пространстве признаков (далее – *отступа*), и заданного критерия неоднородности данных (impurity), что позволяет улучшить обобщающую способность деревьев решений.

1. Деревья решений с многомерными разделителями

Переобучение является одной из основных проблем, связанных с усложнением разделителей в узлах деревьев решений. В статье [4] предложена оценка обобщающей способности деревьев на задаче бинарной классификации. В основе этой оценки лежит понятие «эффективного числа листьев» дерева. Чем ближе эмпирическое распределение обучающих данных по листьям к равномерному, тем ближе это число к фактическому. Показано, что верхняя граница ошибки классификации положительно зависит от эффективного числа листьев и VC-размерности (размерности Вапника-Червоненкиса) алгоритма, применяемого в узлах деревьев [5]. Необходимо отметить, что структура дерева решений, в том числе распределение данных по листьям, зависит от выбора критерия неоднородности данных, оптимизируемого при обучении влияющего на обобщающую способность [6]. В статьях [7, 8] представлена оценка обобщающей способности деревьев решений, согласно которой увеличение отступа в узлах дерева позволяет снизить переобучение.

В работе [1] анализируются случайные леса, показано, что усложнение базовых алгоритмов не приводит к переобучению при условии сохранения низкой корреляции ошибок, которые допускают отдельные деревья. Внесение дополнительной вариативности в структуру деревьев снижает корреляцию и, как следствие, улучшает обобщающую способность композиции. Одним из способов внесения такой вариативности является применение жадных алгоритмов для обучения отдельных деревьев композиции.

Метод опорных векторов (Support Vector Machines, SVM) широко применяется для обучения регуляризованных линейных и нелинейных классификаторов. Он также используется для построения разделителей в деревьях решений. Одна из первых попыток использования SVM представлена в статье [9]. В ней предлагается метод построения разделителей в деревьях с фиксированной структурой. Однако его использование при обучении случайных лесов нежелательно, так как фиксированная структура может приводить к высокой корреляции между ошибками отдельных деревьев композиции. Кроме того, предложенная функция потерь не является выпуклой, что ограничивает применимость существующих вычислительно-эффективных методов оптимизации для обучения.

В статье [10] предлагается вариант случайного леса деревьев решений с линейными разделителями. Для поиска разделяющих гиперплоскостей используется гребневая регрессия (ridge regression). Метод позволяет обучать деревья только для задач бинарной классификации.

В работе [11] отмечается, что критерии неоднородности не отражают пространственного распределения данных. Между тем, учет подобного распределения позволил бы добиться упрощения структуры деревьев без увеличения ошибки на обучающем наборе данных, что привело бы к повышению обобщающей способности.

В [12] предложен метод, в котором для объектов каждого класса строится кластеризующая гиперплоскость. Биссектрисы углов, которые образуют эти гиперплоскости, используют в качестве разделителей. Результаты экспериментальных исследований показывают, что метод позволяет получать компактные деревья решений, имеющие относительно высокую обобщающую способность по сравнению с деревьями, построенными с помощью других подходов.

В статьях [7, 8] представлены теоретические оценки обобщающей способности деревьев решений с нелинейными разделителями, показывающие, что переобучение может быть снижено путем максимизации отступа в узлах и

минимизации следа матрицы Грама [13], применяемой для задания скалярного произведения в нелинейном пространстве. Представлен метод обучения деревьев решений с нелинейными разделителями для задач классификации разреженных данных большой размерности.

Перечисленные работы сфокусированы на оптимизации отступа в узлах деревьев или критерия неоднородности, в то время как оба эти фактора влияют на переобучение [4, 11]. В статье [14] предложены «CO2-деревья» с линейными разделителями. При их построении обучение формулируется в виде задачи построения классификатора структур со скрытыми переменными [15]. Осуществляется оптимизация выпукло-вогнутой функции потерь, которая может быть достаточно эффективно осуществлена с помощью градиентного метода, предложенного в [16]. Функция является верхней гранью эмпирического риска, причем ее поведение зависит от гиперпараметров и масштаба признаков, что ограничивает применимость метода.

