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 дипломной работе:

Метод обнаружения выбросов временных рядов

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

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

Москва 2016

Введение

Задача поиска аномалий является одной из классических задач машинного обучения. 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 вышеперечисленных 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 Обзор предметной области

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:

— пронализировать предметную область и существующие методы обнаружения выбросов

— разработать метод обнаружения выбросов

— создать ПО, собирающее 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, которые занимаются поиском аномалий - это детектирование новизны и обнаружение выбросов. "Объект новизны" - 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 анализировать замеры уровня шума и отбрасывать слишком высокие или слишком низкие значения, то это называется борьбой с выбросами. А если создаётся алгоритм, который для каждого нового замера оценивает, насколько он похож на прошлые, и выбрасывает аномальные, то это назвается "борьбой с новизной" . [1]. Выбросы являются следствием:

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

На рисунке 1.1 находится три вида точек: зеленые, желтые, красные. Множество зеленых точек представляют собой "нормальные"данные. Множество желтых точек означает выбросы в "слабом смысле". Они незначительно отклоняются от основных нормальных данных. Красные же точки являются аномальными - выбросами "в сильном смысле" . Они значительно отклоняются от нормальных 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 - част-

ный случай аномалии). Понятие аномалии 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. Вероятно, первое определение было дано Граббсом[2] в 1969 году: "Относительное наблюдение или выброс - это элемент выборки, который, заметно отличается от других членов выборки, в которых он встречается ". Это определение является актуальным и сегодня, но мотивация для обнаружения аномалий изменилась. Тогда основная причина поиска аномалий заключалась в том, чтобы удалить выбросы из набора данных для обучения, поскольку используемые алгоритмы, были весьма чувствительны к выбросам в данных. Эта процедура также называется очищением данных. После разработки классификаторов устойчивых к наличию аномалий в обучающем наборе данных, интерес к их поиску угас. Однако, в начале 21 века в связи с развитием интернета и значительным увеличением объема собираемых данных для анализа, исследователи стали больше интересоваться аномалиями, поскольку они оказывались часто связаны с особенно интересными событиями. В этом контексте определение Граббса также было расширено, так что сегодня аномалии 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

Классическая система классификации предполагает предварительное обучение на обучающем наборе данных и последующую классификацию на основе этого набора. Данные делятся на "обучающую выборку" - данные, при помощи которых алогритм обучает классификар и, "тестовую выборку" - данные, при анилизе которых, классификатор остается неизменным.Тестовая выборка нужна для того чтобы проверить корректность обучения классификатора.

Однако, в случае с поиском аномалий, возможны варианты, отличающиеся от классического. Подходящий метод классификации выбирается на основе наличия разметки данных. Выделяются три основых класса методов:

а) Обучение с учителем. Для обучения необходимо наличие полностью размеченных данные для обучения и для тестов. Классификатор обучается один раз и

применяться впоследствии.В связи с тем, что для многих наборов данных заранее неизвестно что является аномалией, а что нет, применение этого метода ограничено.

б) Обучение с частичным привлечением учителя. Для обучения необходимо наличие тестового и учебного набора данных. Однако, в отличие от обучения с привелечением учителя, разметка данных не требуется. Все данные, представленные в выборках, считаются нормальными. На основе этих данных строится некая модель. Все данные, отклоняющиеся от этой модели, считаются аномальными. Эта идея также известна как "одноклассовая"классификация [3].

в) Обучение без учителя. Самый гибкий способ, который не требует разметки набора данных. Идея заключается в том, что алгоритм обнаружения аномалий оценивает данные исключительно на основе внутренних свойств набора данных что является нормальным, а что является выбросом. В этой работе основное внимание будет этому именно этому способу. Так же этот способ называют "неконтролируемый способ обнаружения аномалий".

1.4 Результат метода обнаружения аномалий

В результате работы алгоритма обнаружения аномалий с элементом данных связаывается метка или оценка достоверности(показатель аномальности). Метка- показатель, который принимает нулевое значения, в случае если она связана с нормальными данными и единицу в противном случае. Оценка показывает вероятность того, что элемент является аномалией. Для разных алгоритмов используется разные шкалы оценок, поэтому приведение конкретных примеров оценок будет некорректным. В алгоритмах метода обучения с учителем зачастую используются метки как выходные данные, в алгоритмах с частичным привлечением учителя и без учителя обнаружения аномалий чаще встречаются оценки.

