Государственное образовательное учреждение высшего профессионального образования

«Московский государственный технический университет имени Н. Э. Баумана»

(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

Р А С Ч Ё Т Н О - П О Я С Н И Т Е Л Ь Н А Я З А П И С К А

к курсовой работе на тему:

Метод сохранения качества изображений при понижении его разрядности

Студент Капустин А.И. И.О. Фамилия
(Подпись, дата)

Руководитель курсового проекта Оленев А.А. И.О. Фамилия
(Подпись, дата)

Москва 2016

Содержание

Введение 4
1 Аналитический раздел 5
1.1 Цель и задачи работы 5
1.2 Что такое аномалия 5
1.3 Обнаружение аномалий 6
1.3.1 Классификация методов обнаружений аномалий 6
1.3.2 Признакое представление данных 7
1.4 Результат метода обнаружения аномалий 7
1.5 Виды аномалий 8
1.5.1 Нормализация данных 10
1.5.1.1 Основные методы нормализация данных 10
1.6 Неконтролируемые алгоритмы обнаружения аномалий 10
1.6.1 Вероятностный-генеративный подход 10
1.6.2 Линейный подход 11
1.6.3 Метрические методы 11
1.6.3.1 Базовые понятия 12
1.6.3.2 Оптимизация Рамасвани 13
1.6.3.3 Методы Кнора-Реймонда и Байерса-Рейтери 13
1.6.4 Метод Танга 14
1.6.5 Параметрическе методы 14
1.6.6 Локальный коэффициент выбросов(LOF) 15
1.6.6.1 Компонентный коэффициент выбросов(COF) 16
1.7 Методы улучшения алгоритмов 17
1.7.1 Семпилрования 17
1.7.2 Ансамблирование 18
1.7.3 Итеративный отбор 18
1.8 Выводы 19
2 Конструкторский раздел 20
3 Технологический раздел 21
4 Исследовательский раздел 22
4.1 Время дизеринга раличных алгоритмов 22
4.2 Качество получаемого изображения 23
4.3 Размер получаемого изображения 23
Заключение 25
Список использованных источников 26

Глоссарий

Выборка/выборка данных — конечный набор прецедентов (объектов, случаев, событий, испытуемых, образцов, и т.п.), некоторым способом выбранных из множества всех возможных прецедентов, называемого генеральной совокупностью[1].

— Метка(ярлык) - порция данных, идентифицирующая набор данных, описывающая его определенные свойства и обычно хранимая в том же пространстве памяти, что и набор данных[2].

классификатор?

Теория распознавания образа — раздел информатики и смежных дисциплин, развивающий основы и методы классификации и идентификации предметов, явлений, процессов, сигналов, ситуаций и т.п. объектов, которые характеризуются конечным набором некоторых свойств и признаков. ddos-атака? Датасет - набор данных[3]

Введение

Задача поиска аномалий относится к одному из популярных способов машинного обучения - обучению без учителя. В настоящее время задачу поиска аномалий активно решают во многих областях жизнедеятельности:

а) Защита информации и безопастность

б) Социальная сфера и медицина

в) Банковская и финансовая отрасль

г) Распознавание и обработка текста, изображений, речи

д) Другие сферы деятельности(например, мониторинг неисправностей механизмов)

Задачей поиска выбросов, как частный случай задачи поиска аномалий так же занимаются во всех вышеперичисленных отраслях.

Количество данных в мире удваивается примерно каждые два года. Поэтому актуальной задачей является разработка новых методов и усовершенствования старых методов поиска выбросов.

В данной работе предлагается новый метод, позволяющий найти аномалии в выборках данных.

1 Аналитический раздел

1.1 Цель и задачи работы

Целью данной работы является создание программного комплекста для обнаружения выбросов временных рядов в собираемых данных. Для достижения данной цели необходимо решить следующие задачи:

— ПЕРЕПИСАТЬ ИЗ ПРЕЗЕНТАЦИИ

— пронализировать предметную область и существующие методы обнаружения выбросов

— разработать метод обнаружения выбросов

— создать ПО, собирующее данные для анализа

— создать ПО, реализующего разработанный метод обнаружения выбросов

— провести вычислительный эксперимент с использованием разработанного метода

1.2 Что такое аномалия

В анализе данных есть два основных направления, которорые занимаются поиском аномалий - это детектирование новизны и обнаружение выбросов. "Новый объект" - это так же объект, который отличается по своим свойствам от объектов выборки. Однако, в отличие от выброса, его ещё нет в самой выборке и задача анализа сводится к его обнаружению при появление. Например, если вы анализируете замеры уровня шума и отбрасываетете слишком высокие или слишком низкие значения, то вы боретесь с выбросам. А если вы создаёте алгоритм, который для каждого нового замера оценивает, насколько он похож на прошлые, и выбрасывает аномальные — вы "боретесь с новизной"[4]. Выбросы являются следствием:

а) ошибок в данных
б) неверно классифицированных объектов
в) присутсвием объектов других выборок
г) намеренным искажением данных

На рисунке 1.1 можно увидеть желтые точки - выбросы в "слабом смысле". Они незначительно отклоняются от основных данных(зеленые точки). Красные же точки являются аномальными - выбросами "в сильном смысле они значительно отклоняются от основных данных. В данной работе будет изучаться вопрос находждения "сильных выбросов"и критериев отличия сильного выброса от основных данных. В дальнейшем под словом "выброс"будет подразумеваться "сильный выброс а под аномалией - выброс(выброс является частным случаем аномалии). Понятие аномалии зачастую интерпетируют по-разному в зависимости от характера

данных. Обычно аномалией назыют некоторое отклонение от нормы.В дальнейшем будет дано несколько более формальных опредений аномалий, специфичных для метода их определений.