В статье [17] представлены составные деревья решений, внутренние узлы которых являются одномерными, а разделители в терминальных узлах выбираются из множества алгоритмов классификации различной сложности. В работе получена оценка обобщающей способности таких деревьев, которая используется для выбора алгоритмов классификации и каждого терминального узла. Предложены случайные составные деревья решений, которые могут применяться в качестве базовых алгоритмов классификации в случайных лесах деревьев решений (Random Composite Forest, RCF).

Еще одним перспективным направлением является сквозное (end-to-end) обучение деревьев решений с линейными разделителями и их лесов. В статье [18] представлен подход к обучению вероятностных деревьев решений с помощью EM-алгоритма (EM – expectation-maximization). Экспериментальные исследования с использованием размеченных наборов изображений показали, что можно формировать более сложные разбиения, чем детерминированные деревья с линейными разделителями, но EM-алгоритм достаточно медленно сходится на больших наборах данных.

В исследовании [19] предлагается использовать метод обратного распространения ошибки, который обычно используется для обучения многослойных нейронных сетей [21] и обучения деревьев решений с линейными разделителями и фиксированной структурой. Вместе с тем использование фиксированной структуры деревьев может привести к получению избыточно сложных алгоритмов с большим количеством параметров и, как следствие, к ухудшению обобщающей способности.

В статье [22] отмечается, что алгоритмы для построения деревьев решений с линейными разделителями имеют более высокую вычислительную сложность, чем алгоритмы обучения деревьев с одномерными разделителями. Кроме того, результаты обучения часто зависят от инициализации параметров разделяющих гиперплоскостей. В статье предложен метод построения деревьев решений с линейными разделителями (WODT – взвешенное дерево решений с линейными разделителями), в котором применяется гладкая функция потерь. Метод обучения состоит в назначении весов для объектов обучающей выборки и обучении узла дерева решения, путем оптимизации взвешенной информационной энтропии распределения объектов по дочерним узлам. Стоит отметить, что функция потерь WODT является невыпуклой, следовательно, глобальный оптимум не может быть гарантированно найден градиентными методами. В то же время результаты экспериментальных исследований метода показывают конкурентоспособные оценки точности классификации для многих наборов данных.

В статье [23] предлагается алгоритм, который улучшает заданное дерево решений и создает новое дерево с такой же или с сокращенной структурой, но с новыми значениями параметров разделителей, что позволяет повысить точность и полноту анализа данных. Гибридное дерево решений предлагается в [24]. Задача этого исследования – сокращение количества примеров, для разделения которых необходимо применение узлов на основе метода опорных векторов. Для решения выполняется аппроксимация границы между классами, сформированной методом опорных векторов с помощью деревьев решений. В итоге строится гибридное дерево решений, которое имеет как одномерные узлы, так и узлы с линейными разделителями.

В настоящей работе предложен метод обучения деревьев решений с *нелинейными разделителями*, то есть деревьев, в которых в каче-

стве многомерных разделителей в узлах используются нелинейные двоичные классификаторы. Аналогично методу CO2 Forest [14], обучение разделителей в узлах деревьев осуществляется с помощью подходов, применяемых для решения задач классификации структур, однако, в предлагаемом методе эта задача сформулирована в явном виде. Аналогично WODT [22] задаются веса объектов обучающей выборки и минимизируется критерий неоднородности данных, однако в предложенном методе этот критерий оптимизируется совместно с отступом. Построение узлов деревьев решений сводится к решению задачи обучения SVM с масштабированием переменных невязки, то есть выпуклой оптимизации с ограничениями-неравенствами. Для обеспечения высокой скорости оптимизации предложено отказаться от поиска глобального оптимума критерия неоднородности данных в каждом узле. Вместо этого реализован алгоритм поиска квазиоптимальных решений.

2. Метод обучения деревьев решений с нелинейными разделителями

Для построения деревьев решений используется стандартный рекурсивный алгоритм (Алгоритм 1.). На каждом его шаге строится узел дерева, который разделяет обучающие данные, затем эта процедура рекурсивно повторяется для «левого» и «правого» подмножеств обучающих данных, пока не будет достигнута заданная глубина дерева.