1.5 Виды аномалий

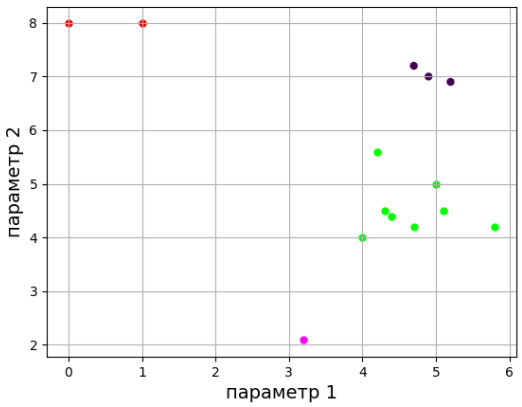

Рисунок 1.1 — Простой двумерный пример

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. Если смотреть на весь набор данных в целом, то фиолетовую точку можно отнести к тому же классу, что и зеленые точки. Однако, если сфокусироваься только на кластере зеленых точек и сравнивать его с фиолетовой точкой, 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 анализа, локальные аномалии могут представлять интерес или нет. Другой вопрос заключается в том,что следует ли рассматривать точки черного кластера как три аномалии или как (небольшой) кластер. Такие небольшие кластеры явления называются микрокластерами. Показатели аномальности у точек этого кластера выше, чем у точек зеленого кластера, но меньше, чем у красных точек. Этот простой пример показывает, что задача нахождения аномалий аномалии не всегда тривиальна, а вычисление показателя аномалности иногда полезнее, чем проставление двоичной метки.

Задача обнаружения одиночных аномальных экземпляров крупном наборе данных называется обнаружением точечных аномалии[4]. Сегодня почти все неконтролируемые алгоритмы обнаружения относятся к этому типу. Если же аномалии составляют заметный процент, от набора данных, то задачу поиска аномалий называют задачей обнаружения коллектнивных аномалий. Пусть аномалии представляют собой некое множество, тогда необязательно каждый элемент этого множества должен быть аномальным. Возможен вариант когда только определенная их комбинация определяет аномалию. Третий вид - контекстуальные аномалии. Элемент выборке в отрыве от своего контекста может казаться нормальным. Однако, если рассмотреть контекст этого элемента, то очевидным станет его аномальная природа. Распространенным контекстом является время. В качестве примера предположим, что измеряется температура в диапазоне от $-30°C$ до $+40°C$ в течение года. Таким образом, температура $25°C$ кажется довольно нормальной, но когда учитывается 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 контекст в алгоритм как параметр алгоритма. В вышеприведенным примере включение месяца как дополнительного параметра поможет обнаружить анома-

лию. Однако в более сложных сценариях может потребоваться один или несколько новых парметров, чтобы преобразовать задачу определения контекстной аномалии в задачу обнаружения точечной аномалии. Для того, чтобы преобразовать задачу поиска коллективной аномалии в задачу поиска одиночной, нужно произвести изменения изначального набора данных. Для этого можно использовать корреляцию, агрегация и группировка. Преобразование может быть нетривиальным.[5] . Преобразование 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.

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, область знаний обнаружения аномалий активно развивается как часть современной науки, остается ещё много простора для исследования алгоритмов, модификации и создания новых.

2 Набор функций разрабатываемого ПО

Разрабатываемое ПО будет получать в качестве входных данных файл, содержащий некоторое неразмеченнных данных. В качестве выходных данных программа будет выводить набор размеченных данных.

Программа для сбора данных будет содержать функционал сбора телеметрии с графического редактора Krita, их агрегации и сохрании в базе данных.

3 Технологические средства

3.1 Выбор средств разработки

3.1.1 Язык программирования и средства разработки

Для реализации данных алгоритмов был выбран язык C++. Данный язык был обоснован следующими причинами: Причины:

а) Компилируемый язык со статической типизацией.
б) Сочетание высокоуровневых и низкоуровневых средств.
в) Реализация ООП.
г) Наличие удобной стандартной библиотеки шаблонов

В качестве средств разработки была выбрана Qt Creator, поддерживающая все возможности языка C++ и имеющий инструментарий для создания как консольных приложений, так и приложений с графическим интерфейсом.

3.1.2 Программа для сбора данных

В качестве языка написания плагина к графическому редактору был выбран язык C++ в силу того, что сам редактор поддерживает только плагины на этом языка. Для написания бекенд-сервера был выбран язык Golang вместе с библиотекой mgo для доступа к базе данных . Плюсы этого языка:

а) Скорость разработки
б) Производительность
в) Удобная реализация легковесных потоков

Список использованных источников

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

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

3. *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.

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

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