Сравнительный анализ методов разбиения подвыборки обучающих данных ансамбля случайных деревьев

А.О. Шибаева[1], О.П. Солдатова[1]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. В работе сравнивается точность классификации данных ансамблями решающих деревьев с различными методами разбиения подвыборки обучающих данных. Идея алгоритма построения ансамбля решающих деревьев заключается в последовательном дроблении выборки выбранным методом на две части (подвыборки) до тех пор, пока не будет выполнено условие остановки. В работе реализованы следующие методы: разбиение по одному параметру, разбиение по двум параметрам (ориентированными прямыми) и разбиение по шести параметрам (эллипсами). В результате проведённых исследований были получены графики зависимости доли правильных ответов от значений параметров метода при различных вариантах разбиения. На основе полученных данных сделан вывод, что усложненные методы разбиения не дают большей точности классификации и требуют больше вычислений, чем более простые аналоги.

1. Введение

Случайный лес (Random forest) – алгоритм машинного обучения, предложенный Лео Брейманом и Адель Катлер, основан на использовании комитета (ансамбля) решающих деревьев [1]. Этот алгоритм выгодно отличается от других алгоритмов машинного обучения, поскольку он превосходит существующие аналоги в точности решения поставленной задачи при малом количестве сложных вычислений и практически не переобучается с ростом числа деревьев в композиции. Основная идея алгоритма заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их объединения результат получается хорошим [2].

В настоящее время алгоритм применяется для широкого класса задач, также появляются новые модификации алгоритма, которые позволяют обрабатывать как большие, так и малые объёмы данных с высокой точностью и быстродействием, которых не всегда можно добиться с использованием других алгоритмов машинного обучения и нейронных сетей.

В данной работе используется классический вариант алгоритма построения ансамбля бинарных решающих деревьев, в котором из исходных тестовых данных берётся случайным образом выборка с повторением [3]. Далее на каждом шаге построения дерева текущая выборка разбивается различными методами на две подвыборки, которые далее разбиваются в следующих узлах дерева. Наилучший предикат выбирается по наилучшему значению меры неоднородности, которая в данной статье представлена энтропией [4].

В качестве метода разбиения можно использовать любые правила для любого количества критериев, однако чаще всего используется простое пороговое разбиение по одному критерию.

Разбиение линиями распространено гораздо меньше, чем разбиение по одному критерию, однако их исследования набирают популярность. Так, например, применение разбиения линиями на данных больших размерностей описывается в статье «Classifying very-high-dimensional data with random forests of oblique decision trees» [5], а в статье «CO2 Forest: Improved Random Forest by Continuous Optimization of Oblique Splits» [6] описывается вариант их дальнейшей оптимизации, основанной на введении верхней гладкой оценки. Любые другие более сложные варианты разбиений не рассматриваются, поэтому в данной работе дополнительно исследуется разбиение по пяти параметрам эллипсами для сравнения с вышеперечисленными методами разбиения.

Введём следующие обозначения:

- x_i – значение i-го критерия вектора;
- a_j – j-й случайно сгенерированный коэффициент.

В данной работе будут рассмотрены следующие варианты разбиения:

- Пороговое разбиение по одному критерию:

$$x_i > a_i$$

- Линейное разбиение по двум критериям:

$$x_2 > a_1 x_1 + a_2$$

- Нелинейное разбиение второго порядка по двум критериям (эллипс):

$$a_1(x_1 - a_2)^2 + a_3(x_2 - a_4)^2 + a_5 > 0$$

При реализации разбиения по одному критерию в качестве номеров критериев будут генерироваться единичные значения критериев с повторением, также случайным образом будет генерироваться единственное значение текущего коэффициента за итерацию цикла. Поэтому выходными данными будут являться один критерий и одно значение данного критерия.

При реализации линейного разбиения по двум критериям в качестве номеров критериев будут генерироваться пары не равных между собой чисел из общего количества критериев вектора с повторением, также для данных будет генерироваться пара значений коэффициентов за цикл, что соответствует формуле линейного разбиения.