1.3 Обнаружение аномалий

В машинном обучении обнаружение "ненормальных"экземпляров в наборах данных всегда представляло большой интерес. Этот процесс широко известен как обнаружение аномалий или обнаружение выбросов. Вероятно, первое определение было дано Граббсом[5] в 1969 году: "Относительное наблюдение или выброс - это элемент выборки, который, заметно отличается от других членов выборки, в которых он встречается ". Хотя это определение по-прежнему актуально и сегодня, мотивация для обнаружения этих выбросов сейчас совсем другая. Тогда основная причина обнаружения заключалась в том, чтобы удалить выбросы из данных для обучения, поскольку алгоритмы распознавания были весьма чувствительны к выбросам в данных. Эта процедура также называется очищением данных. После разработки более надежных классификаторов интерес к обнаружению аномалий значительно снизился. Однако в 2000 году произошел поворотный момент, когда исследователи стали больше интересоваться самими аномалиями, поскольку они часто связаны с особенно интересными событиями или подозрительными данными. С тех пор было разработано много новых алгоритмов, которые оцениваются в этой статье. В этом контексте определение Граббса также было расширено, так что сегодня аномалии, как известно, имеют две важные характеристики:

а) Аномалия отличается от нормы по своим особенностям

б) Аномалия редко встречается в наборе данных по сравнению с "нормальными" данными

1.3.1 Классификация методов обнаружений аномалий

В отличие от хорошо известной системы классификации, где учебные данные используются для обучения классификатора, а результаты измерений данных оцениваются впоследствии, возможно множество вариантов, когда речь идет об обнаружении аномалий. Метод обнаружения аномалий, которая будет использоваться, зависит от ярлыков, доступных в наборе данных, и мы можем выделить три основных типа:

а) Обучение с учителем. Доступны полностью размеченные данные для обучения и для тестов. Обычный классификатор может быть обучен один раз и применяться впоследствии. Это похоже на традиционное распозвание образов, за исключением того, что классы обычно сильно не сбалансированы. Поэтому не все алгоритмы клас-

сификации идеально подходят для этой задачи. Для многих применений аномалии не известны заранее или могут возникать спонтанно в качестве новинок на этапе тестирования.

б) Обучение с частичным привлечением учителя.Обучение использует учебные и тестовые наборы данных. Данные обучения состоят только из нормальных данных без каких-либо аномалий. Основная идея заключается в том, что модель нормального класса изучается, а аномалии могут быть обнаружены впоследствии, отклоняясь от этой модели. Эта идея также известна как «одноклассовая» классификация [6].

в) Обучение без учителя. Самый гибкий способ, который не требует каких-либо меток. Кроме того, нет различия между учебным и тестовым наборами данных. Идея заключается в том, что алгоритм обнаружения аномалии оценивает данные исключительно на основе внутренних свойств набора данных. Как правило, расстояния или плотности используются для оценки того, что является нормальным, а что является выбросом. В этой работе основное внимание будет этому именно этому способу. Так же этот способ иногда называют "неконтролируемый способ обнаружения аномалий".

1.3.2 Признакое представление данных

В дальнейшем будем исходить из предположения что данные имеют признаковое представление, т.е. каждый объект x представлен в виде вектора $\mathbb{R}^d$. В задаче обучения без учителя задача обнаружения аномалий формулируется следующим образом: в заданном множестве X для каждого элемента выдать 0, если этот объект относится. При этом правильных ответов не предоставляется.

В аналогиченой задаче обучения с учителем на некоторой выборке входных данных X_train, называемой тренировочной выборкой, известен правильный ответ, т.е для каждого элемента $x \in X_train$ представлены метки $y \in 0,1$, характирующие является ли объект аномалией или нет. Для выборки данных, для которой метки не предоставлены, задача сводится к задаче бинарной классификации, а значит может решаться при помощи любых алгоритмов машиного обучения с учителем. Возможны и "вырожденные"случаи, когда все метки тренировочного набора данных одинаковы. В таком случае алгоритмы выдают неправильный результат.

1.4 Результат метода обнаружения аномалий

Существует два варианта выходных данных алгоритма обнаружения аномалии. Во-первых, метка может использоваться как результат, указывающий, является ли экземпляр аномалией или нет. Во-вторых, оценка или достоверность могут быть более информативным результатом, указывающим на степень аномалии. А алгоритмах метода обучения с учителем зачастую используются метки как выходные дан-

ные. С другой стороны, для алгоритмах с частичным привлечением учителя и без учителя обнаружения аномалий чаще встречаются оценки.

