УДК 528.854

DOI: https://doi.org/10.26456/2226-7719-2018-3-99-107

МЕТОД СЛУЧАЙНЫХ ЛЕСОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ СПУТНИКОВЫХ СНИМКОВ

Э. М. Купенова[1], А. В.Кашницкий[2]
[1]МГУ им. М. В. Ломоносова, Москва
[2]Институт Космических Исследований РАН, Москва

В ходе данной работы было рассказано о понятии и видах классификации спутниковых снимков, разобран алгоритм построения случайного леса, были показаны основные преимущества данного метода классификации перед множеством различных существующих методах классификаций. Показаны промежуточные результаты работы макета классификатора тестового изображения.

Ключевые слова: *классификация, случайный лес, деревья решений, обучающая выборка, растровое изображение, бутстрэп, бэггинг.*

Классификация изображений – это процесс извлечения классов информации из многоканального растрового изображения. Растр, полученный в результате классификации изображения, можно использовать для создания тематических карт. В зависимости от характера взаимодействия аналитика с компьютером в процессе классификации, различают два типа классификации изображений: классификацию с обучением и классификацию без обучения.

Методы классификации с обучением основаны на обучении решающего алгоритма. Обучение осуществляется путем выбора области интереса (обучающего фрагмента изображения) и исследовании значений характеристик объектов в пространстве признаков. Обучающая выборка создается оператором, она основывается на результатах предыдущих исследований и наличия снимков – эталонов.

Преимущество методов неконтролируемой классификации заключается в том, что на вход практически не требуется вводить данные. Все операции выполняются автоматически, при этом программа анализирует пространство спектральных параметров и на основании определенных критериев разделяет пиксели на классы, рассчитывая для каждого из них средние значения признаков и ковариационные матрицы. После того как все данные распределены по спектральным классам, оператор старается сопоставить их с известными информационными классами пространственных объектов.

Одним из методов классификации с обучением является метод с использованием случайных лесов.

В начале 90-ых годов 20-го века появились первые работы, которые были связаны с построением ансамблей решающих деревьев.

В 1994 году Лео Брейманом был придуман бэггинг (bagging сокр. от bootstrap aggregation) — это один из первых и самых простых видов ансамблей. Бэггинг основан на статистическом методе бутстрэпа, который позволяет оценивать многие статистики сложных распределений.

Метод бутстрэпа заключается в следующем. Пусть имеется выборка X размера N. Равномерно возьмем из выборки N объектов с возвращением. Это означает, что мы будем N раз выбирать произвольный объект выборки (считаем, что каждый объект «достается» с одинаковой вероятностью 1/N), причем каждый раз мы выбираем из всех исходных N объектов. Можно рассмотреть произвольное множество, из которого случайным образом выбирается какой-либо его элемент. Этот элемент множества фиксируется и следующий выбор опять делается равновероятно из того же множества. Очевидно, что благодаря такой системе выбора элементов, у нас окажутся повторы какого-либо элемента со случайной частотой. Обозначим новую выборку через X_1. Повторяя процедуру M раз, сгенерируем M подвыборок $X_1,\ldots,X_M$. Теперь мы имеем достаточно большое число выборок и можем оценивать различные статистики исходного распределения.

Теперь имея представление о бутстрэпе, отметим достоинства бэггинга. Бэггинг позволяет снизить дисперсию обучаемого классификатора, уменьшая величину, на сколько ошибка будет отличаться, если обучать модель на разных наборах данных, или другими словами, предотвращает переобучение. Эффективность бэггинга достигается благодаря тому, что базовые алгоритмы, обученные по различным подвыборкам, получаются достаточно различными, и их ошибки взаимно компенсируются при голосовании, а также за счёт того, что объекты-выбросы могут не попадать в некоторые обучающие подвыборки.

Лео Брейман нашел применение бутстрэпу не только в статистике, но и в машинном обучении. Он вместе с Адель Катлер усовершенствовал алгоритм случайного леса, предложенный Хо, добавив к первоначальному варианту построение не коррелируемых деревьев на основе CART (Classification And Regression Tree), в сочетании с методом случайных подпространств и бэггинга.

