МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В. ЛОМОНОСОВА

ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

КАФЕДРА МАТЕМАТИЧЕСКИХ МЕТОДОВ ПРОГНОЗИРОВАНИЯ

Задание №5(Решение реальной задачи «Topical Classification of Biomedical Research Papers»)

Отчет о проделанной работе

Евгений Зак

317 группа

Москва 2012

Объявив конкурс,
можно за малую часть зарплаты одного
получить увлечённую работу сотен и тысяч.

Елена Ермолова (писатель)

1.Постановка задачи.

Автору отчета было предложено поучаствовать в реальном (всемирном!) соревновании по классификации данных. Суть задания, на первый взгляд, проста: есть 10000 объектов, каждый из которых задан 25 640 признаками плюс 83 целевых бинарных признаков. Требуется для других 10 000 объектах с тем же количеством признаков предсказать целевые признаки.

2.Алгоритм решения.

Как уже было сказано выше, задание кажется простым. Но для того и требуется придумать алгоритмы решения, чтобы решить задачу. И вот тут автор понял всю сложность задачи.

Во-первых, матрица объект- признак очень сильно разряжена. Тут же вторая проблема – целевых признаков 83. Первая попытка классифицировать методом kNN (первое, что приходит в голову автору для классификации сразу 83 признаками (т.е. принимая, что они зависимы)) кончается неудачей. И понятно почему – 10 000 объектов мало, для того чтобы пытаться классифицировать по 83 классам вместе.

Эта неудача вызвала у автора некоторое нежелание пользоваться kNN(а зря, думает сейчас он). Как раз в это время на практикуме автором был изучен пакет WEKA, который показался автору интересным.

Для того, чтобы воспользоваться WEKA, понадобилось:

1) 64 битная система
2) 8 гигабайт оперативной памяти
3) Много терпения и времени
4) Начальные знания Java

Первые 2 пункта автор легко осилил, воспользовавшись стационарным компьютером из профкома факультета ВМК (за это отдельное спасибо). Третий пункт тоже дался автору легко. А вот пункт номер 4 преодолевался с трудом и долго. Знание java понадобилось для того, чтобы решить проблему «out of memory», так как сам пакет написан на java. Но благодаря интернету проблема out of memory осталась позади.

Для продолжения работы с пакетом автор создал 83 файла .arff , по одному для каждого целевого признака. С помощью элемента пакета «Experiments», протестировав различные алгоритмы на различных из 83 файлах с помощью кросс-валидации. Самые лучшие результаты (в среднем 0,38) показал алгоритм линейной классификации SMO (Platt's Sequential Minimal Optimization).

С одной стороны, работа в WEKA облегчает работу. На самом деле для решения реальных задач (таких как данная), так как после обучения получить в каком либо приемлемом виде ответы на тестовую выборку не получится. Для того, чтобы получить ответы пришлось пойти на хитрость (извращение, подсказывают автору его соседи по комнате) – т.к. признаки бинарные, то задать для тестовой выборке целевой признак всеми «0». А после этого отправить на классификацию и сохранить отчет, в котором для каждого объекта есть пометка – если она «+», то наш классификатор ошибся (т.е. предсказывает там «1») например, так:

739	1:0	2:1	+ 0	*1
740	1:0	1:0	*1	0
743	1:0	1:0	*1	0
744	1:0	1:0	*1	0

Эта запись значит, что 739 объекту поставлена метка «1». Написав небольшую программу для поиска «+» в 83 файлах-отчетах, автор получил первый вариант ответа. Проблема в том, что от начала работы в WEKA и получения файла ответа ушло в сумме около 20 часов, которые требовали присутствия у компьютера. В итоге первый ответ системы:

2012-03-24 15:55:39	result.txt	0.387

Еще какое-то время поработав с WEKA, автор понял что даже с помощью нормировок, селекции признаков и других методов предобработки 0,4 – это

потолок чего можно добиться. А вот в таблице лидеров уже были результаты 0,5 и выше. Поэтому автор решил начать все с начала, но уже как нормальные люди – в MatLab.

К тому моменту в рамках обсуждения решения задачи одногруппниками, было накоплено много опыта и за одну ночь автор смог разобраться в пакете Liblinear. Идея алгоритма проста : обучить 83 независимых классификатора и ими предсказать метки для тестовой выборки. Основная сложность заключается в подборе параметров. С помощью нормализации признаков(делением на константу) местами ручного, местами автоматического подбора параметров весов, коэффициента регуляризации и типа линейного классификации (-s 7 -w0 0.26 -w1 0.84 -c 0.715 -e 0.001) получить второй результат: 0,518.

К сожалению, и это оказалось потолком того, чего можно добиться подбором параметров. Различные нормализации, сжатия признаков (например, методом главных компонент и удаления нулевых признаков) улучшения результата не дали.

Направление для дальнейшего развития автору подсказал А.Г. Дьяконов (преподаватель и вдохновитель автора). Идея проста – с помощью имеющегося линейного классификатора получить не метки классов, а предсказания. Так же реализовать еще один(другой, не линейный) классификатор, так же получить предсказания. Сложив результаты обоих алгоритмов и подобрав порог получить итоговые ответы. В качестве второго классификатора был выбран kNN, но уже независимый (в отличие от выше описанного). В его настройке автору очень помогла А. Потапенко (одногруппница автора), которая к тому моменту имела большой опыт в реализации данного алгоритма для данной задачи. Подбор порога был произведен вручную и в итоге был получен результат:

BEST	**2012-03-30 02:20:32**	**resultsm.txt**	**0.523**

И это оказался последний потолок, которого смог добиться автор. Скорее всего, этот результат в конечном счете переобучен, так как финальный ответ системы ниже его:

11	+ Evgeny_Zak	Mar 30, 02:20:32	0.523	0.52110

3.Литература.

[1] - Дьяконов А.Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab – Москва, 2010.

[2] http://www.machinelearning.ru/

[3] http://www.cs.waikato.ac.nz/ml/weka/

4.Заключение.

В заключении хотелось бы привести некоторую статистику:

В ходе работы над решением задачи было:

1) Сделано 193 отправки решения на систему
2) Первая отправка была 19.02.2012 в 11:15
3) Последняя отправка 30.03.2012 в 20:12
4) Задействовано 1 ноутбук и один стационарный компьютер
5) Проведено 6 бессонных ночей
6) Выпита 1 банка кофе
7) Проспано 2 лекции А.Г. Дьяконова (☹)
8) Проведено более 2 часов разговоров в Skype с верным другом и товарищем автора А. Потапенко

И можно смело сказать, что все это не напрасно. В ходе работы автор сделал ряд выводов, а именно:

1) Пакет Weka очень удобен в использования, но для задач с большим количеством признаков, к сожалению, не целесообразен. Он скорее подходит для первоначального осмотра задачи (например, благодаря работы с пакетом автор быстрее подобрал параметры для Liblinear, так как в Weka это делается быстрее и проще).

2) Все таки Matlab – самое эффективное средство для решения задач классификации. Хотя бы с точки зрения скорости работы и исследования. В самом начале автор предпринял попытку решать задачу в среде C#, которая так и не окончилась каким-либо результатом.

3) И чего не сделаешь ради победы в соревновании? Бессонные ночи, нехватка времени на другие предметы – все это того стоит. Соревновательный дух для изучения чего-то нового – самый лучший стимул.

Более локальные выводы:

1) Различные операции с матрицей объект-признак не всегда могут улучшить результат: в данном случае ни нормировка, ни сжатие пространства не давали какого-либо положительного результата

2) Разные алгоритмы классифицируют по-разному. Имеется в виду, что ответы они дают разные, так как классифицируют по-разному. Поэтому смесь разных алгоритмов дает улучшение.

Выводы о том, как проходило решение автором задачи. Предложенная модель первоначального обсуждения с одной стороны хороша. Во-первых, видно кто сколько работает и каких результатов добивается. И это заставляет начинать работать. Это стимул попробовать что-то новое. Но тут кроется и минус данного подхода – желание изобрести что-то новое не всегда приводит к чему то хорошему. Даже чаще оно во вред. Стандартные методы оказались гораздо эффективнее новинок. Этот факт немного расстраивает – получается, что ход решения этой задачи не мотивирует изобретать что-то новое. А значит сводиться лишь к изучению того, что уже известно человечеству. Не романтично как-то, считает автор. Но для образовательного процесса очень полезно – за такой короткий срок было получено очень много опыта, причем именно прикладного, не теоритического. Результаты такого обучения можно «пощупать» на реальном примере. Автор нащупал то, что он ,видимо, обучился не идеально, потому что есть как минимум 10 команд, которые смогли «дощупаться» дальше чем он.

Автор хочет поблагодарить следующих выдающихся людей:

1) **Дьяконова Александра Геннадьевича** – за интересную задачу по практикуму, которая навсегда отложиться в памяти ярким пятном;

2) **Местецкого Леонида Моисеевича** – за проявленное понимание к задержкам с научной работой по спецсеминару;

3) **Потапенко Анну** – за помощь, как идейную, так и моральную;

4) **Ромова Петра** – за то, что своими действиями дал стимул к началу работы, при этом предоставив все инструменты для начала (данные в нормальном виде, ссылки на различные пакеты);

5) **Нижибицкого Евгения, Кондрашкина Дмитрия, Остапца Андрея, Любимцеву Марию, Новикова Максима, Шаймарданова Ильдара–** за то, что нашли в себе силы «раскопать» задачу глубже, и предоставить свои знания другим;

6) **Курова Ивана –** за помощь с поиском мощного компьютера, и его предоставлением для исследования.

А закончить свой отчет автор хочет цитатой писателя Сергея Лукьяненко, которая ассоциируется у автора с тем промежутком времени, когда проходило обсуждение задачи:

> *Не надо никаких гениев, которые хотят сделать весь мир счастливым насильно. Надо лишь помогать тем, кто рядом. Тогда лучше станет всем.*
>
> *Сергей Лукьяненко. Танцы на снегу*