1.5 Виды аномалий

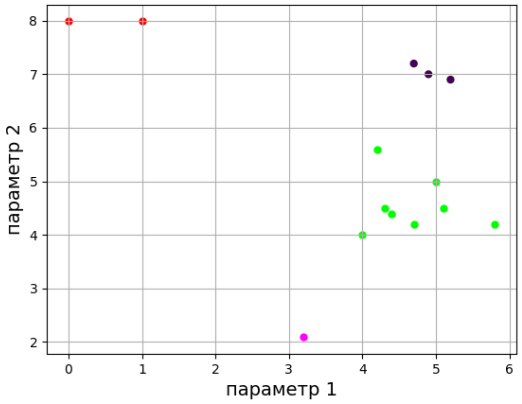

Рисунок 1.1 — Простой двумерный пример

Основная идея алгоритмов обнаружения аномалий заключается в обнаружении экземпляров данных в наборе данных, которые отклоняются от нормы. Однако на практике существует множество случаев, когда это основное предположение является неоднозначным. На рис 1.1 показаны некоторые из этих случаев с использованием простого двумерного набора данных. Две аномалии могут быть легко идентифицированы визуально: красные точки сильно отличаются отличаются значениям параметров от областей плотной группировки точек и поэтому называются глобальными аномалиями.Когда мы смотрим на весь набор данных в целом, то фиолетовую точку можно отнести к тому же классу, что и зеленые тчки. Однако, когда мы фокусируемся только на кластере зеленых точек и сравниваем его с фиолетовой, пренебрегая всеми другими точками, то её можно рассматривать как аномалию. Поэтому фиолетовая точка называется локальной аномалией, так как она аномальна по сравнению с ее близкой окрестностью. В зависимости от цели исследования,нас могут интересовать местные аномалии или нет. Другой интересный вопрос заключается в том, следует ли рассматривать точки черного кластера как три аномалии или как (небольшой) кластер. Эти явления называются микрокластерами, а алгоритмы обнаружения аномалий должны присваиваеть оценку(вероятность того, что точка является аномалией)ему точкам этого кластера значения большие, чем точкам зеленого кластера, но меньше, чем красным точкам. Этот простой пример уже показывает, что аномалии не всегда очевидны, а оценка намного полезнее, чем назначение двоичных меток.

Обычно под аномалией принимают точки красные точки, так как их характеристики значельно отличаются от характеристик датасета, а так же их небольшое

количество.Однако, такой принцип обнаружения иногда терпит неудачу. Например, при хакерских ddos-атаках, большая часть трафика - необычная, аномальная. В этом случае алгоритм обучения без учителя потерпит неудачу и не сможет выделить хакерскую атака как аномальное поведение.

Задача обнаружения одиночных аномальных экземпляров в более крупном наборе данных (как это представлено до сих пор) называется обнаружением точечной аномалии[7]. Сегодня почти все доступные неконтролируемые алгоритмы обнаружения относятся к этому типу. Если аномальная ситуация представлена как множество многих случаев, это называется коллективной аномалией. Каждый из этих экземпляров не обязательно является точечной аномалией, но только определенная их комбинация определяет аномалию. Предыдущим приведенным примером возникновения нескольких специфических шаблонов доступа при обнаружении ddos-атак является такая коллективная аномалия. Третий вид - это контекстуальные аномалии, которые описывают эффект, что точка может рассматриваться как нормальная, но когда данный контекст учитывается,то точка оказывается аномалией. Самым распространенным контекстом является время. В качестве примера предположим, что мы измеряем температуру в диапазоне от $-30°C$ до $+40°C$ в течение года. Таким образом, температура $25°C$ кажется довольно нормальной, но когда мы учитываем контекстное время (например, месяц), такая высокая температура $25°C$ в течение зимы будет рассматриваться как аномалия.

Алгоритмы обнаружения точечных аномалий так же можно использовать для обнаружения контекстуальных и коллективных аномалий. Для этого нужно включить сам контекст как параметр. В вышеприведенным примере включение месяца как дополнительного параметра поможет обнаружить аномалию. Однако в более сложных сценариях может потребоваться один или несколько новых парметров, чтобы преобразовать задачу определения контекстной аномалии в проблему обнаружения точечной аномалии. Преобразование поиска коллективной аномалии в поиск одиночную может быть нетривиальной. Корреляция, агрегация и группировка используются для создания нового набора данных с другим представлением признаков[8] . Преобразование из задачи обнаружени коллективной аномалии в задачу обнаружения точечной аномалии требует глубоких знаний о наборе исходных данных и часто приводит к существенным искажениям при переводе данных в новый формат. Такое семантическое преобразование называется генерированием представления данных(*англ. data view generation*).

Таким образом можно сделать вывод, что многие задачи обнаружения аномалий требуют предварительной обработки данных перед передачей их на вход алгоритму. В противном случае можно получить формально верные, но фактические бесполезные результаты.

1.5.1 Нормализация данных

Когда мы получили предварительно обработанный датасет для поиска точечной аномалии, то последним шагом перед передачей в алгоритм, является нормализация данных. Нормализация данных предназначена для усранения зависимости от выбора единицы измерения и заключается в преобразовании диапазонов значений всех атрибутов к стандартным интервалам([0,1] или [-1,1])[9]. Нормализация данных направлена на придание всем атрибутам одинакового "веса".

1.5.1.1 Основные методы нормализация данных

а) Min-max нормализация заключается в применении к диапазону значений атрибута x линейного преобразования, которое отображает [min(x),max(x)] в [A,B].

$$x'_i = \tau(x_i) = \frac{x_i - min(x)}{max(x) - min(x)} * (B - A) + A \tag{1.1}$$

$$x \in [min(x), max(x)] \Rightarrow \tau() \Rightarrow [A,B] \tag{1.2}$$

Min-max нормализация сохраняет все зависимости и порядок оригинальных значений атрибута. Недостатком этгого метода является то, что выбросы могут сжать основную массу значений к очень маленькому интервалу

б) Z-нормализация основывается на приведении распределения исходного атрибута x к центрированному распределению со стандартным отклоненим, равным 1 [9] .

$$x'_i = \tau(x_i) = \frac{x_i - \overline{x}}{\sigma_x} \tag{1.3}$$

$$M[x'] = 1 \tag{1.4}$$

$$D[\overline{x}'] = 0 \tag{1.5}$$

Метод полезен когда в данных содержатся выбросы.

