Московский Государственный Университет имени М.В.Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Отчет по заданию

«Topical Classification of Biomedical Research Papers»

Остапец Андрей Александрович
группа 317

Москва,2012

Постановка задачи

Задача - классифицировать научные медицинские статьи по 83 пересекающимся рубрикам.

Дана обучающая матрица с ответами, а также имеется матрица, для которой ответы необходимо предоставить.

Каждая статья описывается 25640 признаками. В обучающей матрице 10000 статей.

Аналогичный размер имеет тестовая выборка.

Каждый признак принимает значения из диапазона 0..1000, причем подавляющее число признаков - нулевые.

Для каждой статьи в обучающей выборке известно, к каким рубрикам она относится.

Необходимо найти к каким рубрикам относятся объекты из тестовой матрицы.

Ход решения

Когда я приступил к решению задачи, то первым делом решил попробовать библиотеку liblinear, ссылку на которую дал и предоставил краткое описание работы с ней Петя Ромов.

Неожиданно очень быстро удалось достичь высокого результата, поэтому вся дальнейшая работа была связана, в основном, с подбором оптимальных параметров и настройкой линейных классификаторов.

Также была проведена работа с методом ближайших соседей, который долгое время использовался для классификации тех объектов, для которых линейные классификаторы не выдали ни одной рубрики.

Помимо этого, были попытки использовать систему WEKA для решения данной задачи. Поскольку данная система имеет ограничения на объем используемой памяти, то для каждой рубрики отбирались 500 признаков (брались те признаки, для которых линейный классификатор давал наибольший вес по абсолютному значению) и затем на новой матрице проводился запуск всех доступных там алгоритмов. Увы, высоких результатов это не принесло.

Финальное решение

Финальное решение можно разбить на 4 части:

1) Удаление из матриц тех столбцов, которые в тренировочной матрице полностью нулевые.

2) Нормировка тренировочной матрицы построчно по формуле:

$$X_i = \alpha \sum_{j=1}^{k} X_i^j + \beta \frac{\sum_{j=1}^{k} X_i^j}{\sum_{j=1}^{k} [X_i^j > 0]}, i = 1..m,$$

где $X_i -$ ая строка, $X_i^j -$ элемент в позиции $(i, j), X \in \mathbb{R}^{m \times k}$.

В итоге, параметры были выбраны следующим образом: $\boldsymbol{\alpha = 0.005}$, $\boldsymbol{\beta = 0.5}$.

3) Использование 2-шагового линейного классификатора:

Для каждой рубрики решения о классификации принимаются независимо.

-На первом шаге - обучение первого классификатора, отдельно для каждой из 83 рубрик, с параметрами '**-s 1 -c 0.001 -B 1.1 –e 0.08'.**
Затем удаление из обучения для данной рубрики всех тех объектов, который не принадлежали данной рубрики, но классификатор на них ошибся.
-Затем обучение второго классификатора для каждой из 83 рубрик с параметрами '**-s 1 -c 0.001 -B 1.07 –e 0.08'** на уже подправленной тренировочной выборке.

4) Те документы, которые не были отнесены ни к одной рубрики, относились к 2 рубрикам, для которых классификатор дал значение, наиболее близкое к положительному значению.

Советы новичкам

- Использовать готовые библиотеки для фундаментальных алгоритмов, а не писать что-то свое.
- Находить и читать литературу по тематике конкурса.

Выводы

В итоге, реальная задача мне понравилась намного больше, чем решаемые до этого "искусственные". В реальной задаче некоторые вещи, которые кажутся странными, и на первый взгляд не должны работать, неожиданно приносят улучшения, и бывает наоборот, что-то, что должно хорошо работать, на самом деле не работает. И соревнование помогает добиться более высоких результатов – ты видишь, что соперник тебя обошел, и начинаешь что-то улучшать в своем алгоритме, пытаешься добиться более высокого результата.

Благодарности

Выражаю огромную благодарность Петру Ромову. Им проделана огромная работа в самом начале решения задачи – подготовка страницы, описание задачи на русском языке, код для отправки результатов на сайт. А также очень признателен Евгению Нижибицкому и Дмитрию Кондрашкину за совместную работу в конце конкурса, если бы наша команда просуществовала больше времени, то мы бы точно обошли словенцев!