При реализации разбиения эллипсами по двум критериям в качестве номеров критериев будут генерироваться пары не равных между собой чисел из общего количества критериев вектора с повторением, также будет генерироваться пять значений коэффициентов за цикл, что соответствует формуле разбиения эллипсом.

Под термином вариант разбиения будем понимать один сгенерированный вариант выбора номеров критериев с соответствующими сгенерированными значениями коэффициентов.

Основным недостатком построения деревьев является жадность алгоритма, так как локально оптимальный выбор коэффициентов метода разбиения не является глобально оптимальным. Под локальной оптимизацией будет подразумевать выбор таких номеров критериев и значений коэффициентов, при которых достигается наименьшее значение суммарной энтропии на текущем разбиении. В случае неудачного выбора алгоритм не способен вернуться на уровень вверх и изменить неудачные коэффициенты метода.

Для исследования точности решения задачи классификации с помощью ансамбля решающих деревьев были использованы модельные данные для двух случайных величин с нормальным законом распределения в двумерном пространстве. Набор состоит из двух классов, каждый из которых представляет случайные величины с заданными различными значениями математического ожидания и дисперсии. Каждый класс включает в себя 400 обучающих двумерных векторов. Тестирование проводилось с помощью проверки всех значений в диапазоне [0;5] с шагом 0,04. В статье приведены результаты решения задачи классификации на примере сгенерированных данных для графической иллюстрации поведения решающих деревьев с различными вариантами разбиения.

2. Исследование зависимости доли правильных ответов от параметров метода при разбиении по одному параметру

На рисунке 1 приведены результаты работы алгоритма классификации двух случайных величин с нормальным законом распределения при параметрах, приведённых в таблице 1, а также доли правильных ответов для каждого дерева. Как видно из представленного рисунка, эти классы линейно не разделимы, поэтому эффективное разделение должно быть произведено с некоторой погрешностью, чтобы не допускать переобучения и получать хорошую точность при проверке новых значений.

Как видно из рисунка 1, при меньших параметрах обучения дерево получается более абстрактным, что повышает его способность к обобщению, а при больших параметрах дерево становится более чувствительным к шумам.