в) Масштабирование заключается в изменении длины вектора значений атрибута путем умножения на константу [9] .

$$x'_i = \tau(x_i) = \lambda * x_i \tag{1.6}$$

Длина вектора x уменьшается при $|\lambda| < 1$ и увеличивается при $|\lambda| > 1$

1.6 Неконтролируемые алгоритмы обнаружения аномалий

1.6.1 Вероятностный-генеративный подход

Основная идея генеративного подхода заключается в использование вероятносного смесевого моделирования данных. Предлагается подобрать такую вероятностую модель, из которой было получены нормированные данные. Такие модели обычно называются генеративными моделями, где для каждой точки(элемента

данных) можем посчитать генеративную вероятность(или вероятность правдоподобия).Т.е. задача сводится к нахождению плотности распределения p(x). Аномиями при этом считаются точки(элементы набора данных), имеющию низкое правдоподобие. В качестве показателя аномальности выступает функция p. Для построения генеративной модели нужно решить следующую задачу:

$$\prod_{x \in X_{norm}} p(x,\theta) \rightarrow max_{\theta} \tag{1.7}$$

где X_{norm} - нормальные данные представленного набора данных $p(x,\theta)|\theta \in \omega$ -семейство плотностей вероятностей, параметризованные θ;

Этот метод редко используется на практике, так как тяжело проверить полученную генеративную модель на адекватность, сложно убедится в правильном выборе семейства смесевых распределений. Это связано с тем, что низкое значение функции правдоподобия может означать как и аномальное значение, так и неудачно подобранную модель. Этот метод применяется с опорой на априорную информацию, в случае когда можно проверить полученную модель на адекватность.

1.6.2 Линейный подход

Основной идей линейного подхода является построение некой модели, характеризующей нормальные данные. Точки, которые значительно отклоняются от этой модели, считаются аномалиями.

Предполагается, что нормальные данные находятся в подпространсве пространства атрибутов данных(размер подпространства атирбутов данных равен размерности данных). В свою очередь, задача линейного метода - найти низкоразмерное подпространства, такие что, выборка данных этого подпросранства значительно отличается от остальных точек пространства данных.

Одним из возможных вариантов решения является использование линейной регрессии. Выбирается одна из наблюдаемых переменных набора данных и относительно неё решается задача линейной регрессии оставшихся атрибутов. Итоговым ответом будет является усредненное значения показателя аномалии по всем атрибутам.

Алгоритмы, основанные на линейном подходе, требуют наличия линейной зависимости атрибутов данных.

1.6.3 Метрические методы

Мерические методы пытаются найти в данных точки, в некотором смысле изолированные от остальных[4]. Если в пространстве задана некоторая метрика *p(x1, x2)*, то необходимо задать следующие понятия:

— Аномалии – точки, не попадающие ни в один кластер. К данным применяется один из алгоритмов кластеризации; размер кластера, в котором оказалась точка, объявляется её показатель аномальности.

— Локальная плотность в аномальных точках низкая. Для данной точки показателем аномальности объявляется локальная плотность, которая оценивается некоторым непараметрическим способом.

— Расстояние от данной точки до ближайших соседей велико.

В качестве показателя аномальности может выступать:

— расстояние до k-го ближайшего соседа;

— среднее расстояние до k ближайших соседей;

— медиана расстояний до k ближайших соседей;

— гармоническое среднее до k ближайших соседей;

— доля из k ближайших соседей, для которых данная точка является не более чем k-ым соседом и много другое.

1.6.3.1 Базовые понятия

Метрические методы хорошо подходят в случае когда данные не размечены. Сложность вычисления прямо как пропорциональна размерности данных m,как и их количеству n. При росте набора данных наблюдается экспоненциальный рост сложности вычислений. Однако, эти методы хорошо проявляют себя на ограниченных наборах данных[10]. Следовательно такие методы как k-ближайших соседей(так же известный как обучение на основе примеров, и описанный поздее) с нотацией ассимтпотического роста $O(n^2m)$ недопустимы для наборов данных с большой размерности, если их размерность не может быть уменьшена.

Существуют много различных вариации алгоритма k-ближайших соседей для обнаружения аномалий, но все они основаны на вычислении некой метрики "расстояния до соседей такой как Евклидово расстояние или расстояние Махаланобиса. Евклидово расстояние задается следующей формулой:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1.8}$$

и является просто расстоянием между двумя точками, когда как расстояние Махаланобиса, задаваемое следующей формулой

$$\sqrt{(x-\mu)^T C^{-}1(x-\mu)} \tag{1.9}$$

вычисляет расстояние от точки до центра тяжести (μ), определяемого формулой корреляированных атрибутов, заданных матрицей ковариации (C). Расстояние Махаланобиса рассчитывается значительно дольше по сравнению с евлидковым по сравнению

с евклидовым расстоянием для для больших объемов данных, поскольку оно требует пройти через весь набор данных, чтобы идентифицировать корреляции атрибутов.

