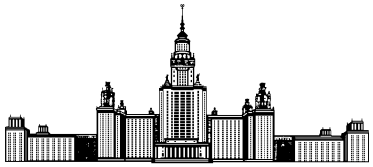
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Какое-то классное название»

Выполнил:

студент 3 курса 317 группы

Фамилия Имя Отчество

Научный руководитель:

д.ф-м.н., профессор

Фамилия Имя Отчество

Заведующий кафедрой

Математических Методов

Прогнозирования, академик РАН _____ Ю. И. Журавлёв

К защите допускаю

«_____» _____ 2010 г.

К защите рекомендую

«_____» _____ 2010 г.

Москва, 2011

Содержание

Аннотация

Аннотация собирается в последнюю очередь путем легкой модификации наиболее важных и удачных фраз из введения и заключения.

Из этого следует, что в работе должны быть удачные фразы.

1 Введение

Введение тоже пишется как-то в конце.

1.1 Определения и обозначения

Формальная постановка задачи. Которые могут очень разнится... Например, одной из интересных формулировок является следующая (источник - [1]): для обучения дана выборка из некоторого распределения. Затем подаются новые точки из него же и какие-то левые, нужно отделять первые от вторых.

Другим определением аномалии может быть "Есть некоторая нормальная модель, описывающая исходная данные. Аномалия - это то, что плохо вписывается в эту модель". Встречаются случаи, когда все аномалии похожи друг на друга и скапливаются в одной точке. В этом случае предположение, например, о том, что у аномалий относительно далеко находятся ближайшие соседи может не зайти. Также: исходя из такого определения, одним из теоретически оптимальных алгоритмов является применить какой-нибудь моделирующий алгоритм машинного обучения на всех данных в целом и выделить в качестве аномалий объекты с наибольшим отступом.

Также помимо аномалий существует такое понятие, как "шум". Отделять одно от другого задача довольно таки гробовая, однако природа этих вещей разная => и подход к ним тоже может оказаться разный.

По мотивам р. 1.5. Допустим, есть данные специфического характера (временной ряд, последовательность символов, т.д.). Согласно общим рассуждениям, решать задачу можно так: 1) построить признаковое пространство по этим данным 2) решить задачу для этого пространства. Естественно, что могут найтись шорткаты - например, в последовательностях символов вдруг обнаружился совершенно новый неизвестный науке иероглиф. Однако глубокого методического смысла в подробном изучение шорткатов я не вижу - это всегда справедливо, что наличие дополнительной информации для общей задачи влечёт некоторую вероятность существования упрощения общего алгоритма. В силу этого, интересно рассматривать только общие задачи (с обучением на неразмеченных данных; с

обучением на нормалиях; возможно, с обучением на аномалиях, хотя у меня не получилось это представить, ведь раз есть данные только об аномалиях, значит их, очевидно, физически больше чем нормалий, и надо просто инвертировать задачу; с обучением на размеченных данных).

1.2 Определение меток по Anomaly Score

Предположим, в задаче с общей постановкой некоторым образом был получен anomaly score (эту штуку надо тоже как-то ёмко называть... "степень аномальности"слишком длинно...), однако разделение на аномалии и нормалии неизвестно. Тогда, можно сказать, просто задача сведена к одномерному случаю. На прямой даны точки, нужно выделить аномалии. Судя по опыту Isolation Forest, самым простым случаем является тот, в котором степень аномальности нормалий сжимается в кластеры. То есть если человеческий глаз посмотрит на такую картинку, он сразу выделит "пустые пространства"вокруг этих кластеров и назовёт очень похожий на правду ответ.

Ноутбук с экспериментами по одномерному случаю

Конечно, наверняка есть более классические способы решать задачу. Например, можно как-то узнать, что вот нормалии распределены по нормальному закону, вот примерный центр, дисперсия, правила трёх сигм... Но вот по какому порогу правдоподобия отделять - неясно, три сигмы это просто число хорошее, как десять или шестьдесят шесть. А ведь интуитивно кажется, что это как раз и есть самое интересное! И, конечно, так получилось, что в ноутбуке мне было интересно поработать в режиме минимальной доп. информации вроде количества кластеров или, тем более, предположений о распределении данных.

Update: применение этого алгоритма к датасету саттелит выдало интересный результат...