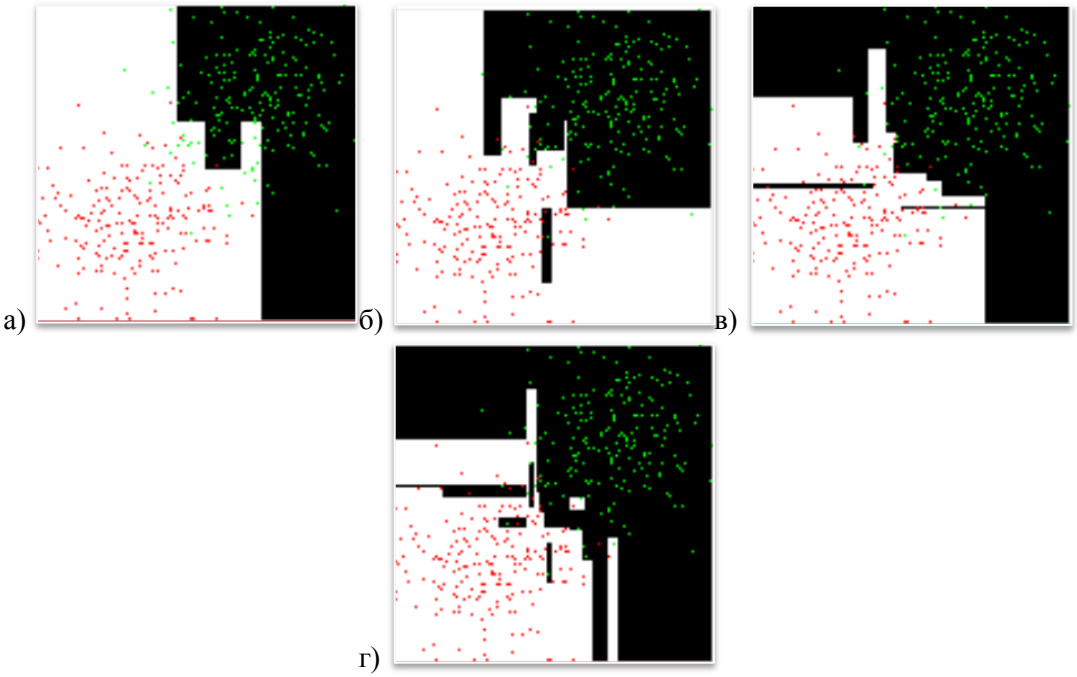


Рисунок 1. Результат работы алгоритма с разбиением по одному критерию: а) доля правильных ответов равна 0,9275; б) доля правильных ответов равна 0,9575; в) доля правильных ответов равна 0,975; г) доля правильных ответов равна 0,975.

Однако ни при каких параметрах одно дерево не даёт удовлетворительного результата, при котором классы разделяются таким образом, чтобы минимально реагировать на шумы при новых значениях, потому что каждое представленное дерево обучалось на зашумлённых значениях без возможности абстрагироваться от шума.

Таблица 1. Значения параметров алгоритма.

Рисунок	Размер подвыборки обучения	Количество значений критериев	Количество вариантов разбиения
1а	50	1	1
1б	50	2	100
1в	400	1	1
1г	400	2	100

Для повышения качества обучения увеличим число деревьев до 10 при размере подвыборки обучения равной 400, количестве критериев равным 2 и количестве вариантов равным 3. На рисунке 2 показан результат работы алгоритма построения ансамбля из 10 деревьев. Слева показан конечный результат тестирования всего ансамбля, справа насыщенностью цвета

показано количество голосов – чем больше деревьев отдало голосов за класс, тем ярче цвет класса.

Качество обучения заметно возросло, так как теперь ансамбль реагирует только на небольшой разброс шумов, и доля правильных ответов составила 0,9975, и, хотя граница конечного разделения классов всё ещё выглядит разорванной, можно заметить, что в области пересечения данных примерно половина деревьев голосует за каждый класс, что соответствует здравому смыслу.

Проверим, как поведёт себя ансамбль при увеличении количества деревьев до 100 при неизменных остальных параметрах. На рисунке 3 приведен результат работы алгоритма построения ансамбля из 100 деревьев. Доля правильных ответов на обучающей выборке равна 1. Так как граница разделения классов стала ещё более гладкой, ансамбль практически не будет реагировать на единичные отклонения новых данных.

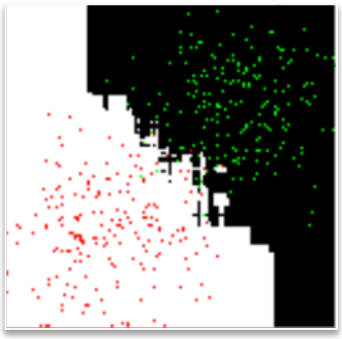
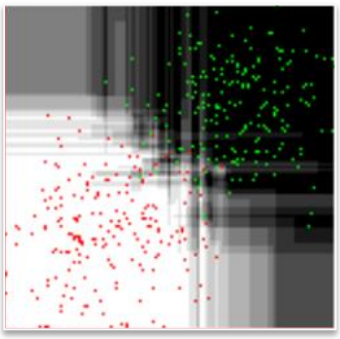
а)  б) 

Рисунок 2. Результат работы алгоритма с разбиением по одному критерию (количество деревьев – 10): а) результат тестирования работы всего ансамбля; б) наложение результатов тестирования отдельных деревьев ансамбля.

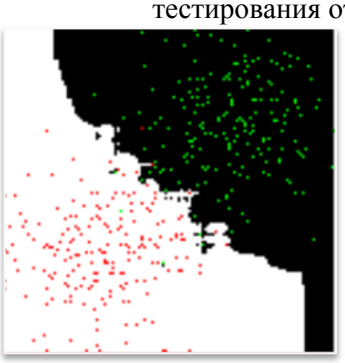
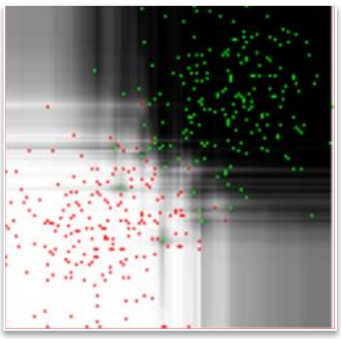
а)  б) 

Рисунок 3. Результат работы алгоритма с разбиением по одному критерию (количество деревьев – 100): а) результат тестирования работы всего ансамбля; б) наложение результатов тестирования отдельных деревьев ансамбля.