Пусть P_X – распределение объектов, P_Y – распределение их меток, P_{XY} – совместное распределение анализируемых данных. Рассмотрим построение разделителя как задачу обучения двоичного классификатора на наборе из m объектов

$$D_m = \{< x_i, y_i > | i = 1, .., m; x_i \in X_m, y_i \in Y_m\}, \quad X_m \sim P_X, Y_m \sim P_Y, D_m \sim P_{XY}$$

с метками классов y_i, которые выбираются из множества U. Пусть этот двоичный классифи-

Алгоритм 1. Обучение дерева с линейными и нелинейными разделителями

Вход: набор данных D_m, параметр регуляризации C, J - порог выбора способа распределения классов по поддеревьям (точный/жадный), K – количество запусков жадной процедуры распределения классов по поддеревьям.

1: **вызвать** *BuildTree*(D_m)
2: BuildTree(D):
3: **Если** D содержит объекты одного класса:
4: **Выход**
5: **Иначе:**
6: **Если** $|U| < J$:
7: s:=*all_distributions(Y,H={+1,-1})*
8: s_{best}:=*sort(Impurity(s))[rnd(1..N)]* \\$U \to H$
9: **Иначе**
10: s_{best}:=*greedy_find_best_impurity(D)*\\ $U \to H$
11: $L^*_1, \ldots, L^*_m := L(h_i, -h_i), h_i \in H_{best}$
12: $w^*, \varepsilon^*_1, \ldots, \varepsilon^*_m := optimize\_node(C, X, L^*_1, \ldots, L^*_m$\\Решить задачу обучения SVM с масштабированными переменными невязки $\frac{\varepsilon^*_1}{L^*_1}, \ldots, \frac{\varepsilon^*_m}{L^*_m}$
13: $D_l := D[classify(D, w^* \geq 0)]$
14: $D_l := D[classify(D, w^* < 0)]$
15: **вызвать** *BuildTree(D_l)*
16: **вызвать** *BuildTree(D_r)*

катор распределяет объекты по поддеревьям; $H = \{-1, +1\}$ – метки этих поддеревьев. На шагах алгоритма 6-10 устанавливается целевое распределение классов в поддеревьях. Некоторые критерии могут разделять объекты одного класса по разным поддеревьям, но в предложенном методе эта особенность игнорируется в угоду скорости обучения. Если количество классов $|U|$ превышает пороговое значение J, являющееся гиперпараметром алгоритма, то перебираются все возможные распределения классов по поддеревьям и вычисляются значения соответствующего критерия неоднородности (шаги 7–8). Затем, полученный список вариантов распределений сортируется по критерию неоднородности, и целевое распределение c_s выбирается случайным образом среди первых N элементов отсортированного списка. Если количество классов $|U|$ больше, чем J, то для выявления целевого распределения классов по поддеревьям применяется жадная процедура (шаг 10). Процедура начинается с генерации случайного распределения классов по поддеревьям. Затем сгенерированное распределение итеративно изменяется и сохраняются модификации, улучшающие критерий неоднородности данных. Эта процедура повторяется K раз, затем выбирается наилучшее отображение $c_s: U \to H$. В результате в процесс обучения узла дерева решений добавляется рандомизация. Затем алгоритм использует полученное распределение c_s для определения поддерева для каждого объекта из обучающего набора данных $D_m: H_{best} = c_s(Y_m)$.

Следующие шаги алгоритма (шаги 11-12) состоят в обучении разделителя в узле дерева решений. В процессе обучения одновременно оптимизируется отступ в узле и критерий неоднородности. Для вычислительно-эффективного построения разделителей, аналогично методам обучения SVM классификации структур, задается непрерывная гладкая функция потерь, которая отражает зависимость между параметрами разделителя и значением критерия неоднородности данных. Обучение разделителя на наборе данных из m обучающих примеров производится путем решения следующей задачи оптимизации:

$$w^*, \xi^* = \operatorname{argmin}_{w,\xi}\left(\frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i\right) \qquad (1)$$

при ограничениях:

$$\forall i = 1..m, w^T x_i h_i \geq 1 - \frac{\xi_i}{L(h_i, -h_i)},$$

где w – параметры разделяющей гиперплоскости; $\xi_1, .., \xi_m$– переменные невязки; w^*– оптимальные значения параметров разделяющей гиперплоскости; ξ^*– оптимальные значения переменных невязки; C – параметр регуляризации; $L(h_i, -h_i)$отражает прирост критерия неоднородности данных в случае отнесения объекта i к некорректному поддереву – h_i вместо целевого поддерева h_i (Рис. 1).

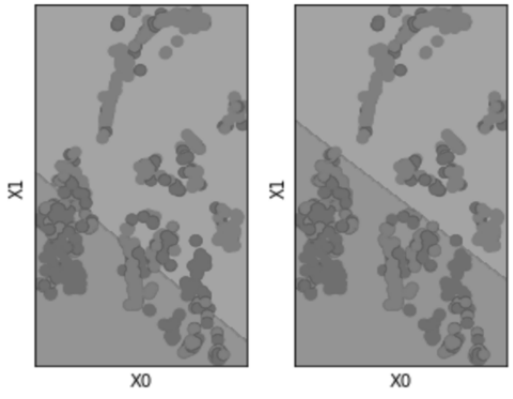
Если к задаче (1) применить условия Каруша-Куна-Таккера, то получится двойственная задача оптимизации:

$$a^* = argmax_a\left(-\frac{1}{2}\sum_{i=1..m}\sum_{j=1..m} a_i a_j K(x_i, x_j) + \sum_{i=1..m} a_i\right) \qquad (2)$$

при ограничениях

$$\sum_{i=1..m} \frac{a_i}{L(h_i, -h_i)} \leq \frac{C}{m},$$

где a_i – вес примера i из обучающей выборки (отличный от нуля для опорных векторов), а $K(x_i, x_j)$ – положительно определенное ядро [13]. В отличие от классификации структур, эта задача эффективно решается в явном виде, поскольку классов всего два (два поддерева). Гиперпараметр регуляризации C должен быть подобран эмпирически для каждого набора данных.



а) б)

Рис. 1. Влияние масштабирования переменных невязки $\xi_1, .., \xi_m$ на результаты построения разделяющей гиперплоскости:
а – без масштабирования, б – с масштабированием

Набор данных – Titanic,
критерий – неоднородность Джини

Шаги 13–16 алгоритма направлены на классификацию всех обучающих данных по поддеревьям с помощью построенного разделителя и рекурсивного обучения новых узлов дерева разделению обучающих данных поддеревьев: D_l и D_r.

Наконец, на основе деревьев, обученных с помощью представленного алгоритма с использованием метода [1], формируется случайный лес.

Покажем, как решение задач (1) или (2) позволяет оптимизировать критерий неоднородности данных. Пусть целевое значение критерия неоднородности данных достигается, если эмпирические оценки вероятностей отнесения объекта обучающей выборки к левому и правому поддеревьям равны соответственно P_L и P_R, а эмпирические оценки вероятностей классов в левом и правом поддеревьях равны: $\boldsymbol{p_L} = \{p_{L_1}, .., p_{L_{|U|}}\}$ и $\boldsymbol{p_R} = \{p_{R_1}, .., p_{R_{|U|}}\}$. Зададим

$$s_{best} = argmin\left(P_L g(\boldsymbol{p_L}) + P_R g(\boldsymbol{p_R})\right), \qquad (3)$$

где s_{best} – целевое значение критерия неоднородности данных (в случае, если данные могут быть разделены без ошибок согласно заданному отображению c_s); $g(\boldsymbol{p})$ – неоднородность Джини, информационная энтропия, или другой критерий.

Пусть w_{best}– параметры (возможно, недостижимые) разделяющей гиперплоскости, которая соответствует значению критерия неоднородности s_{best}. Предположим, что полученный разделитель относит некоторый элемент к поддереву h^*вместо h. Измененные оценки вероятностей (по сравнению с оценками для лучшего разбиения $P_L, P_R, \boldsymbol{p_L}, \boldsymbol{p_r}$) можно обозначить $P_L^*, P_R^*, \boldsymbol{p_L^*}, \boldsymbol{p_R^*}$и прирост критерия неоднородности данных равен:

$$L(h^*, h) = P_L^* g(\boldsymbol{p_L^*}) + P_R^* g(\boldsymbol{p_R^*}) - P_L g(\boldsymbol{p_L}) - P_R g(\boldsymbol{p_R}) \qquad (4)$$

Зададим $\widehat{R}(w) = \frac{1}{m}\sum_{i=1}^{m} L\left(h_i, sign(x_i^T w)\right)$ – эмпирический риск на обучающей выборке. В работе [15] приводятся доказательства того, что решение задач (1) или (2) минимизирует эмпирический риск $\widehat{R}(w)$. Стоит отметить, что эмпирический риск $\widehat{R}(w)$ не является тождественным критерию неоднородности данных. Эмпирический риск может быть переформулирован как

$$\widehat{R}(p_{err}) = \sum_{u \in U} p_{err}(u) L(c_s(u), -c_s(u)),$$

где $p_{err}(u)$ возвращает долю некорректно распределенных объектов класса u (тех, что по итогам классификации распределяются по поддеревьям не в соответствии с c_s).

