## Введение

Будет сформулирована математическая постановка задачи обнаружения аномалий поведения пользователей, а также предложен алгоритм, позволяющий решить поставленную задачу, произведён выбор метрики, используемой в данном алгоритме.

## Математическая постановка задачи

Пусть – множество объектов, определяющих поведение пользователя, – множество оценок. Существует неизвестная целевая зависимость – отображение , значения которой известны только на объектах тестовой выборки .

Дано:

* исходная выборка объектов ;
* тестовая выборка объектов .

Ограничения:

* объекты исходной выборки должны характеризовать обычное поведение пользователя, не содержащее аномальных действий, и быть близки по метрике , определяющей степень сходства объектов.

Требуется:

* построить алгоритм , способный приблизить неизвестную целевую зависимость на всём множестве ;
* из заданного множества метрик выбрать метрику , минимизирующую эмпирический риск на тестовой выборке объектов, формула (2.1):

, (2.1)

где – эмпирический риск, характеризующий среднюю квадратичную ошибку алгоритма , использующего метрику , на выборке , вычисляемый по формуле (2.2).

. (2.2)

## Алгоритм обнаружения аномалий

Входные параметры алгоритма:

* исходная выборка объектов , характерных для обычного поведения пользователя, которая задаёт матрицу исходных данных размера ;
* объект , описывающий неизвестное поведение текущего пользователя.

Ход работы алгоритма:

1. Предобработка данных.
2. Учёт односторонних отклонений.
3. Изменение весов.
4. Измерение расстояния от множества объектов исходной выборки до объекта .
5. Получение оценки .
6. Объяснение полученного результата.

Выходные параметры алгоритма:

* оценка , которая характеризует степень аномальности поведения, описанного объектом ;
* вектор , определяющий вклад каждого признака в итоговую оценку.



### Признаковое описание объекта

В исходной задаче – множество объектов. Под объектом будем понимать поведение конкретного пользователя. Отметим, что подразумевается под понятием «поведение» пользователя. Во-первых, необходимо определить интервал проверки, по происшествии которого можно будет сравнить действия пользователя с эталонными. Это может быть час, день, неделя или любой другой временной интервал. Во-вторых, необходимо определить те действия пользователя, которые будут учитываться при получении оценки. В данной работе в качестве интервала, за который оценивается поведение пользователя, был выбран день, поскольку данное программное обеспечение подразумевает работу в корпоративных информационных системах, а сотрудники обычно работают в течение всего рабочего дня практически непрерывно.

Определим признаковое описание объекта , характеризующего поведение пользователя, которое представляет собой вектор, составленный из набора признаков данного объекта. В качестве признаков возьмём некоторые количественные характеристики объектов. Рассмотрим набор признаков , используемый в данной работе:

* время первого захода в систему;
* время выхода из системы;
* длительность всех сессий;
* суммарный размер исходящего трафика;
* количество уникальных IP-адресов для исходящего трафика;
* суммарный размер входящего трафика;
* количество уникальных IP-адресов для входящего трафика.

Тогда вектор будет называться признаковым описанием объекта , а совокупность признаковых описаний всех объектов исходной выборки будет записана в виде таблицы размера – матрицы исходных данных. Признаковые описания отождествляются с самими объектами.

### Предобработка данных

Перед тем как считать расстояние между объектом , характеризующим неизвестное поведение, и множеством объектов исходной выборки , необходимо найти среднее выборочное выборки по формуле (2.3):

(2.3)

Искомое расстояние будет определяться как расстояние от объекта до выборки со средним значением .

### Метрика Махаланобиса

Имеется признаковое описание объекта, теперь определим, чем характеризуется аномальное поведение. Логично предположить, что аномалия – это событие с низкой вероятностью. Можно посмотреть на график выборочной функции распределения и определить вероятность любого события. Если предположить, что переменная имеет нормальное распределение, то события с низкой вероятностью будут находиться вдали от среднего значения. То есть чем ниже вероятность события, тем дальше от центра оно будет находиться. Поиск событий с низкой вероятностью эквивалентен поиску выбросов, которые располагаются далеко от других значений. Расстояние от среднего значения до выброса является мерой величины аномалии.

Одним из способов поиска выбросов являются метрические методы, они позволяют найти точки, находящиеся дальше всего. В метрических пространствах схожесть признаковых описаний объектов определяется через норму расстояния. В таком случае аномалиями будут точки, которые более всего удалены от центра. Существуют различные виды метрик, рассмотрим метрику Махаланобиса.