Метод быстро получил признание, поскольку, помимо высокой точности классификации, он имеет ряд положительных свойств, описанных ниже:

– Метод исключает возможность от переобучения даже в случае, когда количество признаков заметно превосходит количество наблюдений. Данное свойство сильно выделяет метод случайных лесов

от остальных существующих методов классификации и является очень важным свойством для решения прикладных задач;

– Для построения случайного леса по обучающей выборке требуется задать лишь два параметра, которые требуют минимальной настройки;

– Метод out-of-bag (OOB), предложенный Брейманом, обеспечивает получение естественной оценки вероятности ошибочной классификации случайных лесов на основе наблюдений, не входящих в обучение бутстрэп выборки, используемые для построения деревьев (эти построения называют OOB выборками);

– Случайные леса могут использоваться не только для задач классификации, но и для задач регрессии, кластеризации, выявления наиболее информативных признаков, выделения аномальных наблюдений и определения прототипов классов;

– Обучающая выборка для построения случайного леса может содержать признаки, измеренные в разных шкалах: числовой, порядковой и номинальной, что недопустимо для многих других классификаторов;

– Метод позволяет применять параллельные вычисления, что заметно сокращает время работы при больших объемах обучающей выборки.

Случайный лес состоит из большого числа (ансамбля) решающих деревьев (является одним из основных параметров метода). Для того, чтобы понять, как работает случайный лес, нужно разобраться, что из себя представляют решающие деревья (или деревья решений).

Деревья решений используются для предсказывания значения переменной на основе определенного количества входных параметров. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева записаны параметры, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — параметры, по которым различаются случаи. Для того, чтобы провести классификацию нового случая, необходимо спуститься по дереву до листа и выдать соответствующее значение. В простейшем случае у нас на вход подается всего один признак, для которого хорошо различимы границы между классами (то есть максимальное значение первого класса значительно меньше минимального значения второго класса). Следовательно, достаточно всего одного показателя, чтобы предсказать класс и получить ответ.

В задачах классификации в качестве решения принимается значение, получившее большинство голосов, при условии, что каждое дерево в лесу обладает одним голосом. Деревья в лесу строятся согласно следующей схеме:

– Берется подмножество обучающей выборки, на основе которого строится решающее дерево. Деревья строятся на основе разных подмножеств, что решает проблему построения одинаковых деревьев.

– Используем число случайных признаков для построения и выбора расщеплений в деревьях.

– Выбираем наилучший признак и построенное по нему расщепление.

Тренировочные наборы, на основе которых происходит обучение деревьев, генерируются из исходной обучающей выборки с использованием процедуры бутстрэп: для каждого набора обучения случайным образом выбирается то же количество векторов, что и в исходном наборе. Векторы выбираются с заменой. То есть некоторые векторы будут происходить не один раз, а некоторые будут отсутствовать. В каждом узле каждого обученного дерева не все переменные используются для поиска наилучшего разбиения, а случайное их подмножество. С каждым узлом генерируется новое подмножество. Однако его размер фиксируется для всех узлов и всех деревьев. Этим обучающим параметром является число признаков для выбора расщепления, по умолчанию равный квадратному корню от количества параметров. Данное число признаков используется для построения и выбора расщеплений в деревьях. Далее выбирается наилучший признак и построенное по нему расщепление.

Алгоритм построения случайного леса, состоящего из N деревьев, выглядит следующим образом:

Для каждого n=1,…,N:

Сгенерировать выборку X_n с помощью бутстрэпа;

Построить решающее дерево b_n по выборке X_n:

– по заданному критерию мы выбираем лучший признак, делаем разбиение в дереве по нему и так до исчерпания выборки

– дерево строится, пока в каждом листе не более n_{min} объектов или пока не достигнем определенной высоты дерева

– при каждом разбиении сначала выбирается m случайных признаков из n исходных, и оптимальное разделение выборки ищется только среди них.

Итоговый классификатор $a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x)$, простыми словами — для задачи классификации мы выбираем решение голосованием по большинству.

Рекомендуется в задачах классификации брать m = n, где n – число признаков. Также рекомендуется строить каждое дерево до тех пор, пока в каждом листе не окажется по одному объекту.

Таким образом, случайный лес – это бэггинг над решающими деревьями, при обучении которых для каждого разбиения признаки выбираются из некоторого случайного подмножества признаков.