1.3 Выбор критерия качества

Отдельным вопросом стоит метрика оценивания качества. Например, если это F-мера, то она должна быть взвешанной (cost за пропуск аномалии выше, чем за ложную тревогу)

TODO придумать способ адекватного выбора метрики

1.4 Обзор литературы

«Перечисляются подходы, методы, факты, на которые существенно опирается данная работа» - это в обзоре-то литературы?

2 Новые подходы и результаты

Здесь надо написать много чего-то умного

2.1 Методы решений

The best choice of model is often data-specific (c) Алгоритмы детектирования аномалий обычно имеют свои аналоги среди алгоритмов обучения с учителем. Следующие алгоритмы вызывают любопытство:

- EM-алгоритм. См. ниже подробнее "вероятностный подход"
- "Метод Махаланобиса". По сути, данные нормализуются (центрируются-нормируются), после чего... после чего снова не очень понятно. То есть да, если объём данных велик, можно свести задачу к случаю "мат.ожидание ноль, дисперсия 1 однако для дальнейших шагов всё равно нужны либо метрические эвристики, либо вероятностные предположения.

  (?) в книге заявлено, что его аналогом среди supervised-алгоритмов является некий Rocchio. Впервые слышу такое, что это?

  Также у меня есть подозрение, что это почти EM для случая с одним кластером...

- Isolation Forest, подробнее ниже
- Линейные методы и PCA

  Линейные методы описывались, судя по всему, для того, чтобы громогласно объявить о том, что они не работают, однако заявлено, что может слегка помочь сэмплирование. Чуть интереснее дела обстоят с PCA

  Как и в оригинале, строится оптимальное подпространство заданной размерности $d < D$. В случае нашей задачи рекомендуется брать d близким к D, так как аномальность объектов может проявляться в том числе "вдоль"тех собственных векторов $X^T X$, которые соответствуют малым собственным значениям. Иными словами, оригинальный метод, заключающийся в том, что отбрасывает не самую нужную информацию с целью её сжатия, в контексте нашей задачи может как раз упустить то, что нам интересно. Поэтому опять же сэмплирование. Также для сэмплирования есть незначительные модификации (основное достижение которых состоит в том, что на каждом сэмпле (размер которого предполагается достаточно малым) outlier score считается всё равно для всей выборки).

  С PCA есть такой интересный момент. Для него требуется матрица "похожести"(similarity) XX^T (просто матрица попарных расстояний, по-русски говоря). Например, применение kernel trick для PCA равносильно замене этой матрицы на нечто соответствующее матрице попарных расстояний в новом спрямляющем пространстве, дальнейшая работа алгоритма связана с работой с этой матрицей, поэтому матрицу придётся задавать явно, ну то есть алгоритм эту матрицу будет получать на некотором этапе работы. Если предполагать, что сэмплирования не происходит, это означает, что в алгоритме заложено вычисление матрицы попарных расстояний. Так если эта матрица подсчитана, то почему бы не применить метрические алгоритмы (раз уж всё предподсчитано!)? Ведь понятно же, что у PCA "задача"другая, так что он такого даёт, что потенциально лучше kNNоподобных алгоритмов? Сам по себе kernel trick - способ генерации признаков, на нём, может быть, можно получить хороший результат на и без того работающих алго-

ритмах... но допустим, пространство признаков такого, что в нём аномалии отличимы, тогда, конечно, для улучшения работы ядерный переход делать можно, но по идее, всё ж должно работать и без него.

(?) Есть следующая мысль неизвестной степени содержательности. Взять случайный объект в выборки. И применить PCA со следующими двумя изменениями: 1) объекты имеют веса в зависимости от близости к выбранному объекту 2) пространство должно проходить через выбранный объект (как реализовать без мучений - ?). Получится условно такая "касательная"к подпространству, в котором лежит выборка. Применить так много раз (можно дополнительно проверять, не оказался ли выбранный объект-фиксатор аномалией по расстояниям до соседей, чтоб касательные были правильней) и усреднить расстояния до подпространств главных компонент.

- One-class SVM (вызывает любопытство не у меня)

  Первый же вопрос - а не проще ли строить минимальную выпуклую оболочку с тем же успехом?

- Replication Neural Networks - кажется, это нейронные сети, бежим отсюда в страхе

- FP-Outlier - что это? в книге это упрятано в последние главы.

