ВМК МГУ
Кафедра математических методов прогнозирования

Отчет о выполнении задания 5 «Topical Classification of Biomedical Research Papers — решение реальной задачи»

Студент: Исмагилов Тимур Ниязович

2012

Постановка задачи

Требуется решить реальную задачу «Topical Classification of Biomedical Research Papers».

Это задача классификации медицинских статей с 83 пересекающимися классами. Каждая статья описана 25640 целочисленными признаками, обучающая и контрольная выборки состоят из 10000 статей.

Оценка качества производится с помощью F-меры.

Ход работы

0. Я решил полениться и дождаться, пока по результатам одногруппников станет более-менее очевидно, какой именно класс алгоритмов следует разрабатывать, и затем уже работать только над этим классом. Лучше тратить временной ресурс на получение тех сведений, которыми со мной делиться никто не будет.

1. По полученным одногруппниками результатам более-менее очевидно, что следует пытаться доработать прежде всего линейный классификатор. Обучив разные алгоритмы с разными методами нормировки данных, я выбрал L2-reg L2-loss SVM с параметрами c = 0.00000000591, w0 = 1, w1 = 4.1.
   Такой алгоритм позволяет добиться результата в 0.52.
   Теперь следует искать способы улучшить этот алгоритм.
2. Шум. SVM, как известно, плохо работает с зашумленными данными.
   Был опробован следующий алгоритм: из выборки удалялись все объекты, которые классифицировались достаточно плохо.
   Положительных результатов достичь не удалось.
3. Преобразование значений признаков. Руководствуясь той логикой, что значение признака 10 может быть ближе к 1000, чем к 0, были опробованы преобразования признаков *value = [value ~= 0]* и *value = [value ~= 0]*(500 + value].* Алгоритмы были обучены заново.
   Положительных результатов достичь не удалось.

   Я решил рассмотреть в качестве приоритетного направления исследования те линейные классификаторы, где для классификации принадлежности каждому классу используется свой набор параметров. Далее — различные способы обучения этого набора.
4. В качестве оценки набора использовалась F-мера для столбца, соответствующего классу. Результат получился 0.5 — хуже, чем у алгоритма из пункта 1), то есть алгоритм был сильно переобученным.
5. В качестве оценки набора использовалась F-мера для данных, соответствующих правильному ответу, но с одним столбцом, замененным на результат классификации. Результаты получились плохими.
6. Были рассмотрены также разные гибридные методы, но они тоже не дали результатов.

Я предположил, что, вероятно, алгоритм из пункта 4) не совсем ужасен (все же 0.5 — это не так уж и плохо), и переобучение происходит на классах, для которых мало примеров объектов, им принадлежащих. Так что были рассмотрены гибридные алгоритмы, где для разных классов ответ давали разные алгоритмы. Далее следует конкретное описание разбиений:

7. Ответ для большинства классов давал алгоритм 1), но для тех классов, которые хорошо классифицировались алгоритмом 4), ответ давал он. Алгоритм получил результат в 0.522.
8. Я жадным образом заменял результаты классификации алгоритма 1) на результаты алгоритма 4), пока F-мера улучшалась. Аналогичную замену столбцов я выполнил для контрольной выборки.
   Положительных результатов достичь не удалось.
9. В алгоритме 8) помимо столбцов алгоритма 4) я добавил возможность использовать нулевые столбцы и столбцы, полученные алгоритмами, похожими на 1), но с немного другими параметрами. Положительных результатов достичь не удалось.

10. Я попробовал удалить признаки, которые были ненулевыми только у небольшого количества объектов.
    Положительных результатов достичь не удалось.
11. Возможно, имеющиеся 10000 текстов уже дают адекватное представление о том, какими бывают тексты в целом? Я попробовал заменить каждый результат классификации алгоритма 1) на ближайший из списка правильных ответов для обучающей выборки в соответствии с нехитрой метрикой, почти равной манхэттенскому расстоянию между векторами.
    Положительных результатов достичь не удалось.
12. Мне примерно известно, с какой частотой документ принадлежит тому или иному классу. Может быть, следует подгонять параметры, полученные при обучении алгоритма 4), чтобы эти значения совпадали? Я делал это следующим образом — если в результате применения параметров алгоритма 4) положительный ответ давался для сильно меньшего/большего числа классов, то я увеличивал/уменьшал вес класса 1, пока процент не становился похож на требуемый. Этот метод позволил в разы поднять F-меру для множества столбцов с плохой F-мерой, и не ухудшить ее для всех остальных, но, что странно, итоговая F- мера только ухудшилась. Этот факт позволяет предположить, что результаты алгоритма 4) и подобных плохи не только из-за переобучения, но и из-за того, что максимизация F-меры по столбцам отнюдь не максимизирует ее в целом.

Я пытался понять, почему это так, чтобы понять, что же стоит максимизировать вместо F-меры в алгоритме 4, или же придумать алгоритмы с оценкой наподобие той, что в методе 5, но положительных результатов достичь не удалось.

Результирующее решение

Увы, ни один пробуемый метод не дал хороших результатов.

В качестве итогового алгоритма был взят алгоритм 7) с результатом 0.522, что мало отличается от взятого за основу алгоритма 1).

Комментарии по поводу группового обсуждения

При выполнении задания я разбивал его на две основные части: первая — выбор метода решения из широкого множества классов, вторая — применение к выбранному методу всевозможных нетривиальных и специфичных для задачи «хитростей».

К сожалению, с групповым обсуждением у меня дела не заладились совершенно, т.к. по поводу первой части задачи я сразу решил, что в ней не участвую, а вторая часть — просто игра заключенного.

Вместе с тем, мне было бы очень интересно решать эту задачу в команде, имея возможность устно (что, очевидно, более продуктивно) обсуждать вышеупомянутые «хитрости» между собой.

Что касается пользы группового обсуждения для меня, то она заключалась в выборе класса алгоритмов, который имело смысл разрабатывать, соответственно, я опирался на результаты тех, кто проводил исследования этих классов, т.е.
— для линейных классификаторов — Петр Ромов и Андрей Остапец,
— для метрического алгоритма — Аня Потапенко и Евгений Нижибицкий,
— для латентного распределения Дирихле — Петр Ромов и Аня Потапенко.

Результаты и выводы

1. К сожалению, мне не удалось разработать алгоритм, адекватно превосходящий взятый за основу линейный классификатор по качеству работы (единственное толковое улучшение дает всего +0.02 к F-мере), поэтому мой результат следует считать неудачным. Тем не менее, работа, которую я проделал, мне кажется вполне полезной с точки зрения приобретения опыта решения задач машинного обучения и работы в MatLab-е (и даже немножко в других пакетах).

2. Что касается допущенных мной ошибок, то в первую очередь стоит отметить то, что я недооценил время, которое мне потребовалось для проработки всех своих модификаций алгоритмов. Каждый алгоритм работает довольно долго, а эффективно думать над следующей модификацией получается, только зная результаты предыдущей. У меня осталось множество непроверенных вариантов модификаций, что плохо.

3. В целом, само решение задания мне показалось очень интересным, было бы здорово решать что-то подобное еще раз, в идеале — в команде.

4. Я не отправлял жюри конкурса отчет, т.к. не нашел в своем результирующем алгоритме ничего такого, чем мог бы гордиться, а спецификации нашего университетского задания прослушал, так что у меня нет значения F-меры для результатов работы алгоритма. Впрочем, организаторы обещают выложить ответы для контрольной выборки в начале мая, так что, если это потребуется, я смогу узнать это значение.