1.6.3.2 Оптимизация Рамасвани

Точка p является выбросом если не более n - 1 других точек в наборе данных имеют более высокий D_m(расстояние до m соседей), где m задается. Например на рисунке 1.2 черная точка является наиболее удаленной от соседей, следовательно она является выбросом. Красные точки расположены рядом друг с другом, однако расстояние до других точек велико, следовательно они тоже являются аномалиями. Такой подход вприимчив к вычислительному росту, потому что должна быть вычислена матрица расстояний точек друг от друга, поэтому Рамасвани в 2000 году предложил оптимизацию метода k-ближайших соседей(с англ. k-Nearest Neighbour - k-NN) в виде составления ранжированного списка потенциальных выбросов.

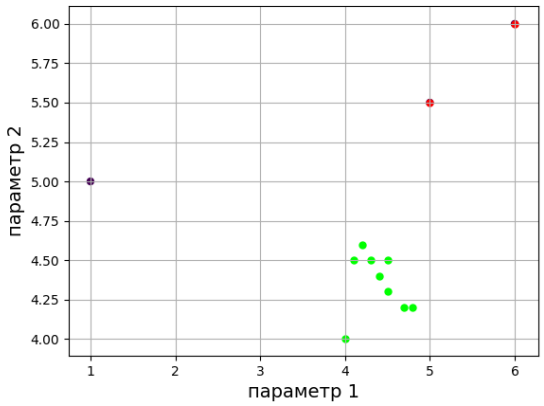

Рисунок 1.2 — Пример k-ближайших соседей

Оптимизация Рамасвани заключается в разбиении данных на ячейки.Если какая-либо ячейка и ее ближайшие соседи содержат больше, чем k точек, то точки в ячейке считаются лежащими в плотной области поэтому содержащиеся точки вряд ли будут выбросами. Если же почти все ячейки содержат больше,чем k точек, а какие-то ячейки содержат меньше, чем k точек, то тогда все точки, лежащие в ячейках, которые содержат менее k элементов, помечаются аномальным. Следовательно, необходимо обработать только небольшое количество ячеек, которые ранее не были помечены и только относительно небольшое количество расстояний необходимовычислить для обнаружений аномалий.

1.6.3.3 Методы Кнора-Реймонда и Байерса-Рейтери

Кнор и Реймонд предложили свой эффективный метод KNN подхода обучения без учителя[11]. Если m из k ближайших соседей (где m<k) лежат в пределах определенного порогового значения d, тогда считается, что данные точки лежат в

достаточно плотной области распределения данных, подлежащей классификации и подлежат классификации как нормальные, в противном случае они помечаются как аномальные.

Очень похожий метод был придуман для идентефикации наземный мин на спутниковых снимках поверхности Земли Байеросом с соавторстве с Рейтери[12](этот метод можно использовать и для других целей) Он заключается в том, что берется m точек, для них ищется расстояние D_m. Если расстояние меньшего некого порогового значеня d,тогда считается, что данные точки лежат в достаточно плотной области распределения данных, подлежащей классификации и подлежат классификации как нормальные, в противном случае они помечаются как аномальные. Этот метод уменьшается количество варьируемых параметров, по сравнению с методом Кнора-Реймонда: остаются только параметры d и m, параметр k убирается. подход оригинальный подход k-NN, поскольку только k ближайших соседей должны быть вычислены для каждой точки, а не всей матрицы расстояния для всех точек

1.6.4 Метод Танга

Метод Танга заключается в вычислении среднюю цепочку расстояний между точку p и k её соседями. Ранним расстояниям присваиваются более высокие веса, поэтому, если точка находится в разреженном область как черная точка на рисунке1.2, то путь до ее ближайши соседей будет относительно далеким, а среднее расстояние цепочки будет высоким. Этот метод выгодной отлчичается от вышеописанных тем, что учитывает как плотность, так и изоляцию. Рассмотрим рисокок 1.3. Очевидно, что черные точки являются аномалиями, а скопление зеленых точек - множеством "нормальных"точек. Однако, алгоритмы k-NN классификации могут столкнутся с проблемой того, что расстояние от черных точек до зеленого кластера примерно равно, значит эти точки можно отнести к одной группе и при определенных значениях параметров алгоритма эти точки не будут считаться аномалиями. Метод Танга поможет избежать таких ошибок при обнаружении выбросов. Однако метод является вычислительно сложным с временем выполнения как у оринального k-NN, поскольку он полагается на вычисление путей между всеми точками и их k соседей.

1.6.5 Параметрическе методы

Вышеописанные методы плохо подходят для работы с большим объемом данных. Параметрические методы позволяют очень быстро пересчитывать модель для новых данных и подходит для больших наборов данных; модель растет только с сложностью модели, а не размером данных. Однако они ограничивают применимость, применяя предварительно выбранную модель распределения для проверки данных на аномальность. Т.е. предварительно априорно подбирается модель правдо-

подобности данных. Элементы , которые значительно отклоняются от этой модели считаются аномальными. Параметрический подход схож с линейным по описанию, но значительно отличается от него по принципу работы.

Одним из таких подходов является оценка эллипсоидой минимального объема[13], которая соответствует наименьшему допустимому эллипсоиду, покрывающему не меньше 50% точек выборки.

1.6.6 Локальный коэффициент выбросов(LOF)

Этот метод является одним из самых известных алгоритмов обнаружения локальных аномалий. Недостком метрических методов является тот факт, что все лежащие в их основе предположения верны лишь в дополнении друг с другом: локальная плотность точки, лежащей в центре небольшого кластера аномалий, может оказаться выше, чем для любой точки из большого кластера нормальных данных. Возможно и обратное: изолированная точка-аномалия мо- жет располагаться, например, в центре масс кластера нормалий, и тогда среднее расстояние от неё до соседей будет меньше, чем для нормальных точек. Это "свойство"метрических алгоритмов пытается учесть алгоритм локального коэффициентов выбросов(англ.Local Outlier Factor).