  (?) хочется выяснить, что это

- Метрические аналогии, в первую очередь, kNN-а. Полагаю, что их не одна штука, и здесь доступно много вариаций. Одно из главных преимуществ этого подхода - интерпретируемость и логичность результата. Так что с этим можно покопаться.

  TODO выяснить подробнее об алгоритмах LOF, LOCI, которые заявлены как раз метрическими алгоритмами

  TODO Немного про морфологический спектр. Давайте найдём для точки расстояние до первого, второго, третьего... k-го соседа, k - большое. Это много, конечно, чисел. Зато ответ спрятан же где-то в них, так? И эти числа,

заметим, не очень зависят от датасета. Может, взять матожидание или ещё что от этого набора? Или и тут ножницы впихнуть?

Помимо алгоритмов выявления аномалий, выделяются алгоритмы нахождения "экстремальных значений как это по-русски-то сказать, Extreme-value Analysis - в общем, поиск "граничных"объектов в выборке. Насколько они полезны, пока сомневаюсь, но пока сделаю краткий конспект (р.2.3):

- Depth-based) строим минимальную выпуклую оболочку. Выкидываем то, что оказалось элементами границы. Повторяем. Таким образом, каждому объекту будет соответствовать натуральное число, "слой"границы, чем меньше - тем "граничнее"объект.

- Deviation-based) идея заключается в нахождении такого подмножества объектов, при выкидывании которых дисперсия выборки уменьшается. Однако странная затея, как по мне, ведь если проводить какую-нибудь централизацию выборки, то для уменьшения дисперсии достаточно просто "убирать"граничные элементы - иными словами, когда остановится, всё равно непонятно, с этой логики выгодно убирать граничные элементы, но как это помогает их найти - непонятно.

- Angle-based) у граничных элементов "угол в пределах которого лежат все остальные объекты, существенно меньше пи (ну это для двумерного случая, что там в многомерном не очень понятно). Есть эвристики через сэмплирование, позволяющие оценить этот угол, благо точность тут не нужна: чем "шире"угол, тем менее граничным объявляется объект... ну окей

2.2 Вероятностный подход

Основное моё смущение - предложить параметризованную модель распределения это уже решить половину задачи. Однако, в обучении с учителем гипотезу можно проверить и качественно оценить, скажем, на контроле. В случае же обучения без учителя, вероятностное моделирование может применяться только с опорой на априорную информацию, потому что иначе выдаваемый алгоритмом результат безоснователен.

Смущает также то, что алгоритм подбора параметров в том же EM-алгоритме (а предполагается использовать, как я понимаю, оригинальный алгоритм без каких-либо модификаций) неустойчив к "выбросам"... То есть если в выборке 40% значений, алгоритм будет настраиваться на шумы и сходить из-за этого с ума. Причём мы никак не можем его спасти, потому что тогда бы решили исходную задачу))))). Вторым смущением является то, что необходима априорная информация о количестве кластеров в датасете.

Допустим, модель выбрана, а параметры некоторым образом оценены. Тогда выдать результат можно опираясь на квантили распределения (в некотором смысле, условная вероятность $a(x) = p(y|x)$ и будет anomaly score).

Однако, можно применить и другой подход, оценивая моменты, мат.ожидание и дисперсию, после чего применяя всякий теорвер вроде неравенств Маркова, Чебышева и ещё много всякого.

Примечание: если данных мало (скажем, 20 объектов), то на помощь могут прийти распределения Стьюдента и прочие статистические способы проверки гипотез... я притворюсь, что это не мой случай.

2.3 Isolation Forest

Основная статья: Ссылка. Хм, а в итоговом тексте же не будет ссылок! То есть когда я сошлюсь на эту статью [1], читатели пусть сами долго и страдающе гуглят.

Краткий перессказ статьи - Isolation Forest круче всех. Работая в предположении "аномальные точки - изолированные, то есть с низкой плотностью априорного распределения можно за линейное время (с константой, не пропорциональной размерности пространства) научиться выдавать объектам адекватный скор аномальности, превосходящей по метрике AUC алгоритмы-предшественники

(?) якобы, особенно в многомерном случае - почему так, мне непонятно.