3. Исследование зависимости доли правильных ответов от параметров метода при разбиении по двум параметрам (разбиение ориентированными прямыми)

Анализируя полученные при разбиении по одному параметру результаты, можно прийти к выводу, что разбиение ориентированной линией может дать более простой и более точный результат. Для проверки данной гипотезы были проведены соответствующие исследования.

На рисунке 4 приведены результаты работы алгоритма разбиения ориентированными линиями и доли правильных ответов по двум критериям при значениях параметров, приведённых в таблице 2.

Из рисунков видно, что граница получается разорванной и даже сильно отличающейся от удовлетворительного разбиения, и варьирование значений параметров алгоритма не оказывают значительного влияния на результат.

Такой результат связан в первую очередь с тем, что для генерации значений коэффициентов используются, согласно классическому алгоритму, случайные числа с локальной оптимизацией на каждом узле. Для улучшения качества классификации увеличим количество деревьев до 10

при размере подвыборки обучения равной 400, количестве критериев равным 2 и количестве вариантов равным 3, что соответствует параметрам, которые использовались при разбиении первым методом.

Таблица 2. Значения параметров алгоритма.

Рисунок	Размер подвыборки обучения	Количество вариантов разбиения
1а	50	1
1б	50	100
1в	400	1
1г	400	100

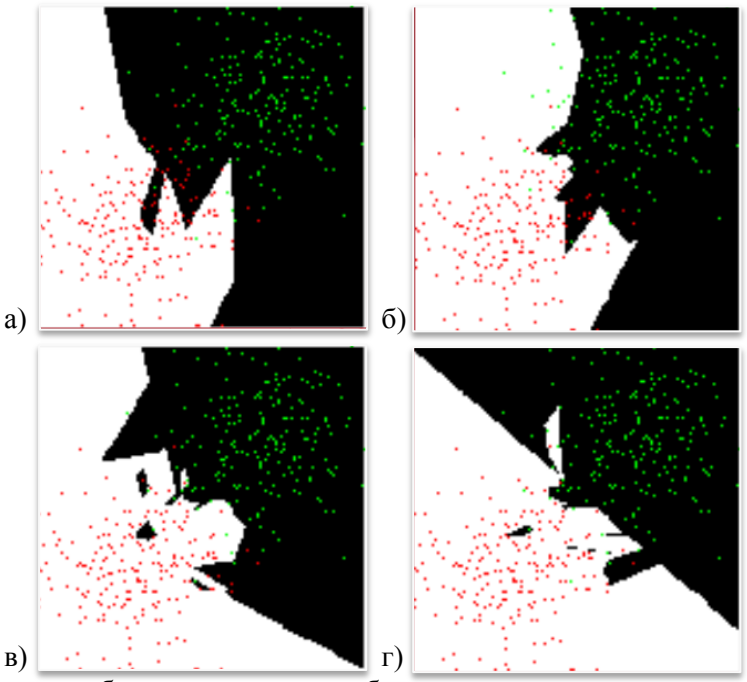


Рисунок 4. Результат работы алгоритма с разбиением по двум критериям ориентированными прямыми: а) доля правильных ответов равна 0,92; б) доля правильных ответов равна 0,925; в) доля правильных ответов равна 0,9725; г) доля правильных ответов равна 0,9875.

На рисунке 5 приведен результат работы алгоритма. Как видно из рисунка справа, хотя на тестирующей выборке ошибка сведена к минимуму (доля правильных ответов равна 1), при генерации других данных по этому закону распределения, ошибка может существенно возрасти, потому что разбиение классов сильно реагирует на шум, и линия разбиения является ломаной.