Чтобы вычислить LOF необходимо произвести следующие действия:

а) Для каждой записи найти всех соседей, расстояния до которых не превышает k. Их количество может быть больше, чем k.

б) Используя эти записи для каждой точки N_k, вычислить локальную плотность точки, основанную на локальной плотности достижимости(англ. local reachability density (LRD)):

$$LRD_k(x) = 1/(\frac{\sum_{o \in N_k(x)} d_k(x,o)}{|N_k(x)|}) \tag{1.10}$$

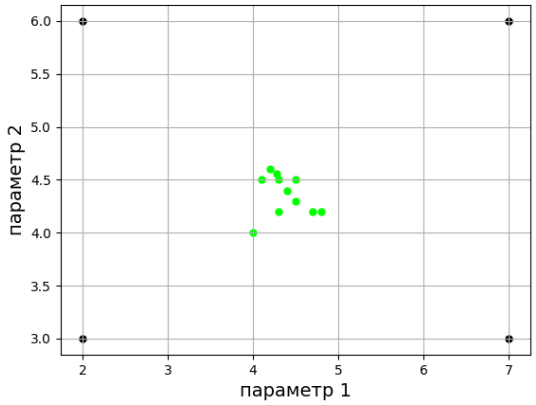

Рисунок 1.3 — Пример для метода Танга

где $d_(x,o)$ расстояние достигаемости. За редким исключеним в качестве расстояния достигаемости используется евклидово расстояние [14]

в) Вычисляем LOF путем сравнения LRD записи с LRD соседей.

$$LOF(x) = \frac{\sum_{o \in N_k(x)} \frac{LRD_k(o)}{LRD_k(x)}}{|N_k(x)|} \tag{1.11}$$

Таким образом LOF является отношением локальных плотностей. Нормальные записи, плотности которых равны плотности их соседей, получают оценку около 1,0. Аномалии, которые имеют низкую локальную плотность, получат значительно более высокую оценку. После этого становится, почему этот алгоритм называется локальным: он полагается только на свою прямую окрестность, а оценка - это величина, основанная основанное только на k-соседях. Конечно, глобальные аномалии также могут быть обнаружены, так как они также имеют низкую LRD,по сравнению со своими соседями. Важно отметить, что в задачах обнаружения аномалий, где местные аномалии не представляют интереса, этот алгоритм будет генерировать множество ложных аномалий, как мы выяснили во время нашей оценки. Опять же, настройка k имеет решающее значение для этого алгоритма.

Авторы алгоритма LOF рекомендуют использовать для вычисления k стратегию ансамблирования(алгоритм описан ниже). Берется интервал возможных значений k и с некоторым шагом для всех возможных значений k вычисляется показатели аномальности для каждого элемента выборки. Путем голосования определяется является ли эта ли эта точка аномалий. Однако, на практике такие рекомендации редко исользуют из-за их значительной вычислительной сложности.

1.6.6.1 Компонентный коэффициент выбросов(COF)

Компонентный коэффициент выбросов аналогичен LOF, но оценка плотности для записей выполняется иным способм В LOF k-ближайших соседей выбирают

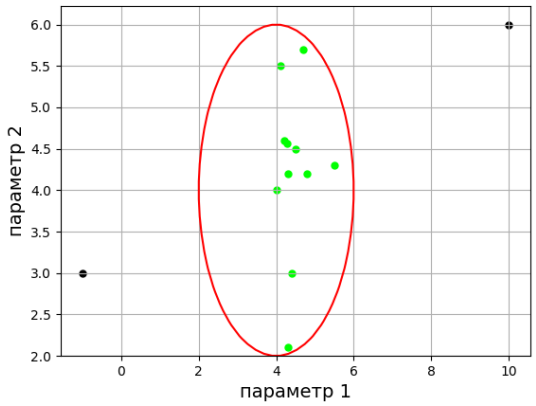

Рисунок 1.4 — Двухмерная проекция эллипсоиды минимального объема

на основе евклидова расстояния. Это косвенно предполагает, что данные распределяются сферическим образом вокруг экземпляра. Если это допущение нарушено, например, если функции имеют прямую линейную корреляцию,то оценка плотности неверна. COF исправляет этот недостаток и оценивает локальную плотность окрестности с использованием метода кратчайшего пути, называемого расстоянием цепочки. Математически это расстояние цепочки является минимумом суммы всех расстояний, соединяющих все k соседей точки и саму точку. Например, когда функции, очевидно, коррелированы, этот подход оценки плотности работает значительно лучше [15]. ..... На рисунке 5 показан результат для LOF и COF в прямом сравнении для простого двумерного набора данных, где атрибуты имеют линейную зависимость. Можно видеть, что оценка сферической плотности LOF не может обнаружить выброс, но COF удалось подключить нормальные записи друг к другу для оценки локальной плотности.

