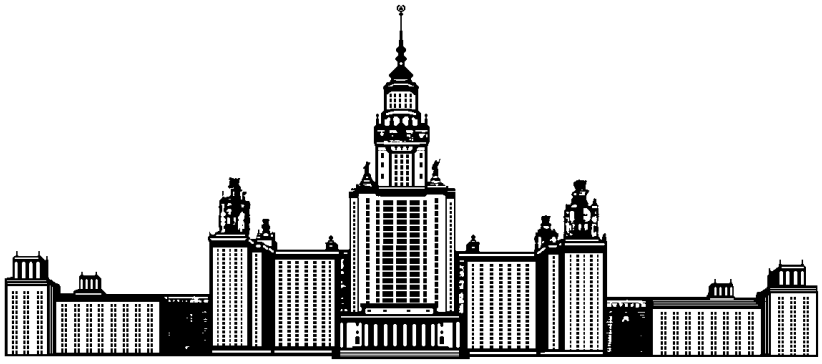
Московский Государственный Университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

JRS 2012 DATA MINING COMPETITION: TOPICAL CLASSIFICATION OF BIOMEDICAL RESEARCH PAPERS

Отчет по участию в конкурсе

Нижибицкий Евгений

Группа: 317

Москва 2012

Постановка задачи

Задача состояла в автоматической рубрикации медицинских текстов на основе предоставленных факторов «силы связанности» статей с каждым из 25640 терминов (в числовом выражении каждый термин имел «связь» от 0 до 1000 с каждым документом), определенных неким алгоритмом заранее.

Необходимо было проставить рубрики (всего таковых было 83) для выданных 10000 документов, для каждого из которых необходимо было указать как минимум одну — сверху ограничений не было.

Для оценки итогового алгоритма была использована **F-macro** мера.

Данные были представлены в виде csv-файлов.

Начало исследования

Предобработка данных

Данные были получены сразу в используемом виде (`.mat`-файл с разреженными матрицами вектора обучения, контрольные вектора и матрица ответов для первых) благодаря Петру Ромову. В дальнейшем из них были удалены столбцы, соответствующие нулевым оным в тренировочной матрице.

Предпосылки к использованию алгоритмов

Так как задача ставится как классификация текстов, первое, что пришло в голову попробовать, это применить стандартные для таких задач алгоритмы классификации — классическое **TF×IDF** преобразование матрицы данных и центроидный алгоритм на них. Но ввиду секретной предобработки данных они были мало похожи на реальные данные о количестве тех или иных слов, поэтому такие алгоритмы каких-либо значительных успехов не имели.

Исходя из постановки задачи возникает идея использовать линейный классификатор, т.к. вообще все координаты вносят пропорциональный вклад в рубрикацию по той или иной тематике (Чем сильнее документ связан с термином «перелом», тем вероятнее он относится к хирургии, и менее вероятно к стоматологии).

Так же, скорее всего, выполняется гипотеза компактности — статьи на схожие тематики, скорее всего, содержат схожие наборы терминов — следовательно, вполне логично попробовать метрические алгоритмы.

Метрические алгоритмы

Наилучшие результаты были достигнуты алгоритмом **k-NN** при **k** равном около 40 и квадратичных весах ($w_i=i^2$). Максимум — **0.473** на локальной выборке.

Так как довольно быстро лучшие результаты получались использованием линейных классификаторов, в дальнейшем метрические алгоритмы использовались лишь в качестве «затычки» для остальных — они запускались на тех векторах, на которых линейные алгоритмы отказывались от классификации по всем 83 рубрикам.

Линейные классификаторы

Несмотря на пессимистичные заявления в обсуждении сего задания на соответствующей странице на сайте `machinelearning.ru` было решено попробовать различные нормировки данных и подбор параметров для классификатора `liblinear`.

Нормировка данных

Лучше всего показали себя нормировка по максимуму в строке и нормировка по среднему арифметическому ненулевых (или больших некоего порога) координат столбцов. Соответствующие результаты были порядка **0.508** и **0.520**.

Синтез алгоритмов

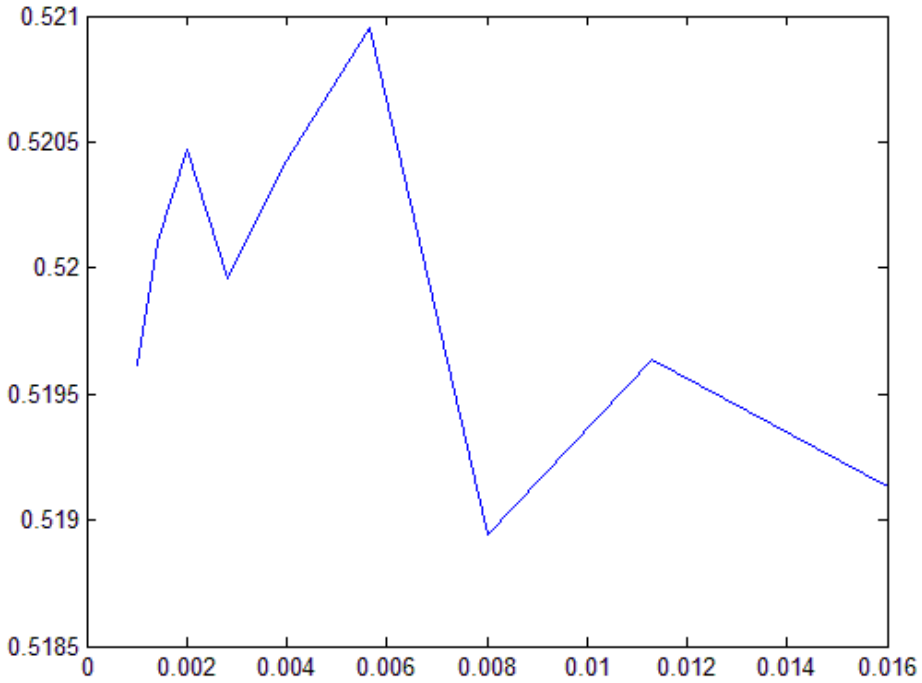
На последнем этапе конкурса было решено на базе моего аккаунта сделать команду с Андреем Остапцом и Дмитрием Кондрашкиным для дальнейших улучшений имеющихся алгоритмов и реализации на их основе новых.

Наилучшим результатом до этого момента была планка в **0.525**. Через полчаса после объединения был получен результат **0.528** благодаря «*инновационным решениям в области поиска параметров*» автора — алгоритм был запущен на сетке 4×4 с варьированием параметров **C** и **b** в линейном классификаторе, и были найдены лучшие параметры, давшие улучшение результата.

В дальнейшем предпринимались попытки использования не непосредственно ответов классификатора, а т.н. *decision values*, т.е. степени уверенности классификатора в принадлежности данного объекта к первому классу.

Рассматривались линейные комбинации весов, и на основе них уже принимались решения о принадлежности к тем или иным рубрикам.

На локальной выборке получались улучшения порядка **+0.002** на кросс-валидационной проверке, но на итоговых результатах это не отражалось (проверялись лишь 10% данных — не самых, видимо, репрезентативных):



подбор коэффициента для одного из двух алгоритмов в линейной комбинации

Результаты

Наилучший алгоритм в итоге был основан на алгоритме Андрея с 2 стадиями классификации — после первой из выборки выкидывались объекты, на которых наш алгоритм ставил слишком много меток на ней же. Работа также велась не с непосредственно выводом классификатора, а теми самыми *decision values*.