Известно, что точность (вероятность корректной классификации) ансамблей классификаторов существенно зависит от разнообразия классификаторов, составляющих ансамбль (на сколько коррелированы их решения). То есть, чем более разнообразны классификаторы ансамбля, тем выше вероятность корректной классификации. В случайных лесах решения составляющих их деревьях слабо коррелированы за счет наличия стадий бутстрэп и случайного отбора признаков, используемых при расщеплении вершин деревьев.

Промежуточные результаты

В настоящее время для информационной системы «ВЕГА-Science» разрабатывается программа, позволяющая проводить классификацию спутниковых изображений на основе метода случайных лесов. Макет классификатора был написан на языке Python 2.7. Для Python в библиотеке skilit-learn есть реализация классификатора RandomForestClassifier().

```
class sklearn.ensemble.RandomForestClassifier(n_estimators = 10,
        criterion = 'gini', max_depth = None, min_samples_split = 2,
        min_samples_leaf = 1, min_weight_fraction_leaf = 0.0,
        max_features = 'auto', max_leaf_nodes = None, min_impurity_decrease = 0.0,
        min_impurity_split = None, bootstrap = True, oob_score = False,
        n_jobs = 1, random_state = None, verbose = 0, warm_start = False,
        class_weight = None)
```

Рассмотрим основные параметры этой функции.

1) n_estimators

Данный параметр отвечает за количество деревьев в лесу. С увеличением количества деревьев увеличивается качество классификации, но также возрастает время работы;

2) criterion

Функция оценки качества разбиения. Но в этой функции для задач классификации поддерживаются критерии «Джини» для примеси Джини и «энтропия» для получения информации;

3) max_depth

Максимальная глубина деревьев. Ясно, что чем меньше глубина, тем быстрее строится и работает случайный лес. При увеличении глубины резко возрастает качество на обучении, но и на контроле оно, как правило, увеличивается. Рекомендуется использовать максимальную глубину (кроме случаев, когда объектов слишком много и получаются очень глубокие деревья, построение которых занимает значительное время). Неглубокие деревья рекомендуют использовать в задачах с большим числом шумовых объектов (выбросов);

4) min_samples_split

Минимальное количество образцов, необходимых для разделения внутреннего узла. По умолчанию, этому параметру присваивается значение 2. Если тип параметра int, то рассматриваем min_samples_split как минимальное число;

5) min_samples_leaf

Ограничение на количество объектов в листьях;

6) max_features

Количество функций, которые следует учитывать при поиске наилучшего разделения. Если int, то рассматриваем функции max_features при каждом разбиении;

7) bootstrap

Используются ли bootstrap образцы при построении деревьев. Может принимать занчения True и False;

8) n_jobs

Количество заданий, выполняемых параллельно как для fit, так и для predict. Если значение -1, то число заданий устанавливается равным числу ядер. По умолчанию берется значение 1;

9) oob_score

Следует ли использовать образцы oob для оценки точности обобщения. Так же, как и аргумент bootstrap, принимает значения True и False;

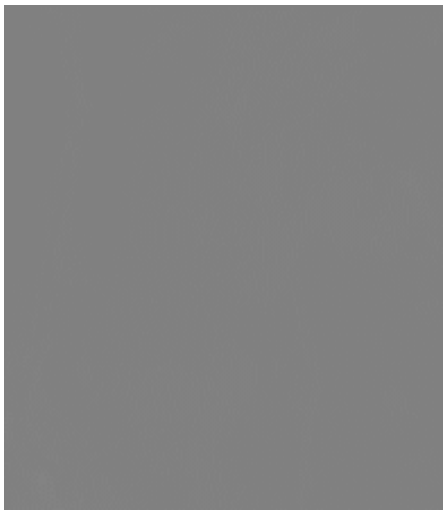
10) class_weight

Веса классов. Если значение этого параметра не указано, все классы должны иметь вес единицу. Для задач с несколькими выходами список диктовок может быть предоставлен в том же порядке, что и столбцы y;

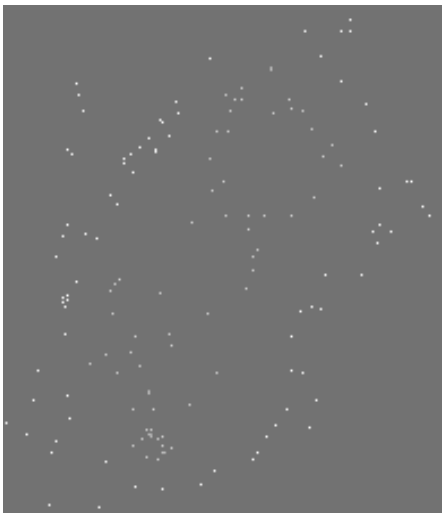
11) min_weight_fraction_leaf

Минимальная взвешенная доля от общей суммы весов (всех входных выборок), требуемая для конечного узла. Образцы имеют одинаковый вес, если параметр sample_weight не указан.