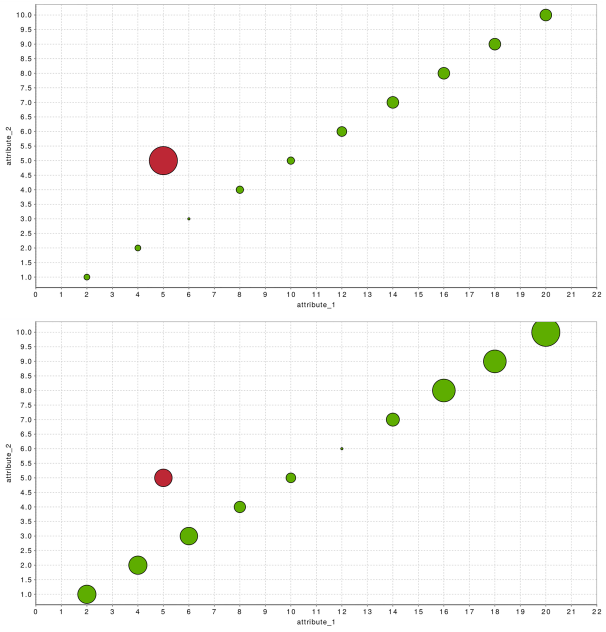

Рисунок 1.5 — Сравнение COF (сверху) с LOF (внизу) с использованием простого набора данных с линейной корреляцией двух атрибутов

1.7 Методы улучшения алгоритмов

1.7.1 Семпилрования

Большинство алгоритмов распознавания аномалий успешно работают на выборках малых размеров. Поэтому предлагается разбить начальный набор данных на несколько случайных выборок и усреднить результат. Размер этих выборок может быть как и случайным, там и фиксированного размера, но, как правило, он отличается от размеров исходного набора данных не меньше чем на порядок. Идея такого выбора заключается в том, что шумовые объекты попадут в выборки с низкой веро-

ятностью; кластера нормальных данных будут представлены несколькими представителями, а кластера аномалий выродятся в изолированные точки. На основе этих выборок алгоритмы строят функции показателя аномальности, незначительно уступающему результату, полученному на основе анализа всех исходных данных.

Этот метод помогает значительно сократить вычислительную сложность, а так же уменьшить вероятность "подгона"алгоритма под конкретный набор данных. В силу особенностей задачи, необходимое условие отсут- ствия параметризации алгоритмов зачастую означает их детерминированность(в отсутсвиее стохастичности показатель аномальности однозначно определяется по заданной выборке). В общем случае при добавлении новых данных в общий набор данных, можно не пересчитывать заново показатель аномальности для всего набора данных, а добавить запуски алгоритма на новых данных в ансамбль(так называемый warm start[16])

1.7.2 Ансамблирование

Ансамблированием в задаче поиска аномалий называют использование нескольких различных алгоритмов с последующи усреднением их показателя аномальности. При использовании различных алгоритмов можно столкнуться с проблемой того, что показатель аномальности выглядит по-разному в различных алгоритмах и сравнить напрямую эти показатели некорректно. Поэтому традиционное приведение показателей значений различных функций к одному диапазону, например,к [0,1], будет некорректным. На практике наиболшего успеха можно добиться ансамблированием одного и того же алгоритма(как предлагали авторы алгоритма LOF). Объединение результатов разнородных алгоритмов стоит проводить при помощи "голосования": для каждой точки исходных данных каждый алогритм решает аномалия это или нет и на основе решения большинства принимается решение о аномальности конкретно точки

1.7.3 Итеративный отбор

Итеративный отбор основам на идее многократного применения алгоритмов ансамблирования. Преположим, построеная некоторая модель, описывающая нормальные данные. Эта модель построена на основе всех имеющихся данных, но точность этой модели невелика, она умеет определить только явные аномалии. Отсортировав все точки по показателю аномальности, мож- но выбрать k самых аномальных объекта в данных и исключить из данных. После этого можно перестроить модель и повторить вышеуказанные действия несколько раз, пока не будут достигнуты некоторые условия. При каждой итерации точность модели будет увеличиваться.

Идея итеративного отбора может быть обобщена различными способами. Результат работы одного алгоритма может быть использован для отсеивания явных

аномалий и настройки нового алгоритма, не обязательно совпадающего с предыдущим, на оставшихся данных. х. Возможна и противоположная механика: по результатам работы одного алгоритма отбираются явные, гарантированные представители нормальных данных, и исключительно на них строится модель, их описывающая

1.8 Выводы

Существует большое число алгоритмов для нахождения аномалий. Некоторые из них опирается на априорные данные, некоторые не опираются. Для выбора подходящего алгоритма нахождения аномалий зачастую стоит учитывать характер данных, их размер и доступную априорую информацию. Несмотря на то, знаний обнаружения аномалий активно развивается как часть современной науки, остается ещё много простора для исследования алгоритмов, модификации и создания новых.

2 Конструкторский раздел

3 Технологический раздел

4 Исследовательский раздел

4.1 Время дизеринга раличных алгоритмов

Рассмотрим время работы различных алгоритмов для различных размеров изображения.

	Размер, пиксели	Время, мкс
White noise	133x90	862
Blue noise	133x90	930
Brown noise	133x90	934
Violet noise	133x90	937
Pink noise noise	133x90	930
Floyd-SD	133x90	1200
F. Floyd-SDe	133x90	1093
JJN	133x90	1909
White noise	458x458	15735
Blue noise	458x458	19374
Brown noise	458x458	19432
Violet noise	458x458	18787
Pink noise noise	458x458	18129
Floyd-SD	458x458	27173
F. Floyd-SDe	458x458	26424
JJN	458x458	47201
White noise	458x458	194376
Blue noise	458x458	200577
Brown noise	458x458	208400
Violet noise	458x458	251294
Pink noise noise	458x458	258775
Floyd-SD	458x458	251294
F. Floyd-SDe	458x458	387104
JJN	458x458	857481

Из рассмотрения вынесены алгоритм Юлиомы в вследствие того, что он значительно медленней других алгоритмов(2732568 мкс для изображения 113x90) в и алгоритм Байера, реализованный при помощи шейдеров, вследствии того, что он не не укладывается в рамки требуемой палитры (при этом он работает очень быстро 64 мс для изображени 640x480).