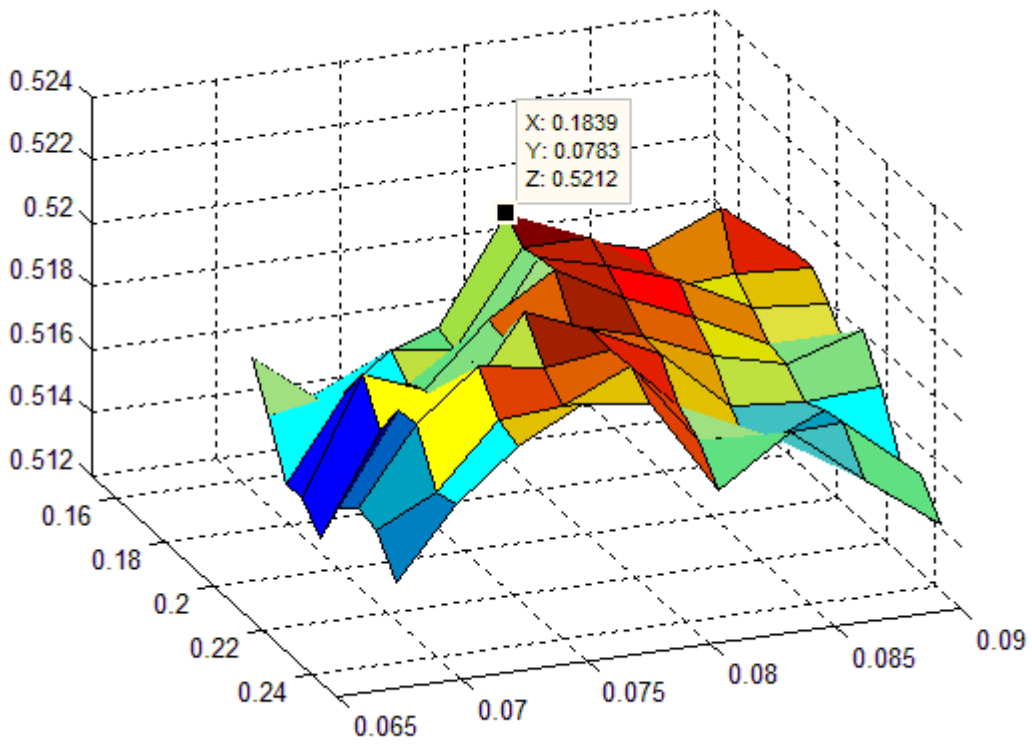
Предварительный результат (на 10%) — **0.531**.

После опубликования финальных итогов — **0.52939**.

Все исходные коды опубликованы на сайте `machinelearning.ru`.

Советы новичкам

- Продумывать структуру файлов, разделение функций, чтобы можно было независимо менять нормировки/способ работы непосредственно с классификатором/способ постобработки полученных результатов.
- Тем не менее, периодически делать что-то вроде “milestones” — при получении какого-то результата собирать все файлы воедино в некий скрипт *result0531.m*, чтобы можно было позже вернуться к старым идеям и взглянуть на них по-новому.
- При тестировании параметров использовать графическое представление получаемых результатов — например, рассматривать `meshgrid` с центром в текущих параметрах и варьировать два параметра по его точкам для того, чтобы понять направление дальнейших поисков, если таковые разумно еще делать. При достижении устойчивого максимума следует снова продумывать дальнейшие уже структурные улучшения алгоритма.



пример подбора параметров **C** и **B**

Впечатления

- Природа данных была довольно странной — от текстов как таковых там ничего не осталось. Вследствие этого трудно было придумать алгоритм «на идею», все уловки были лишь чисто техническими. Единственный такой алгоритм был представлен Александром Геннадьевичем (предположение о существовании простого линейного отображения из пространства данных в пространство ответов), но и это идейное решение суть из общих соображений, а не данной задачи ☹

- Репрезентативность количества и качества данных, на которых проводились предварительные замеры на сайте, вызывает сомнение — расхождения с хорошо просчитанным (t×q-fold кросс-валидация) локальным результатом достигали 2%.

- Тем не менее, этот конкурс был хорошим поводом познакомиться с темой классификации с множественными ответами и темой классификации текстов в частности — автором за время его проведения было прочитано около 30 статей на данную тематику. К сожалению, для многих из них требовались либо преобразования непосредственно текстов, либо большие вычислительные мощности при применении них к этим данным (была нехватка памяти при отборе признаков, либо требовались ядра для SVM, что влекло за собой использование на порядки более медленной библиотеки `libsvm`).

Послесловие

Выражаю благодарность Петру Ромову за предоставленные данные в удобоваримом виде, а также Андрею Остапцу и Дмитрию Кондрашкину за плодотворное сотрудничество на заключительной стадии конкурса.

Из собственных же попыток помочь команде могу выделить выкладывание кода для удаления ненужных признаков, реализацию центроидного алгоритма, исправление кода конвертации в arff-формат для данной задачи, реализацию идеи выделения иерархии среди рубрик, периодические *викификации* страницы обсуждения для удобства последнего.