Московский Государственный Университет им. М.В. Ломоносова
факультет Вычислительной математики и кибернетики
кафедра Математических методов прогнозирования

Отчёт по решению реальной задачи «Topical Classification of Biomedical Research Papers»

Выполнила: *Огнева Дарья,*
317 группа

Преподаватель: *Дьяконов Александр Геннадьевич*

2012

1. Постановка задачи.

Первоочередная задача состояла в применении полученных теоретических знаний по методам классификации к решению реальной задачи. Для этого было выбрано соревнование по классификации медицинских текстов[1].
В состязании каждый объект представлен журнальной статьёй, описывающейся 25640 признаками: целыми числами 0...1000, означающими, насколько сильно журнальная статья связана с соответствующим медицинским термином. Тематик статей 83. С помощью тренировочной выборки объёмом 10000 (матрица 10000 x 25640) необходимо классифицировать тестовую выборку также объёмом 10000. Ответом для каждого объекта - подмножество целых чисел 1..83.
Таким образом, было предложено решить задачу классификации текстов с 83 пересекающимися классами.

2. Ход решения.

2. 0. Статьи.

Изначально надо было ограничить круг методов решения задачи классификации текстов и теоретически понять их плюсы и минусы при решении общих задач.
Список литературы:

- Machine Learning in Automated Text Categorization. Fabrizio Sevastiani
- Text Classication from Labeled and Unlabeled Documents using EM. Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell
- Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Thorsten Joachims
- Improved Use of Continuous Attributes in C4.5. J. R. Quinlan
- Text Classification using String Kernels. Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, Chris Watkins
- Text Categorization, Fabrizio Sebastiani
- Логические алгоритмы классификации, К. В. Воронцов
- Package ‘randomForest’

2. 1. Среда и основной метод.

Первоначально была предпринята попытка произвести работу по анализу данных. Для этого были созданы arff-файлы и произведена коррекция README.txt для возможности работы системы Weka с большими объёмами данных. Затем был запущен алгоритм kNN со стандартным набором параметров. Вероятней всего вследствие большого объёма данных, алгоритм проработал сутки, подвесив систему. Таким образом, была установлена необходимость решать задачу в классических средах, подготовленных к большим объёмам разреженных данных.

В качестве базовой была выбрана среда МатЛаб. Зная, что время выполнения встроенного kNN больше 4 часов, что достаточно неприятно для подбора коэффициентов, а написанные циклы в МатЛабе будут выполняться не сильно

быстрее, за основной метод был взят SVM, эффективно реализованный в библиотеке liblinear.

2. 2. Предобработка данных.

Начальная обработка данных состояла в удалении нулевых признаков обучающей выборки и соответствующих столбцов в контрольной.
Затем были предприняты преобразования каждого столбца путём деления столбца на :

1. значение максимального элемента столбца
2. значение суммы элементов столбца
3. значение медианы столбца
4. значение медианы ненулевых элементов столбца
5. значение евклидовой метрики столбца (корень из суммы квадратов элементов)

В результате экспериментов с каждым из преобразований соответственно:

0 (без преобразования)	1	2	3	4	5
0.4258	0.4071	0.2790	0.3446	0.4285	0.4392

По итогам экспериментов по изменению типа классификаторов два лучших преобразования остались неизменны. В результате была выбрана “евклидова нормировка”, хотя для самоконтроля результаты экспериментов повторялись и для медианы ненулевых элементов столбца.

2. 3. Использование ответов классификации.

Так как классификация с 83 пересекающимися классами была реализована как последовательные 83 классификации с 2 непересекающимися классами, то была возможность опробовать эвристику использования уже полученных результатов классификации для классификации последующих классов.
Для этого была составлена матрица условных вероятностей по описанному в “Обсуждении” коду. Классы сортировались по возрастанию максимальной вероятности встречи любого другого класса по данному.

```
% get probability matrix sup and returns the class order
function [order] = orderY (sup)
order = max(sup, [], 2)';
[order, ~] = sort(order);
end
```

Проведённые эксперименты с различными вариантами добавления к признаковому пространству векторов ответов: добавление сразу после классификации, добавление только в классификацию начиная с некоторого номера, - показали, что в связи с небольшой точностью полученных векторов, качество классификации падает на 0.01-0.04.