Увеличим количество деревьев до 100 при неизменных остальных параметрах. На рисунке 6 приведен результат работы алгоритма. Алгоритм безошибочно работает на тестовой выборке (доля правильных ответов равна 1). Из рисунка видно, что теперь граница разделения близка к гладкой, что означает, что алгоритм будет также работать адекватно при появлении новых данных.

4. Исследование зависимости доли правильных ответов от параметров метода при разбиении по пяти параметрам (разбиение эллипсами)

Анализируя внешний вид данных можно заметить, что представленные данные можно вписать в окружности с некоторым центром и радиусом. Используем разделение эллипсом по двум критериям с 5 коэффициентами, согласно формуле, указанной в пункте 1. Построим по одному дереву с параметрами, указанными в таблице 2.

На рисунке 7 приведены результаты работы алгоритма. Отметим, что при количестве вариантов разбиения равным 1, во всех листах дерева оказывается только один класс. Это происходит

потому, что сгенерированными значениями коэффициентов являются 5 случайных значений, разброс которых может быть любым. Отсюда следует, что чем больше вариантов разбиения, тем эффективнее можно выбрать разбиение на конкретном шаге и тем точнее получится результат.

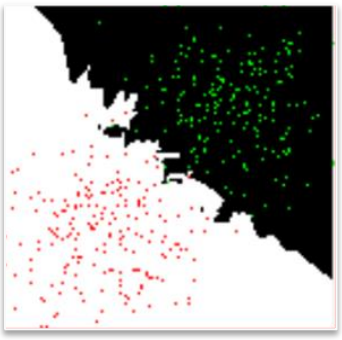
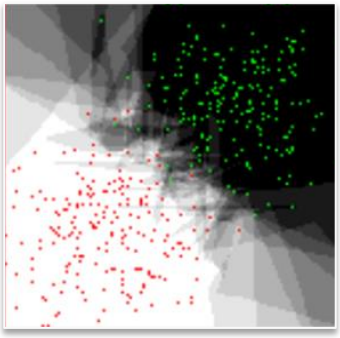
а)  б) 

Рисунок 5. Результат работы алгоритма с разбиением по двум критериям ориентированными прямыми: (количество деревьев – 10): а) результат тестирования работы всего ансамбля; б) наложение результатов тестирования отдельных деревьев ансамбля.

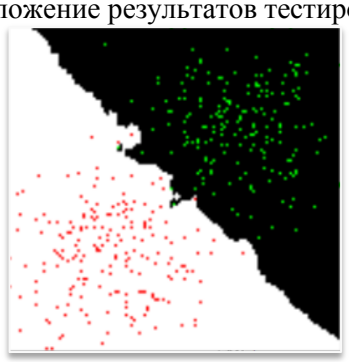
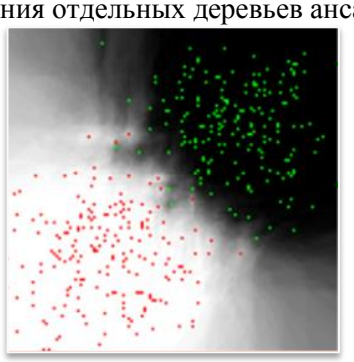
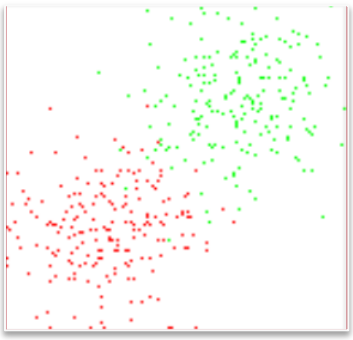
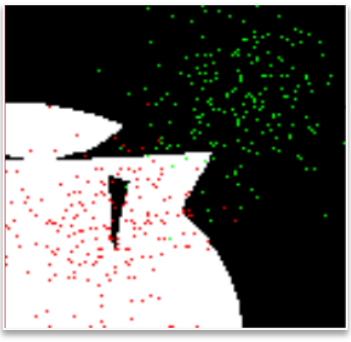
а)  б) 

Рисунок 6. Результат работы алгоритма с разбиением по двум критериям ориентированными прямыми: (количество деревьев – 100): а) результат тестирования работы всего ансамбля; б) наложение результатов тестирования отдельных деревьев ансамбля.

а)  б) 

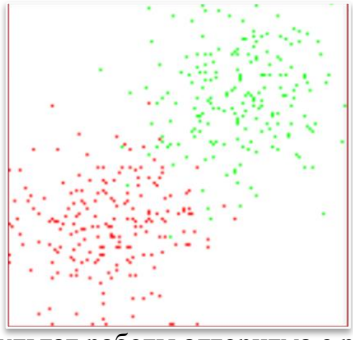
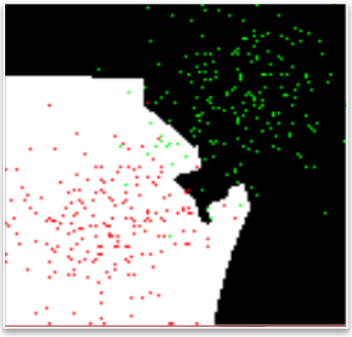
в)  г) 

Рисунок 7. Результат работы алгоритма с разбиением по двум критериям эллипсами: а) доля правильных ответов равна 0,5; б) доля правильных ответов равна 0,935; в) доля правильных ответов равна 0,5; г) доля правильных ответов равна 0,9575.

Однако, число вариантов, при котором будет достигнут наилучший результат на каждом шаге построения дерева, будет очень большим из-за того, что значения коэффициентов являются случайными числами с большим разбросом значений.

Увеличим число деревьев до 10 при следующих параметрах: размер подвыборки обучения равен 400, количестве критериев равно 2 и количестве вариантов равно 3. На рисунке 8 представлены результаты работы алгоритма. Заметим, что одно дерево (часть эллипса на рисунке справа обозначено более светлым тоном), показывает наиболее приемлемый результат, но остальные деревья являются неудачными и портят результат классификации (так как относят все значения к одному классу), поэтому итоговая доля правильных ответов равна 0,8725. Увеличим число деревьев до 100.

На рисунке 9 представлен результат работы алгоритма. Несмотря на то, что остаётся множество деревьев с неудачными параметрами (это можно утверждать по преобладающему серому цвету среднего тона на рисунке справа), общая способность классификации всего ансамбля заметно возросла (доля правильных ответов равна 0,9125) и теперь такой ансамбль почти не уступает ансамблям с такими же значениями параметров для других разбиений.

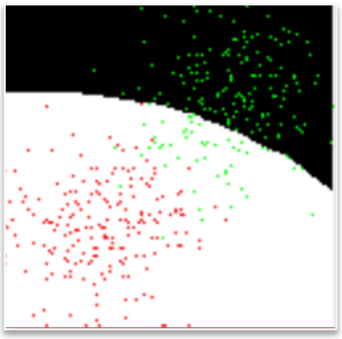
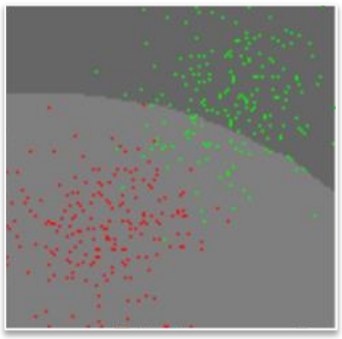
а)  б) 

Рисунок 8. Результат работы алгоритма с разбиением по двум критериям эллипсами: (количество деревьев – 10): а) результат тестирования работы всего ансамбля; б) наложение результатов тестирования отдельных деревьев ансамбля.