(?) Сам по себе алгоритм вызывает следующие вопросы: зависимость качества работы алгоритма от размера случайных подвыборок для каждого дерева. По заявлениям авторов статьи, этот параметр существенен, но не требует тонкой настройки ("обычно 250 это норм вариант")

Кстати, некоторое теоретическое обоснование, почему работает тот же Isolation Forest: вся теория о bias-variance разложении верна и для обучения без учителя, все стандартные утверждения об ансамблировании справедливы и для задачи обнаружения аномалий.

(?) а если все признаки бинарны? Вообще, это следующий особый случай после одномерного, вот что делать, если у каждого признака только два значения? Интересно, можно ли модифицировать тут Isolation Forest так, чтобы он работал и в таком случае?

Кстати, хотел попробовать такой вариант. Режем не в случайном месте, а там, где расстояние наибольшее. Isolation Forest режет там с наибольшей вероятностью, что добавляет стохастичности. Однако, если наибольших диапазона, например, два, то логичнее выбирать именно случайным образом. Тем более, если набор таких расстояний без выраженного максимума... нет, оригинал лучше этой идеи, однако вдруг такой анализ тоже окажется интересным - в конце концов, на каждом узле выбирается случайный признак, и, возможно, в каких-то случаях так делать лучше, например, когда признаков много.

2.4 Isolation Forest и вращения!

После знакомства появилось страстное желание сделать для каждого дерева случайный выбор базиса (это называется умными словами rotated bagging). Ссылка на эксперименты. Во всех случаях контуры одинакового скора аномальности становились эстетически красивее. По мере auc-roc эти искусственные задачи Isolation Forest решает с высоким результатом и за линейное время. Встал выбор адекватной метрики оценивания качества работы детектора аномалий - а это нетривиальный вопрос.

Искусственной реализацией rotated bagging может служить следующий костыль: поскольку афинное преобразование - это домножение на невырожденную матрицу плюс сдвиг, то невырожденную матрицу можно генерить как случайную (гипотеза: вырожденная матрица будет получаться с пренебрежительно низкой вероятностью), а сдвиг неважен, так как к нему изолирующее дерево инвариантно. Более красивое решение подсмотрено вот здесь, берётся случайная матрица

и от неё QR-разложение. Кажется, это проходилось на первом курсе, но, как и половина всего линала, где-то мимо.

(?) На датасетах с немасштабированными признаками работает плохо! Действительно, при таком вращении признак с большим разбросом будет всегда играть большую роль. Нормировка не работает... возможно, нужно нормировать дисперсию или как-то хитрее вращать.

TODO протестировать и сравнить IsolationForest и IsolationForestWithRotatedBagging (название сократить) на известных датасетах

Отдельный интерес представляет то, как алгоритм выдаёт оценку плотности распределения.

TODO сравнить результат с чистыми density-based алгоритмами.

3 Вычислительные эксперименты

...

3.1 Исходные данные и условия эксперимента

Описывается прикладная задача, сколько объектов, сколько признаков, каких они типов, параметры эксперимента, как там контроль, например

3.2 Результаты эксперимента

Результаты экспериментов представляются в виде таблиц и графиков. Объясняется точный смысл всех обозначений на графиках, строк и столбцов в таблицах.

3.3 Обсуждение и выводы

Приводятся выводы: в какой степени результаты экспериментов согласуются с теорией? Достигнут ли желаемый результат? Обнаружены ли какие-либо факты, не нашедшие объяснения, и которые нельзя списать на «грязный» эксперимент?

Обсуждаются основные отличия предложенных методов от известных ранее. В чем их преимущества? Каковы границы их применимости? Какие проблемы

удалось решить, а какие остались открытыми? Какие возникли новые постановки задач?

4 Заключение

Пафосные заявления о том, как я молодец, шаблонами вида

- Предложен новый подход к…
- Разработан новый метод…, позволяющий…
- Доказан ряд теорем, подтверждающих (опровергающих), что…
- Проведены вычислительные эксперименты…, которые подтвердили / опровергли / привели к новым постановкам задач.

5 Хочу прочитать

- Про алгоритмы, основанные на расстояниях между объектами
- Люто много математики по задаче скора аномальности (какая-то страшная статья)
- Пресловутый One-class SVM и винда

Список литературы

[1] Isolation Forest — Fei Tony Liu, Kai Ming Ting, ????.