2. 4. Подбор коэффициентов.
Первый коэффициент в функции train -s -- тип классификатора:
при двух различных наборов остальных параметров:

-s0	-s1	-s2	-s3	-s4	-s5	-s6	-s7
0.4493	0.3945	0.4292	0.3792	0.3896	0.3967	0.4147	0.4224
0.4701	0.4021	0.4381	0.4183	0.4135	0.3998	0.4157	0.4296

Таким образов, для последующих попыток классификации использовались s0 и s2.

Эксперименты по остальным параметрам показали сильную зависимость между ними. Таким образом, для качественного подбора необходимо было строить четырёхмерную сетку.

В связи с ограниченностью по времени на данном этапе были опробованы несколько десятков вариантов, и выбран лучший: - c 10 -B 10 -w0 1 -w1 10

3. Финальное решение.
В конечном итоге из начальных данных были удалены нулевые столбцы тренировочной выборки и им соответствующие в тестовой, каждый из оставшихся столбцов был поделен на корень из суммы своих элементов (с проверкой на 0 на тестовой выборке). Затем был запущен линейный классификатор с параметрами ‘-s0 -c10 -B10 -w0 1 -w1 10’.
Предварительный результат: 0.453.
Итоговый результат: 0.47036.

4. Выбор финального решения.
Финальное решение было выбрано как лучшее по скользящему контролю(10:1).
Несмотря на сравнительно низкий результат 0.453 на тестовой подвыборке, итоговый результат в точностью совпал с оценкой по F-мере на кросс-валидации - 0.470. В результате, методика сравнения алгоритмов была оправдана.

5. Материалы по решению.
В ходе решения задачи были получены arff-файлы, содержащие полную матрицу признаков и каждый класс в отдельности; коды нормировок (преобразований) признаков и упорядочения классов для последующего их использования в классификации; код финального решения.

6. Советы новичкам

- Начните решать задачу с первого дня. Ибо любое дело занимает больше времени, чем вы предполагаете. (Закон Мерфи.)
- Попробуйте всё. Лучше попробовать и сожалеть, чем сожалеть, что никогда не пробовал.
- Мыслите нестандартно. Только те, кто предпринимают абсурдные попытки, смогут достичь невозможного. (Альберт Эйнштейн.)
- Будьте открыты для общения. Учиться можно и нужно даже на примере конкурентов. Люди склонны хвалиться своими достижениями и довольно часто рассказывают вещи, до которых сам можешь не додуматься. (Олег Бойко.)
- Не забывайте о маленьких радостях. Если программа работает больше получаса - прилягте поспать или перекусите. Получайте удовольствие от процесса.

7. Новое. И немного о желаниях.

Открытием при работе над задачей было наличие качественно написанных всех алгоритмов классификации, которые только могли быть применимы к решению данной задачи. Таким образом удалось установить и разобраться в библиотеке liblinear.
Узнала о способе уменьшения признакового пространства с помощью сингулярного разложения матрицы (команды svd в МатЛабе).

При наличии времени более качественно подобрала коэффициенты, попробовала другие эвристики преобразования признаков и объектов, разобралась с пакетом “randomForest” в системе R.

Для реализации этих планов хотелось бы ещё решить реальную задачу по практикуму. Practice makes perfect.

8. Работа в группе.
1. Помощь группе.
Публикация ссылки на статью по классификации текстов (с акцентом на способы

уменьшения размерности пространства признаков) и на пакеты для работы с Random Forest.

2. Помощь группы.

Благодарю всех участников обсуждения, чьи коды и идеи помогли в достижении результата: Пётра Ромова(!), Евгения Нижибицкого, Дмитрия Кондрашкина, Андрея Остапца, Максима Новикова, Ильдара Шаймарданова, Марию Любимцеву. За аппаратную поддержку: Евгения Зака, Пётра Ромова и Ильдара Шаймарданова. За моральную поддержку: Екатерину Малышеву и Екатерину Лобачёву.

Ссылки:

[1] **JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers**