ДЕТЕКРИРОВАНИЕ ОБЪЕКТОВ НА ИЗОБРАЖЕНИИ НА ОСНОВЕ МЕТОДА ДЕРЕВЬЕВ РЕШЕНИЙ

Д.И. Коваль, К.В. Вик, Ю.А. Иванова
Томский политехнический университет
E-mail: dik9@tpu.ru

Введение

Классическое, общее (и не больно-то строгое) определение машинного обучения звучит так (T. Mitchell "Machine learning", 1997):

Говорят, что компьютерная программа *обучается* при решении какой-то задачи из класса *T*, если ее производительность, согласно метрике *P*, улучшается при накоплении опыта *E*.

Далее в разных сценариях под *T*, *P*, и *E* подразумеваются совершенно разные вещи. Среди самых популярных задач T в машинном обучении**:**

- классификация – отнесение объекта к одной из категорий на основании его признаков
- регрессия – прогнозирование количественного признака объекта на основании прочих его признаков
- кластеризация – разбиение множества объектов на группы на основании признаков этих объектов так, чтобы внутри групп объекты были похожи между собой, а вне одной группы – менее похожи
- детекция аномалий – поиск объектов, "сильно непохожих" на все остальные в выборке либо на какую-то группу объектов

Под опытом E понимаются данные (без них никуда), и в зависимости от этого алгоритмы машинного обучения могут быть поделены на те, что обучаются с учителем и без учителя (supervised & unsupervised learning). В задачах обучения без учителя имеется выборка, состоящая из объектов, описываемых набором признаков. В задачах обучения с учителем вдобавок к этому для каждого объекта некоторой выборки, называемой обучающей, известен целевой признак – по сути это то, что хотелось бы прогнозировать для прочих объектов, не из обучающей выборки [1].

Задачи классификации и регрессии – это задачи обучения с учителем. В качестве примера будем представлять задачу кредитного скоринга: на основе накопленных кредитной организацией данных о своих клиентах хочется прогнозировать невозврат кредита. Здесь для алгоритма опыт E – это имеющаяся обучающая выборка: набор объектов (людей), каждый из которых характеризуется набором признаков (таких как возраст, зарплата, тип кредита, невозвраты в прошлом и т.д.), а также целевым признаком. Если этот целевой признак – просто факт невозврата кредита (1 или 0, т.е. банк знает о своих клиентах, кто вернул кредит, а кто – нет), то это задача (бинарной) классификации. Если известно, на сколько по времени клиент затянул с возвратом кредита и хочется то же самое прогнозировать для новых клиентов, то это будет задачей регрессии.

Наконец, третья абстракция в определении машинного обучения – это метрика оценки производительности алгоритма P. Такие метрики различаются для разных задач и алгоритмов, и про них мы будим говорить по мере изучения алгоритмов. Пока скажем, что самая простая метрика качества алгоритма, решающего задачу классификации – это доля правильных ответов (accuracy, не называйте ее точностью, этот перевод зарезервирован под другую метрику, precision) – то есть попросту доля верных прогнозов алгоритма на тестовой выборке [2,3].

Задачи

Основная сфера применения деревьев решений — поддержка процессов принятия управленческих решений, используемая в статистике, анализе данных и машинном обучении. Задачами, решаемыми с помощью данного аппарата, являются [4]:

- Классификация — отнесение объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.
- Регрессия (численное предсказание) — предсказание числового значения независимой переменной для заданного входного вектора.
- Описание объектов — набор правил в дереве решений позволяет компактно описывать объекты. Поэтому вместо сложных структур, описывающих объекты, можно хранить деревья решений.

Основные этапы построения

В ходе построения дерева решений нужно решить несколько основных проблем, с каждой из которых связан соответствующий шаг процесса обучения:

1. Выбор атрибута, по которому будет производиться разбиение в данном узле (атрибута разбиения).
2. Выбор критерия остановки обучения.
3. Выбор метода отсечения ветвей (упрощения).
4. Оценка точности построенного дерева.

Реализация

Приложение написано на языке программирования высокого уровня Python для версии 3.x.

Для реализации задачи использовалось два метода:

- Метод главных компонент (PCA)
- ID3

Метод PCA - Метод главных компонент (англ. principal component analysis, PCA) —метод снижение размерности путем выделения n главных компонент.

ID3 — это один из наиболее популярных алгоритмов обучения деревьев решений. В основе идеи алгоритма лежит рекурсивное разбиение обучающего множества, размещаемого в корневом узле дерева решений, на подмножества с помощью решающих правил.

Разбиение продолжается до тех пор, пока в результирующих подмножествах не останутся примеры только одного класса, после чего процесс обучения остановится, а подмножества будут объявлены листами дерева, содержащими решения.

Каждый атрибут обучающего множества отражает некоторое свойство классифицируемых объектов. При этом атрибуты могут иметь разную значимость с точки зрения классификации. Например, атрибут, все значения которого одинаковы, вообще бесполезен для различия классов.

Классифицирующая сила других атрибутов может быть разной. Целью алгоритма является выбор атрибутов для разбиения таким образом, чтобы полученное дерево было компактным, простым для понимания и при этом достаточно точным.

Алгоритм начинает работу с корневого узла дерева, который содержит все примеры обучающего множества. На каждой итерации алгоритма выбирается один из атрибутов, по которому производится разбиение множества примеров в узле на подмножества. При этом для дискретных и непрерывных атрибутов процесс отличается [5,6].

Выборка

При обучении и тестировании сети использовалась выборка (рисунок 2) из 3000 изображений размером 1x1. Размер изображения 1 на 1 был выбран, исходя из того, что для изображений большего размера необходимы большие мощности вычислительных устройств.

Рис. 1. Пример изображений из выборки.

Для обучения использовалось 2000 изображений 60% от всей выборки. Общее количество элементов, используемых в тестовой выборке, составило 1000 изображений, оставшиеся 30% изначальных данных.

Размерность входных изображений может быть различной, и от размера изображения будет зависеть только скорость вычислений. Так как с помощью дерева большие изображения обрабатываются долго, мы меняем размер на 1x1.

Основным недостатком дерева является то, что данный метод машинного обучения не подходит для решения поставленной задачи. Соответственно, точность классификации составляет порядка 50 %.

Заключение

Также низкий показатель метрики качества работы алгоритма обусловлен тем, что для обучения использовались изображения размером 1x1 для ускорения вычислений.

Для проверки, насколько данный алгоритм является уместным применительно к этой задаче, был использован метод снижения размерности PCA. Все пространство признаков было снижено до 2 компонент и размещено на двумерной плоскости. На рисунке таком-то можно увидеть, что пространство признаков не является линейно разделимым, соответственно данная задача не может быть решена с помощью простых методов машинного обучения.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-08-00977 А и в рамках Программы повышения конкурентоспособности ТПУ.

Список использованных источников

1. Деревья решений в задачах распознавания образов. URL: https://www.dissercat.com/content/derevya-reshenii-v-zadachakh-raspoznavaniya-obrazov
2. Алгоритм ID3. URL: https://wiki.loginom.ru/articles/algorithm-id3.html
3. Метод главных компонентов. URL: https://habr.com/ru/post/304214/
4. Обучающая выборка. URL: https://www.kaggle.com/
5. Классификация. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F
6. Машинное обучение. URL: https://habr.com/ru/post/319288/