(2.10)

где – матрица ковариации, а её элементы вычисляют по формуле (2.11):

(2.11)

В том случае, если матрица ковариации является единичной, то расстояние Махаланобиса является расстоянием Евклида, если матрица диагональная – нормализованным расстоянием Евклида.

### Оптимизации алгоритма

Рассмотрим некоторые оптимизации, используемые в предложенном выше алгоритме. Они предоставляют возможность более точно настроить параметры алгоритма расчёта оценки для того, чтобы наиболее эффективно осуществлять обнаружение аномалий. Кроме этого, полученное расстояние необходимо перевести в оценку, удобную для понимания человеком.

#### Односторонние отклонения

Поиск аномалий в некоторых случаях заключается не в том, чтобы найти любые подозрительные значения, а только те, которые отклоняются от нормы в одну сторону. Например, при наблюдении за пользователем не так важно знать, что сотрудник скачал файлов меньше чем обычно, но если полученное значение будет значительно превышать обычную активность, то необходимо это учесть. Подсчёт расстояния между объектами не позволяет отличить случаи, когда величина имеет слишком высокое или низкое значение, поэтому необходимо дополнить существующий алгоритм.

Каждому из признаков объекта будет соответствовать значение, которое определяет, необходимо ли игнорировать положительные или отрицательные отклонения от выборочного среднего. Определим дополнительный вектор длины , который будет содержать значения -1, 0 или 1. Если -й элемент данного вектора равен единице, то значение -го признака, если оно превышает выборочное среднее, не будет учитываться при расчёте расстояния. Если -й элемент вектора равен -1, то не будет учитываться значение меньшее выборочного среднего. Если 0 – все величины будут учитываться при подсчёте расстояния. Обозначим данный вектор .

Для того чтобы учесть, в какую сторону отклоняется текущее поведение пользователя, необходимо первоначально вычесть из вектора признаков объекта среднее выборочное значение выборки согласно формуле (2.14).

(2.14)

Теперь, перед тем как рассчитывать расстояние, добавим дополнительный шаг, который определяется следующим выражением (2.15):

(2.15)

В данной формуле происходит проверка, отклоняется ли значение признака в ту сторону, которая не характеризует аномальное поведение. Если элемент вектора указывает на то, что данный признак не должен влиять на итоговую оценку, то значение признака приравнивается к нулю.

#### Вес переменных

Часто возникает необходимость изменения веса разных признаков для того, чтобы пользователь смог более точно настроить алгоритм. Например, необходимо уделить большее внимание времени доступа пользователей к рабочему месту, но при этом нам не так важна другая активность, как объём скачанных файлов. Изменение веса переменной приведёт к тому, что она будет вносить больший вклад в расстояние, чем другие признаки. При этом расположение объектов относительно друг друга останется прежним, изменится лишь масштаб. В качестве весов могут выступать экспертные оценки.

Обозначим вектор весов , где – вес -го признака. После того, как будут учтены односторонние отклонения, умножим значение -го признака на вес (2.16):

(2.16)

Изменение веса также может использоваться в качестве обратной связи в том случае, если полученная оценка является слишком высокой или низкой. Рассмотрим два случая: ошибки первого и второго рода. Поскольку предполагается, что обучающая выборка не содержит никаких объектов, соответствующих аномальному поведению пользователя, то в случае ошибки первого рода можно добавить объект, характеризующий ложное аномальное поведение, в исходную выборку. Тем самым, мы будем использовать его для дальнейшего анализа, чтобы похожие данные не определялись как аномальные. Данный метод не получится реализовать в случае ошибок второго рода, поэтому рассмотрим другой подход.

Чтобы оба случая обрабатывались корректно, необходимо отслеживать вклад признаков в общую оценку и изменять их вес. Например, если пользователь замечает ошибку, то необходимо найти признаки, которые вносят наибольший вклад, и изменить их веса. Для того, чтобы веса не были слишком большими или слишком маленькими, можно их ограничить максимальным и минимальным значениями.

#### Получение оценки

Расстояние не ограничено сверху, поэтому полученное значение может быть достаточно велико. Поскольку итоговая оценка должна быть в интервале от 0 до 100, то отобразим полученное расстояние на сигмоидальную функцию. Пусть – рассчитанное расстояние, тогда итоговая оценка будет выражена формулой (2.17):

, (2.17)

где – крутизна кривой;

– расстояние, которое соответствует оценке .

Для того, чтобы получить правильное отображение, необходимо зафиксировать величины и . Для этого воспользуемся зависимостью (2.18), взяв желаемую оценку для двух разных величин .

