Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Отчёт о решении реальной задачи «Topical Classification of Biomedical Research Papers»

Выполнила:
Морозова Дарья 317

2012

Постановка задачи:

Была предложена задача классификации в рамках конкурса “JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers”. Нужно было определить, к каким классам (тематикам) относятся объекты (научные статьи). Были даны 83 пересекающихся класса и 25640 признаков, где каждый признак показывает, насколько сильно научная статья связана с данным медицинским термином. У каждого объекта большинство признаков были равны 0, а это значит, что такой объект связан лишь с небольшим числом медицинских терминов.

Ход решения и финальная реализация:

Я реализовала в среде MATLAB метрический классификатор. Сначала я убрала нулевые и ненужные признаки. Потом применила метод Парзеновского окна к этой реальной задаче. В результате экспериментов я выбрала оптимальные параметры (например, тип метрического алгоритма, функцию ядра K, высоту h). У меня получился вектор из 0,1 и ‘ ‘. ‘ ‘ возникало тогда, когда оценка принадлежности объекта к классу 0 совпадала с оценкой принадлежности объекта к классу 1. В ходе исследования я обнаружила, что эти оценки равны 0, а это значит, что объект расположен далеко от любого класса. Для таких 109 объектов (1-104, 107, 110-112, 316) я далее выбрала классы, которые встречаются чаще для этих объектов (18, 40, 41, 44, 62). Программа должна была работать около недели, поэтому я решила пробегать цикл не по всем объектам, а по сотне объектов, в которых обязательно встречалась бы и ‘1’, и ’0’. Мне удалось сократить время до ~40 минут. В итоге я получила результат 0.10570.

Что я ещё могу предоставить по решению этой задачи:
-код
-более подробное объяснение
-отчёты о результатах экспериментов

Советы новичкам:
Не стоит бояться объёма задания, масштабов матриц и способностей конкурентов, а стоит действовать.

Что я узнала нового:
Что как всегда неэффективно расходую время, а именно уделяю много времени на реализацию методов, плохо решающих данную задачу. Таких методов, за которые не берутся остальные студенты.

Что могла бы ещё сделать при наличии времени:
Я могла бы попробовать решить задачу другими методами, воспользоваться помощью ребят со страницы “Обсуждение” и оптимизировать их код, подобрать параметры.

Какие задания по практикуму хотела бы я ещё выполнить?
Я хотела бы решить аналогичную задачу. Уделить ей больше времени и добиться лучшего результата.

Мне принёс максимальную пользу:
-Пётр Ромов с понятным разъяснением задания и хода решения.
-Новиков Максим с его кодом по преобразованию файлов из Matlab в ARFF/XML.
-Женя Нижибицкий и Анна Потапенко с их идеями по удалению ненужных признаков.
-Андрей Остапец с примером работы с LIBLINEAR

Список используемой литературы:
-“Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RAPIDMINER и MATLAB” Дьяконов А.Г.
-страничка “Обсуждение”
- http://www.machinelearning.ru/

Как я ещё пыталась решить это задание:

Я попыталась решить эту реальную задачу с помощью систем WEKA и RapidMiner. В WEKA мне удалось загрузить разреженные матрицы, но она падала через некоторое время каждый раз. В RapidMiner я реализовала схему, по которой я получала матрицу, в которой последний столбец показывал, принадлежит ли данный объект к определённому классу (0 и 1). И загрузив данные, мне удалось получить огромную нужную матрицу. Но так как я выбрала не слишком удачный классификатор, результаты меня не порадовали.