Теперь перейдем к результатам работы макета классификатора. Была рассмотрена задача классификации лесной гари. Программа была протестирована на примере снимка со спутника Landsat-8. В качестве входных параметров было взято изображение в формате GeoTIFF (рис. 1) с двумя каналами (RED и NIR), а также растеризованное (на основе GeoJSON файла с обучающей выборкой) изображение с метками классов (рис. 2).



Р и с. 1.



Р и с. 2.

Классу «Гарь» был присвоен красный цвет, классу «Лес» - зеленый цвет. В итоге было получено классифицированное изображение (Рис. 3)

Р и с. 3.

Список литературы

1. Картиев С.Б., Курейчик В.М. 2016. Алгоритм классификации, основанный на принципах случайного леса, для решения задачи прогнозирования // Программные продукты и системы / Software & Systems № 2 (114). С. 11–15.
2. Шовенгердт Р. А. 2010. Дистанционное зондирование. Модели и методы обработки изображений М.: Техносфера. –560 с.

3. Breiman L. 2001. Random forests // Machine learning. V. 45. №. 1. P. 5–32.
4. Campbell J. B., Wynne R. H. 2011. Introduction to remote sensing. 5th ed. New York, London: The Guilford Press. 667 p.
5. Барталев С. А., Егоров В. А., Жарко В. О., Лупян Е. А., Плотников Д. Е., Хвостиков С. А., Шабанов Н. В. 2016. Спутниковое картографирование растительного покрова России М.: ИКИ РАН 208 с.
6. Лупян Е.А., Барталев С.А., Толпин В.А., Жарко В.О., Крашенинникова Ю.С., Оксюкевич А.Ю. Использование спутникового сервиса ВЕГА в региональных системах дистанционного мониторинга // Современные проблемы дистанционного зондирования Земли из космоса, 2014. Т. 11. №. 3. С.215–232.
7. Кашницкий А.В., Балашов И.В., Лупян Е.А., Толпин В.А., Уваров И.А. Создание инструментов для удаленной обработки спутниковых данных в современных информационных системах // Современные проблемы дистанционного зондирования Земли из космоса. 2015. Т.12. № 1. С.156–170.
8. Информационная система «ВЕГА-Science». Электронный ресурс. URL: http://sci-vega.ru/.
9. Лупян Е. А., Прошин А. А., Бурцев М. А., Балашов И. В., Барталев С. А., Ефремов В. Ю., Кашницкий А. В., Мазуров А. А., Матвеев А. М., Суднева О. А., Сычугов И. Г., Толпин В. А., Уваров И. А. Центр коллективного пользования системами архивации, обработки и анализа спутниковых данных ИКИ РАН для решения задач изучения и мониторинга окружающей среды // Современные проблемы дистанционного зондирования Земли из космоса. 2015. Т. 12. № 5. С. 263–284.

THE METHOD OF RANDOM FORESTS IN THE CLASSIFICATION TASKS OF SATELLITE IMAGES

E. M.Kupenova, A.V. Kashnitskiy [2]

[1]Lomonosov Moscow State University, Moscow
[2]Space Research Institute Russian Academy of Sciences, Moscow

In the course of this work, the concept and types of classification of satellite images were discussed, the algorithm of constructing a random forest was analyzed, the main advantages of this method of classification were shown over many different existing methods of classification. The intermediate results of the test image classifier layout are shown.

***Keywords**: classification, random forest, decision trees, training sample, raster image, bootstrap, bagging.*

Об авторах:

КУПЕНОВА Эльвира Максатовна, студент-магистр, МГУ им. Ломоносова, Москва, Ленинские горы, д.1, 119234, kupenova_elya@mail.ru

КАШНИЦКИЙ Александр Витальевич, к.т.н., Институт космических исследований Российской академии наук, Москва, Профсоюзная 84/32, kashnizky@gmail.com