. (2.18)

#### Объяснение результатов

Важной частью алгоритма выявления аномалий является не только их обнаружение, но и предоставление информации о том, почему была получена именно такая оценка. Для этого необходимо определить, какой вклад в итоговую оценку вносит каждый из признаков. Предварительно проведём стандартизацию признаков объекта по формуле (2.19):

(2.19)

Теперь посчитаем значение , выраженное формулой (2.20), которое определяет вклад каждого признака объекта.

. (2.20)

### Блок-схема алгоритма

На рисунке 2.1 представлена блок-схема описанного выше алгоритма.

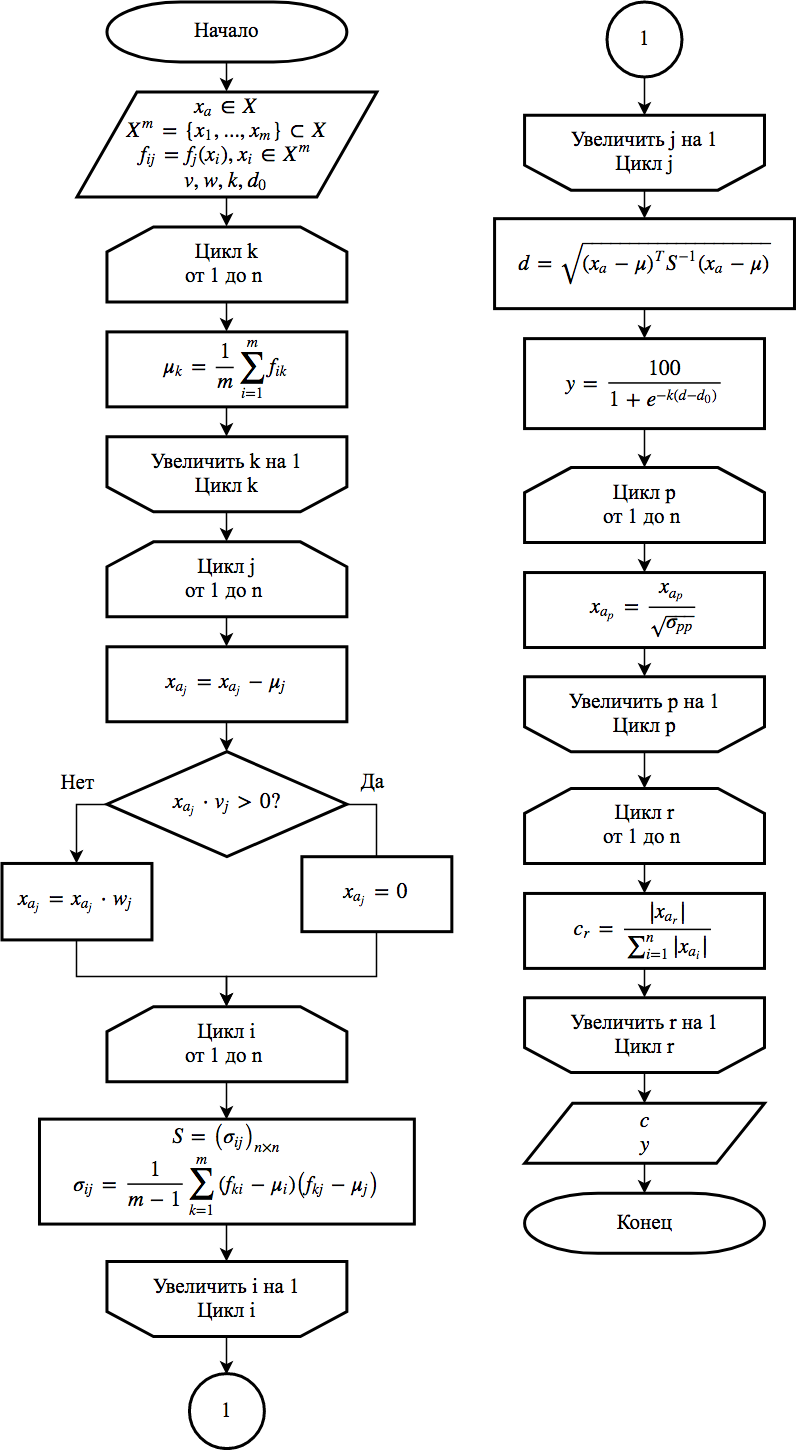


Рисунок 2.1 – Блок-схема алгоритма