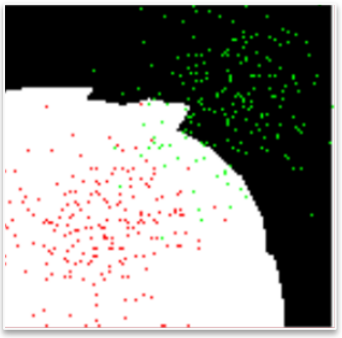
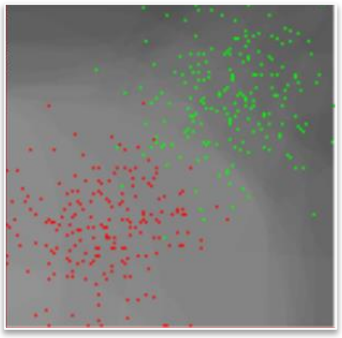
а)  б) 

Рисунок 9. Результат работы алгоритма с разбиением по двум критериям эллипсами: (количество деревьев – 100): а) результат тестирования работы всего ансамбля; б) наложение результатов тестирования отдельных деревьев ансамбля.

5. Выводы

В работе были рассмотрены три разных метода разбиения подвыборки модельных данных, сгенерированных в соответствии с нормальным законом распределения. Разбиение по одному параметру и разбиение ориентированными линиями по двум параметрам показали примерно одинаковые значения доли правильных ответов и имеют малый объём вычислений. Это связано с тем, что на каждой итерации генерируется только одно значение коэффициента при методе с одним параметром и два при методе с двумя параметрами, в дальнейшем используются простые формулы вычисления (см. пункт 1), поэтому их рекомендуется использовать при решении задачи классификации.

Из полученных результатов можно сделать вывод, что для эффективной работы более сложных алгоритмов, необходимо использовать больше вариантов разбиения, или более сложные

алгоритмы генерации коэффициентов разбиения, однако это требует большего объёма вычислений, чем простое разбиение. Из-за жадности алгоритма никакая локальная оптимизация не может гарантировать удовлетворительный результат, особенно при пересекающихся значениях классов или при сильном зашумлении данных.
Из вышесказанного следует, что для достижения большей точности без усложнения вычислительных методов разбиения в дальнейших исследованиях необходимо обратить внимание на другие параметры построения случайных деревьев, как например, в статье [7], где предложенные методы оценки качества разбиения в узле дерева показывают лучший результат, чем широко распространённая энтропия.

6. Литература

[1] Breiman, L. Random forests // Machine Learning, 2001. – P. 5-32.
[2] Введение в машинное обучение и анализ данных [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c6/SIS_MachineLearning_260812.pdf (20.10.2018).
[3] Воронцов, К.В. Лекции по логическим алгоритмам классификации [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/3/3e/Voron-ML-Logic.pdf (05.10.2018).
[4] Заметки по R [Электронный ресурс]. – Режим доступа: https:// bdemeshev.github.io/r_cycle/cycle_files/22_forest.html (28.10.2018).
[5] Classifying very-high-dimensional data with random forests of oblique decision trees [Электронный ресурс]. – Режим доступа: https://pdfs.semanticscholar.org/b18c/8760f0eae111d48aa71b3178a98e67854daf.pdf.
[6] CO2 Forest: Improved Random Forest by Continuous Optimization of Oblique Splits [Электронный ресурс]. – Режим доступа:https://arxiv.org/pdf/1506.06155v2.pdf.
[7] Semi-supervised Node Splitting for Random Forest Construction [Электронный ресурс]. – Режим доступа: https://www.cv-foundation.org/openaccess/content_cvpr_2013/papers/Liu_Semi-supervised_Node_Splitting_2013_CVPR_paper.pdf.

Comparative analysis of subset splitting methods for training data in decision tree ensemble

A.O. Shibaeva[1], O.P. Soldatova[1]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

Abstract. In this article is compared the accuracy of data classifying by Decision Tree Ensembles with different methods of subset splitting methods for training data. The idea of the algorithm for constructing an Decision Tree Ensemble is the sequential splitting of the sample by the selected method into two parts (subsets) until the stopping condition is satisfied. The methods are splitting by one feature, splitting by two features (oriented straights), and splitting by six features (ellipses).As a result of the study, was obtained graphs of the proportion of correct answers to the variation of different parameters of the method for different variants of the splitting. Based on the data obtained, it was concluded that the complicated splitting methods do not provide greater classification accuracy and require more calculations than simpler analogues.