Все рассматриваемые критерии неоднородности данных (Джини, прирост информации, информационная энтропия и др.) являются выпуклыми функциями от вероятностей классов в поддеревьях. Можно преобразовать $L(c_s(u), -c_s(u))$ в функцию потерь от изменений вероятностей классов по сравнению с лучшим разбиением c_s (см. выражение (4)). Функция $L(h_i, sign(x_i^T w))$ возвращает значение частной производной критерия неоднородности данных в точке, соответствующей критерию s_{best}. График функции $\widehat{R}$ является касательной к графику критерия неоднородности в точке s_{best}, которой соответствуют параметры разделяющей гиперплоскости w_{best}. При минимизации (1) производится спуск по этой касательной настолько близко к точке, соответствующей s_{best} и w_{best}, насколько позволяют распределение классифицируемых объектов в признаковом пространстве и выбранная функция ядра (Рис. 2).

Гарантируется, что для заданного отображения классов на поддеревья, разделяющая гиперплоскость будет построена так, чтобы минимизировать $\widehat{R}(w)$ (Рис. 1). Теоретически для минимизации критерия неоднородности необходимо рассмотреть все отображения $U \to H$. Однако было решено отказаться от достижения глобального оптимума, ввиду того, что сам алгоритм построения деревьев является жадным. Такой подход позволяет дополнительно рандомизировать структуру деревьев, что положительно сказывается на обобщающей способности случайных лесов, формируемых из них.

3. Результаты экспериментального исследования методов построения деревьев решений

Экспериментальные исследования проводились на размеченных наборах данных Titanic, CIFAR-10 [25] и «Youtube Channels» [26], а также на четырех выборках из коллекции UCI: SatImage, USPS, Letter, MNIST [27]. Наборы данных из UCI применяются для обучения и оценки методов распознавания изображений.

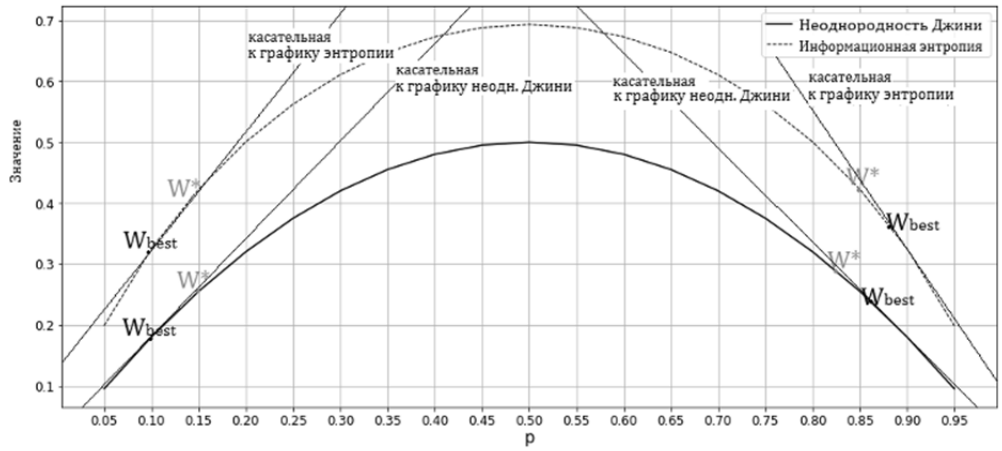


Рис. 2. Критерии построения разделителей и эмпирический риск $\hat{R}(w)$

Точке w^* соответствуют параметры разбиения, достижимые на практике на анализируемых данных

MNIST и USPS содержат изображения рукописных цифр, SatImage – фотографии различных видов земной поверхности, а Letter – латинские буквы. Набор CIFAR-10 содержит изображения, относящиеся к одному из 10 классов (самолет, лошадь, птица и др.). Youtube Channels – набор данных, содержащий психолингвистические маркеры текстов дискуссий трех категорий российских Youtube-каналов: политических проправительственных, политических оппозиционных, аполитичных. В качестве интервала времени для сбора дискуссий был выбран период с 30 апреля 2020 г. по 30 апреля 2021 г. Набор содержит маркеры комментариев к 4807 видео: 2629 видео с политическими дискуссиями (5 млн сообщений) и 2178 аполитичных видео (1,2 млн сообщений).

