

Лекция. Искусственный интеллект и машинное обучение

Павлова

22 марта 2025 г.

ВОПРОСЫ:

- Основные понятия метода случайного леса. Особенности алгоритма
- Достоинства и недостатки алгоритма случайный лес

1 Основные понятия метода случайного леса

Случайный лес (англ. Random forest)- алгоритм машинного обучения, заключающийся в использовании комитета (ансамбля) решающих деревьев.

В статистике и машинном обучении под **ансамблем моделей** понимают комбинацию нескольких алгоритмов обучения, позволяющую создать модель машинного обучения (ММО) более эффективную и точную. Данную ММО называют **метамоделью**.

Термин случайный лес используется в дисциплинах: машинное обучение, распознавание образов, Data Mining, прикладная статистика. Алгоритм применяется для решения задач классификации, регрессии и кластеризации.

Алгоритм машинного обучения, предложенный Л. Брейманом и А. Катлером, использует **ансамбль** решающих деревьев (деревьев решений) (один из параметров метода). Ансамбль моделей требует больше вычислительных затрат в отличие от одной модели машинного обучения. Поэтому работу ансамбля можно рассматривать как способ компенсации «плохих» алгоритмов обучения путем дополнительных вычислений. Алгоритмы обучения с учителем используют для решения задачи поиска в пространстве гипотез. Ансамбли моделей объединяют несколько гипотез. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству.

Использование ансамбля моделей позволяет получать более точные результаты. Однако при построении ансамблей классификаторов во многих случаях наилучшим оказывается число моделей, равное числу классов.

Существует несколько методов объединения моделей машинного обучения в ансамбли:

Оптимальный байесовский классификатор — ансамбль, состоящий из простых классификаторов Байеса, взвешенных их апостериорными вероятностями.

Бэггинг — ансамбль моделей, обучающихся параллельно на различных случайных выборках одного и того же обучающего множества. Определение конечного результата определяется путем голосования классификаторов ансамбля. Деревья решений очень чувствительны к данным, на которых обучаются: небольшие изменения в обучающем наборе могут привести к существенно разным древовидным структурам. Случайный лес использует это преимущество, позволяя каждому отдельному дереву случайным образом выбирать из

набора данных, в результате чего получаются разные деревья. Этот процесс известен как бэггинг.

Бустинг — ансамбль моделей, обучающихся последовательно. Каждый последующий алгоритм обучается на примерах, в которых предыдущий классификатор допустил ошибку. В этом случае бустинг имеет более точные результаты, в отличие от бэггинга. Бэггинг склонен к переобучению.

Стекинг — обучающее множество разбивается на N блоков, и на $N-1$ обучается столько же базовых моделей. Далее N -я модель обучается на оставшемся блоке. В качестве целевой переменной используются выходные значения основных классификаторов, образующие метапризнак. Недостаток стекинга: метод значительно улучшает результаты базовых классификаторов при относительно большом числе обучающих примеров (несколько десятков тысяч).

Методы построения ансамблей позволяют работать с различными видами моделей: регрессией, искусственными нейронными сетями, деревьями решений, алгоритмами кластеризации. В машинном обучении используют ансамбли, разработанные специально для моделей одного типа. Например, метод случайного леса разработан на основе **ансамблей деревьев решений**.

Алгоритм случайный лес реализует: **метод бэггинга Бреймана** и **метод случайных подпространств**, предложенный Тин Кам Хо.

В случайных лесах решения составляющих их деревьев слабо коррелированы вследствие двойной "инъекции случайности" в алгоритм построения случайного леса — на стадии бутстрепа и на стадии случайного отбора признаков, используемых при расщеплении вершин деревьев.

В отличие от классических алгоритмов построения деревьев решений в методе случайного леса при построении каждого дерева на стадиях расщепления вершин используется только фиксированное число случайно отбираемых признаков обучающей выборки (второй

параметр метода) и строится полное дерево (без усечения дерева решений). Каждый лист дерева включает примеры только одного класса объектов.

Случайность признаков. В обычном дереве решений, когда приходит время деления, используются все возможные признаки и выбирается признак, который дает наилучшее разделение между наблюдениями в левом узле и наблюдениями в правом узле дерева решений. В случайном лесе дерево создается из случайного подмножества функций-признаков. Это создает большее разнообразие деревьев в модели и приводит к более низкой корреляции между деревьями решений.

В алгоритме случайный лес деревья решений строятся по следующей схеме:

- Выбирается подвыборка на основе обучающей выборки размером `samplesize` и строится дерево (для каждого дерева своя подвыборка).
- Для построения каждого расщепления в дереве просматриваем `max features` случайных признаков (для каждого нового расщепления используются случайные признаки).
- Выбирается наилучший признак и расщепление по нему (по заранее заданному критерию). Дерево строится до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Алгоритм случайный лес основан на построении большого числа (ансамбля) деревьев решений (это число является параметром метода), каждое из которых строится по выборке, получаемой из исход-

ной обучающей выборки с помощью бутстрепа (т. е. выборки с возвращением). Каждое дерево решений строится на основе выборки, получаемой из исходной обучающей выборки **с помощью бутстрепа (т. е. выборки с возвращением)**. Классификация осуществляется с помощью голосования классификаторов, определяемых отдельными деревьями решений, а оценка регрессии производится путем усреднения оценок регрессии всех деревьев решений. Поэтому точность (вероятность корректной классификации) ансамблей классификаторов существенно зависит от разнообразия (diversity) классификаторов, составляющих ансамбль.

2 Достоинства и недостатки алгоритма случайный лес

Достоинства алгоритма:

- метод гарантирует защиту от переобучения (overfitting) в случае, когда количество признаков значительно превышает количество наблюдений;
- для построения случайного леса по обучающей выборке требуется задание всего двух параметров, которые требуют минимальной настройки (tuning);
- способность эффективно обрабатывать данные с большим числом признаков и классов;
- нечувствительность к любым монотонным преобразованиям значений признаков;
- обучающая выборка для построения случайного леса может содержать признаки, измеренные в разных шкалах: числовой, по-

рядковой и номинальной, что недопустимо для многих других классификаторов;

- случайные леса могут использоваться не только для задач классификации и регрессии, для выявления наиболее информативных признаков, кластеризации, выделения аномальных наблюдений и определения прототипов классов;
- внутренняя оценка способности модели к обобщению;
- высокая параллелизуемость и масштабируемость;
- случайные леса очень гибки и обладают высокой точностью.

Недостатки алгоритма:

- большой размер получающихся моделей;
- построение леса сложнее и отнимает больше времени;
- чем больше объем данных, тем сложнее интуитивное понимание результатов работы алгоритма;
- алгоритм склонен к переобучению при использовании зашумленных данных.