4.2 Качество получаемого изображения

	PSNR	SSIM
White noise	33.2894	0.914778
Blue noise	36.1756	0.971626
Brown noise	33.32370	0.915767
Violet noise	37.63480	0.984574
Pink noise	36.4484	0.974718
Floyd-SD	37.0553	0.979173
F. Floyd-SDe	36.8401	0.976452
JJN	37.30740	0.981688
Yliouma	36.2359	0.967796
Without dithering	37.6348	0.984574

Несмотря на то, что некоторые сложные алгоритмы дизеринга диффузии ошибок обещат получения хорошего качества изображений,некоторые алгоритмы случайного дизеринга на конкретных изображениях дают лучший результат. Для того чтобы получить наилучший результат дизеринга, следует проанализировать результаты дизеринга нескольких изображений и выбрать среди них наилучшее. Так же следует отметить некоторую необъективность метрик: результат метрик не всегда совпадает с человеческим восприятием картинки.

4.3 Размер получаемого изображения

	Разрешение, пикс	Размер, кб	Исх. раз., кб
White noise	900x675	186	2373 bmp,1779 png
Blue noise	900x675	135	
Brown noise	900x6750	186	
Violet noise	900x675	98	
Pink noise	900x675	1158	
Floyd-SD	900x675	1273	
F. Floyd-SDe	900x675	143	
JJN	900x675	117	
White noise	3984x32355	3431	50344 bmp,37758 png
Blue noise	3984x3235	2570	
Brown noise	3984x3235	3432	
Violet noise	3984x3235	1950	
Pink noise	3984x3235	2406	
Floyd-SD	3984x32355	3605	
F. Floyd-SDe	3984x3235	4269	
JJN	3984x3235	3716	

Из вышеприведенной таблицы, можно заметить, размер изображения после дизеринга значительно уменьшиается,достирается выигрыш в размере изображение до 15 раз, в зависимости от иходного контейнера изображения и выбранного способра дизеринга.

Заключение

В данной работе были реализованы различные алгоритмы дизеринга, было произведено сравнение и анализ этих алгоритмов. Программа позволяет получить изображение схожего визуального качества при значительном уменьшении размера. Был получен вывод о том, что для различных целей следует использовать различные алгоритмы дизеринга, универсального алгоритма дизеринга не существует. Программа не привязана к какой-то конкретной операционной системе и может быть скомпилирована и запущена на всех популярных ОС.

Список использованных источников

1. *machinelearning.ru/*. Выборка. https://goo.gl/7gjJ6p.

2. *определения, ГОСТ 20886-85: Организация данных в системах обработки данных. Термины и.* http://www.gostrf.com/normadata/1/4294832/4294832686.pdf.

3. *Википедия*. https://en.wikipedia.org/wiki/Data_set.

4. *Дьяконов, Александр*. Поиск аномалий (Anomaly Detection) / Александр Дьяконов. — 2017. https://goo.gl/Z43Ne9.

5. *F.E., Grubbs*. Procedures for Detecting Outlying Observations in Samples. Technometrics / Grubbs F.E. — 1969.

6. *Moya M.M., Hush D.R.* Network Constraints and Multi-objective Optimization for One-class Classification. Neural Networks / Hush D.R. Moya M.M. — 1996.

7. *Chandola V Banerjee A, Kumar V.* Anomaly Detection: A Survey / Kumar V. Chandola V, Banerjee A. — ACM Computing, 2009.

8. *Goldstein M, Uchida S.* Behavior Analysis Using Unsupervised Anomaly Detection / Uchida S. Goldstein M. — The 10th Joint Workshop on Machine Perception and Robotics, 2014.

9. *Андрей, Гахов*. Интеллектуальный анализ данных / Гахов Андрей. — Харьковский национальный университет имени В.Н. Карамзина, 2014.

10. *Hodge V., Austin J.* A survey of outlier detection methodologies / Austin J. Hodge V. — Artificial intelligence review, 2004.

11. *Knox, Edwin M.* Algorithms for Mining Distance-Based Outliers in Large Datasets / Edwin M. Knox, Raymond T. Ng. — University of British Columbia, 1998.

12. *S. Bayers, A.Raftery*. Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes / A.Raftery S. Bayers. — Journal of the American Statistical Association, 1998.

13. *Rousseeuw, Leroy*. Robust Regression and Outlier Detection / Leroy Rousseeuw. — John Wiley and Sons, 1996.

14. *Goldstein, Markus*. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data / Markus Goldstein. — Seiichi Uchida, 2016.

15. *M., Goldstein*. Anomaly Detection in Large Datasets / Goldstein M. — University of Kaiserslauterna, 2014.

16. *B.Chu Chia-Hua Ho, Cheng-Hao Tsa*. Warm Start for Parameter Selection of Linear Classifiers / Cheng-Hao Tsa B.Chu, Chia-Hua Ho. — National Taiwan University, 2015.