Исследовались случайные леса, деревья которых построены с помощью методов C4.5 (Random Forest), CO2 (CO2 Forest), WODT, а также подхода, предложенного в настоящей работе (Kernel Forest). Для подбора гиперпараметров использовалась статистическая процедура перекрестного скользящего контроля. Оценка точности (accuracy) классификации проводилась на отложенной выборке. Для предложенного подхода подбирались следующие гиперпараметры: количество деревьев *T={30,100,300}*, параметр регуляризации при построении разделителей *C={100, 1000, 3000, 5000}*, максимальная глубина дерева *n={3,4,5,6,7}*, доля признаков, которые необходимо учитывать в каждом узле *f={0.08,0.1,0.2,0.3,0.4,0.5}*, а также параметры ядра *gamma={10,100}* для Гауссовского ядра и *degree=3* для полиномиального ядра.

В Табл. 1 представлены результаты экспериментального исследования, где полужирным шрифтом выделены наилучшие результаты методов на исследуемых наборах данных. Видно, что Kernel Forest обеспечивает лучшую точность на всех выборках, кроме Letter, при использовании разделителей с Гауссовским ядром. Отметим также, что наилучшие результаты классификации достигаются при значении гиперпараметра регуляризации *C>1000*, что приводит к увеличению времени оптимизации.

Табл. 1. Результаты экспериментального исследования точности методов классификации на основе деревьев решений

Набор данных/ Метод	**MNIST**	**USPS**	**Letter**	**SatImage**	**Cifar-10**	**Youtube Channels**	**Titanic**
Random Forest	0.972	0.936	0.963	0.911	0.501	0.938	0.810
CO2 Forest	0.981	0.945	**0.982**	0.911	-	0.821	0.816
WODT	0.943	0.905	0.879	0.876	-	0.914	0.804
Kernel Forest	**0.991**	**0.946**	0.975	**0.918**	**0.581**	**0.944**	**0.820**

Заключение

Предложенная в настоящей работе совместная оптимизация отступа и критерия неоднородности данных позволяет улучшить обобщающую способность деревьев, что приводит к повышению точности классификации на тестовой выборке по сравнению с имеющимися аналогами. Дальнейшие исследования будут направлены на изучение случайных лесов деревьев решений с нелинейными разделителями в качестве компонентов алгоритмов, которые позволят выявлять сложные неявные зависимости между классифицируемыми объектами и их отдельными признаками. Кроме этого необходимы дополнительные исследования, направленные на создание быстрых алгоритмов построения деревьев решений с нелинейными разделителями

