Московский Государственный Университет имени М.В.Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Отчет по решению реальной задачи

“Topical Classification of Biomedical Research Papers”

Выполнила: студентка 317 группы

Лобачева Екатерина

Москва

2012

Постановка задачи

Перед нами поставлена задача классификации медицинских статей по рубрикам. Каждая статья описывается 25640 признаками, принимающими значения от 0 до 1000 (предположительно это частота встречаемости некоторых терминов), и может относиться к одной или нескольким рубрикам (всего их 83). Дана обучающая выборка из 10000 объектов и контрольная, также состоящая из 10000 объектов. Качество классификации оценивается по f-мере.

Ход решения

1. Предобработка признаков

В первую очередь я удалила все признаки, которые на обучающей выборке принимают только нулевые значения, так как они не могут внести осмысленного вклада в классификацию.

Далее были испробованы различные нормировки признаков по строкам и столбцам, а именно:

- на сумму элементов,
- на среднее значение,
- на среднее значение ненулевых элементов,
- на максимальное значение,
- на некоторую константу.

Так же рассматривались варианты сокращения признакового пространства за счет признаков, которые отличны от нуля на очень маленьком количестве объектов.

Как ни странно, но наилучшего результата мне удалось достичь при нормировке на константу (однако разница в наилучших результатах, которых удалось достичь с нормировкой на константу, с нормировкой на сумму элементов строки и с нормировкой на максимальный элемент столбца отличаются очень незначительно – не более чем на 2%).

2. Алгоритмы

В основном я работала с линейным алгоритмом. Так же был испробован метод ближайших соседей, однако он работал сильно дольше, а сильно лучших результатов добиться группе не удалось, поэтому я сосредоточилась на линейном алгоритме.

Для классификации использовались 83 независимых линейных классификатора one-vs-all. Параметры выбирались с помощью CV.

После настройки параметров я решила учесть взаимосвязь между рубриками. Так как классификация по рубрикам происходит независимо, я решила попробовать

использовать ответы, полученные при классификации для предыдущих рубрик, как дополнительные данные при классификациях для последующих. Однако этот подход только все ухудшает из-за появления шума по причине неточной классификации, поэтому я несколько модифицировала данный подход. Для этого я запустила свой алгоритм на нескольких случайных подвыборках и посчитала среднюю ошибку алгоритма на каждом классе. Далее отсортировала классы в порядке увеличения ошибки и отобрала «цепочку» тех, в которых ошибка не превышала некоторого порога.

Классификация тестовых объектов делилась на два этапа:

- Классификация для классов из «цепочки»: классификация производилась в порядке встречаемости классов в «цепочке», при этом каждому последующему классификатору выходы предыдущих подавались как дополнительные признаки.
- Классификация для остальных классов: в качестве признаков подавались описания объектов и их классификации на классах из «цепочки».

Такой подход к учету связи рубрик уже дает некоторое улучшение к результату классификации.

Объекты, которые линейный алгоритм не отнес ни к одной из рубрик, относились к наиболее часто встречаемым рубрикам.

Финальное решение

Финальное решение я в общем-то не выбирала, его за меня выбрал сайт, подсчитывающий промежуточный результат. Оно представляет собой следующую последовательность действий:

- Удаление нулевых столбцов.
- Нормировка данных – деление всех признаков на константу 1200.
- 83 независимых линейных классификатора из библиотеки liblinear со следующими параметрами: '-s 3 -c 0.0393 -B 1 -w1 3.75689'.
- При этом использовалась описанная выше процедура учета зависимости рубрик с порогом 0.5%, в которой «цепочка» имела следующий вид: 77, 33, 31, 16, 10, 37, 72, 55, 53, 27, 3, 7.
- Объекты, которые линейный алгоритм не отнес ни к одной из рубрик, относились к 5 наиболее часто встречаемым рубрикам: 40, 41, 44, 18, 62.

Данное решение дало предварительный результат 0.518 и финальный 0.520.

При необходимости могу предоставить все коды функций в Matlab.

Советы новичкам

- Не откладывайте на завтра то, что можно и нужно делать сегодня.

- Обсуждайте задачу с друзьями в непринужденной обстановке – это способствует появлению множества интересных идей.
- Получайте от решения удовольствие, а не гонитесь за результатами, тогда вы попробуете много разных методов и многому научитесь, а не будете нудно подбирать параметры.

Что было бы, если бы…

Я бы последовала своим советам из предыдущего пункта.

Отзывы и предложения

Сама задача мне понравилась – это новый и полезный опыт. Однако я считаю, что необходимо обсуждение идей на семинарах в устной форме как с одногруппниками, так и с преподавателем. Тогда мы бы узнали намного больше интересных идей и подходов к решению такого рода задач. А так, получилось что задачу в основном все решили довольно стандартными методами и кроме опыта в решении именно реальной задачи ничего интересного то и не было :-(

Работа в группе

Я не достаточно ответственно подошла к заданию, поэтому группе ничем конструктивным не помогла, однако оказывала моральную поддержку некоторым ее членам.

Мне же помогли следующие люди:

- Петр Ромов – ему вообще отдельная благодарность и вкусняшка с меня ☺,
- Потапенко Анна,
- Огнева Дарья,
- Евгений Зак.