Киреев Руслан

Отчет БДЗ

1. В качестве исходных данных были использованы данные по автомобилям (http://archive.ics.uci.edu/ml/machine-learning-databases/car/) , которые были разбиты на два класса: positive(acc, good, vgood) и negative(unacc). Они составляют 30% и 70% от общего числа объектов соответственно. Применялось естественное шкалирование, уже реализованное в представленном алгоритме. То есть каждое значение атрибута преобразовывалось в новый признак.

2. Использовался метод на основе генераторов. Характеристика, участвующая в итоговом правиле классификации, – поддержка. Агрегатная функция – сумма по всему «+» контексту и отдельно по «-» контексту. Чтобы найти пороговые значения поддержки, допускающие некоторую ошибку на обучении, мы проходим сначала по тестовой выборке и находим средние значения отдельно по положительным и отрицательным примерам. В итоге, получаются две пары чисел: средние значения агрегатной функции по положительным примерам и средние значения по отрицательным примерам. Для данных об автомобилях эти значения оказались следующими:

- Положительный пример: среднее значение агрегатной функции на положительном контексте: 0.313, на отрицательном: 0.158.
- Отрицательный пример: среднее значение агрегатной функции на положительном контексте: 0.192, на отрицательном: 0.276.

В качестве решающего правила выступило соображение, к каким значениям (положительного или отрицательного) ближе значения функции текущего примера. Например, если оба числа одновременно ближе к 0.313, чем к 0.192 и к 0.158, чем к 0.276, то пример классифицируется как положительный. Аналогично с классификацией отрицательного примера.

Для проверки качества классификации использовалась 10-fold cross validation. То есть обучение проходит на 90% от данных (1557 объектов) и тестирование на 10% (172 объекта). И в итоге усреднялось по 10 запускам. В каждом запуске находилось число правильно определенных примеров, деленное на общее их число. Для описанного выше правила средняя точность(accuracy) 79.2%. Что на 9% больше, чем в случае, если мы бы, например, просто определили бы все примеры негативными (70%). Следовательно, классификация работает!

Так же пробовался вариант, когда выставлялся порог минимальной поддержки. В таком случае наиболее подходящими оказались значения 0.27 для положительной классификации и 0.23 для отрицательной (находились перебором с шагом 0.1, начиная со средних значений, шаг большой, потому что алгоритм медленный). В этом случае результаты cross validation около 77%

3. В качестве одной из модификаций предлагается корректировать эти средние значения непосредственно во время тестирования. Если пример определяется как положительный, но на самом деле он отрицательный (можно создать интерфейс взаимодействия с экспертом, который по признаковому описанию объекта отвечает на вопрос, к какому классу он принадлежит, но на деле тестовая выборка уже промаркирована), то этот отрицательный пример добавляется к отрицательным примерам обучающей выборки, и пересчитывается соответствующее среднее значение. Аналогично для случая, когда положительный пример определяется как отрицательный. С учетом этой модификации результаты кросс валидации составили 79%, то есть почти не изменились. Чтобы удостовериться в обоснованности этого метода необходимо тестировать его на различных данных, возможно, где-нибудь эти небольшие флуктуации параметров окажутся критичными.