Литература

1. Breiman L. et al. Classification and regression trees. Routledge. 2017.
2. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. P. 785-794.
3. Breiman L. Random forests //Machine learning. 2001. Т. 45. №. 1. P. 5-32.
4. Golea M. et al. Generalization in decision trees and DNF: Does size matter? //Advances in Neural Information Processing Systems. 1997. Т. 10.
5. Vapnik V. N. An overview of statistical learning theory //IEEE transactions on neural networks. 1999. Т. 10. №. 5. P. 988-999.
6. Breiman L. Some properties of splitting criteria //Machine learning. 1996. Т. 24. №. 1. P. 41-47.
7. Liu W., Tsang I. W. Sparse perceptron decision tree for millions of dimensions //Thirtieth AAAI Conference on Artificial Intelligence. 2016.
8. Liu W., Tsang I. W. Making decision trees feasible in ultrahigh feature and label dimensions //Journal of Machine Learning Research. 2017.
9. Bennett K. P., Blue J. A. A support vector machine approach to decision trees //1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227). IEEE. 1998. Т. 3. P. 2396-2401.
10. Menze B. H. et al. On oblique random forests //Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg. 2011. P. 453-469.
11. Tibshirani R., Hastie T. Margin Trees for High-dimensional Classification //Journal of Machine Learning Research. 2007. Т. 8. №. 3.
12. Manwani N., Sastry P. S. Geometric decision tree //IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2011. Т. 42. №. 1. P. 181-192.
13. Hofmann T., Schölkopf B., Smola A. J. Kernel methods in machine learning //The annals of statistics. 2008. Т. 36. №. 3. P. 1171-1220.
14. Norouzi M. et al. Co2 forest: Improved random forest by continuous optimization of oblique splits //arXiv preprint arXiv:1506.06155. 2015.
15. Tsochantaridis I. et al. Large margin methods for structured and interdependent output variables //Journal of machine learning research. 2005. Т. 6. №. 9.
16. Yuille A. L., Rangarajan A. The concave-convex procedure //Neural computation. 2003. Т. 15. №. 4. P. 915-936.
17. DeSalvo G., Mohri M. Random composite forests //Proceedings of the AAAI Conference on Artificial Intelligence. 2016. Т. 30. №. 1.
18. Hehn T. M., Kooij J. F. P., Hamprecht F. A. End-to-end learning of decision trees and forests //International Journal of Computer Vision. 2020. Т. 128. №. 4. P. 997-1011.
19. Irsoy O., Alpaydin E. Autoencoder trees //Asian conference on machine learning. PMLR. 2016. P. 378-390.
20. Chai Z., Zhao C. Multiclass oblique random forests with dual-incremental learning capacity //IEEE transactions on neural networks and learning systems. 2020. Т. 31. №. 12. P. 5192-5203.
21. Hecht-Nielsen R. Theory of the backpropagation neural network //Neural networks for perception. Academic Press. 1992. P. 65-93.
22. Yang B. B., Shen S. Q., Gao W. Weighted oblique decision trees //Proceedings of the AAAI Conference on Artificial Intelligence. 2019. Т. 33. №. 01. P. 5621-5627.
23. Carreira-Perpinán M. A., Tavallali P. Alternating optimization of decision trees, with application to learning sparse oblique trees //Advances in neural information processing systems. 2018. Т. 31.
24. Kumar M. A., Gopal M. A hybrid SVM based decision tree //Pattern Recognition. 2010. Т. 43. №. 12. P. 3977-3987.
25. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images //Master's thesis, University of Tront. 2009.
26. Набор данных «Youtube channels dataset». URL: http://keen.isa.ru/youtube (дата обращения: 14.07.2022).
27. Blake C. UCI repository of machine learning databases. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html (дата обращения: 14.07.2022).

Девяткин Дмитрий Алексеевич. Научный сотрудник. Федеральный исследовательский центр «Информатика и управление» РАН. Области исследований: машинное обучение, анализ больших массивов данных, компьютерный анализ естественного языка, методы извлечения информации из текстов, методы оценки эффективности научной деятельности. E-mail: devyatkin@isa.ru

Григорьев Олег Георгиевич. Доктор технических наук. Главный научный сотрудник. Федеральный исследовательский центр «Информатика и управление» РАН. Области исследований: машинное обучение, анализ больших массивов данных, обработка естественного языка, извлечение знаний. E-mail: oleggpolikvart@yandex.ru

Method for Training Decision Trees with Non-Linear Splitters

D. A. Devyatkin, O. G. Grigoriev

Federal Research Center “Computer Science and Control” of Russian Academy of Science, Moscow, Russia

Abstract. Univariate decision trees, used in the processing of sparse large dimentional data, have low computational efficiency. Multivariate decision trees are more expressive when classifying data, but overfit on small datasets. The paper proposes a method for learning trees with multidimensional non-linear splitters, which improves the accuracy of classification on sets of images and texts. This is achieved by jointly optimizing the distance from the objects of the training dataset to the separating hyperplane and the data impurity criterion when building each node of the tree. Test results confirm the effectiveness of the method.

Keywords: decision trees, kernel splits, kernel trees, slack re-scaling, random forests.

DOI 10.14357/20718594220308

References

1. Breiman L. et al. Classification and regression trees. – Routledge, 2017.
2. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – pp. 785-794.
3. Breiman L. Random forests //Machine learning. – 2001. – Vol. 45. – No. 1. – pp. 5-32.
4. Golea M. et al. Generalization in decision trees and DNF: Does size matter? //Advances in Neural Information Processing Systems. – 1997. – Vol. 10.
5. Vapnik V. N. An overview of statistical learning theory //IEEE transactions on neural networks. – 1999. – Vol. 10. – No. 5. – pp. 988-999.
6. Breiman L. Some properties of splitting criteria //Machine learning. – 1996. – Vol. 24. – No. 1. – pp. 41-47.
7. Liu W., Tsang I. W. Sparse perceptron decision tree for millions of dimensions //Thirtieth AAAI Conference on Artificial Intelligence. – 2016.
8. Liu W., Tsang I. W. Making decision trees feasible in ultrahigh feature and label dimensions //Journal of Machine Learning Research. – 2017.
9. Bennett K. P., Blue J. A. A support vector machine approach to decision trees //1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227). – IEEE, 1998. – Vol. 3. – pp. 2396-2401.
10. Menze B. H. et al. On oblique random forests //Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – Springer, Berlin, Heidelberg, 2011. – pp. 453-469.
11. Tibshirani R., Hastie T. Margin Trees for High-dimensional Classification //Journal of Machine Learning Research. – 2007. – Vol. 8. – No. 3.
12. Manwani N., Sastry P. S. Geometric decision tree //IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). – 2011. – Vol. 42. – No. 1. – pp. 181-192.
13. Hofmann T., Schölkopf B., Smola A. J. Kernel methods in machine learning //The annals of statistics. – 2008. – Т. 36. – №. 3. – С. 1171-1220
14. Norouzi M. et al. Co2 forest: Improved random forest by continuous optimization of oblique splits //arXiv preprint arXiv:1506.06155. – 2015.
15. Tsochantaridis I. et al. Large margin methods for structured and interdependent output variables //Journal of machine learning research. – 2005. – Vol. 6. – No. 9.
16. Yuille A. L., Rangarajan A. The concave-convex procedure //Neural computation. – 2003. – Vol. 15. – No. 4. – pp. 915-936.
17. DeSalvo G., Mohri M. Random composite forests //Proceedings of the AAAI Conference on Artificial Intelligence. – 2016. – Vol. 30. – No. 1.
18. Hehn T. M., Kooij J. F. P., Hamprecht F. A. End-to-end learning of decision trees and forests //International Journal of Computer Vision. – 2020. – Vol. 128. – No. 4. – pp. 997-1011.
19. Irsoy O., Alpaydin E. Autoencoder trees //Asian conference on machine learning. – PMLR, 2016. – pp. 378-390.
20. Chai Z., Zhao C. Multiclass oblique random forests with dual-incremental learning capacity //IEEE transactions on neural networks and learning systems. – 2020. – Vol. 31. – No. 12. – pp. 5192-5203.
21. Hecht-Nielsen R. Theory of the backpropagation neural network //Neural networks for perception. – Academic Press, 1992. – pp. 65-93.
22. Yang B. B., Shen S. Q., Gao W. Weighted oblique decision trees //Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33. – №. 01. – pp. 5621-5627.
23. Carreira-Perpinán M. A., Tavallali P. Alternating optimization of decision trees, with application to learning sparse oblique trees //Advances in neural information processing systems. – 2018. – Vol. 31.
24. Kumar M. A., Gopal M. A hybrid SVM based decision tree //Pattern Recognition. – 2010. – Vol. 43. – No. 12. – pp. 3977-3987.
25. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images //Master's thesis, University of Tront. – 2009.
26. Youtube channels dataset. URL: http://keen.isa.ru/youtube (accessed 14.07.2022).
27. Blake C. UCI repository of machine learning databases. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html (accessed: 14.07.2022).

Devyatkin Dmitry A. Researcher. Federal Research Center "Computer Science and Control", the Russian Academy of Sciences. Research areas: machine learning, natural language processing, information extraction, big data, methods for assessing the effectiveness of scientific activities. E-mail: devyatkin@isa.ru

Grigoriev Oleg G. Doctor of Technical Sciences. Chief Researcher. Federal Research Center “Computer Science and Control”, the Russian Academy of Sciences. Research areas: machine learning, big data, natural language processing, knowledge extraction. E-mail: oleggpolikvart@yandex.ru