ния в объекте исследований. Для аварийного сброса давления используется дополнительный клапан.

Устройство было опробировано на экспериментальных животных (мелких гризунах) и использовалось для создания внутрибрюшного давления порядка 15-20 мм рт. ст. и поддержание его в течение 30 минут. Начальное давление сотавляло 370-400 мм.рт.ст. при продолжительности заполнения брюшной полости животного около 2 минут.

Перспективой работы является создание микроконтроллерной системы с компьютерным управлением параметрами подаваемого воздушного потока.

Кобзев В.Г., Чернов А.Г.

АНАЛИЗ ВЫБРОСОВ И ПОИСК АНОМАЛИИ ДАННЫХ

В век современных информационных технологий и обилия информации актуальной является не только проблема хранения самих данных, но также (и прежде всего) проблема их анализа и обработки. При обработке больших массивов данных часто используются методы теории временных рядов.

В общем случае, временной ряд – это упорядоченная последовательность значений, описывающая протекание во времени какого-либо длительного процесса. Значениями временного ряда могут быть показания датчиков, цены на какой-либо продукт, курс валюты и т.п.

В данной работе решается задача определения аномалий в наборах временных рядов. Проблема определения или обнаружения аномалий формулируется как задача поиска в наборах данных отдельных образцов, не удовлетворяющих предполагаемому типовому поведению.

Определение аномалий может быть актуальным в различных сферах, например, для выявления дефектов оборудования, вторжений в инфокоммуникационные системы, банковского мошенничества, нарушений экосистемы, при анализе медицинских показателей и мониторинге исправности систем различного назначения. При решении задачи предотвращения вторжений выявление аномалий позволяет устанавливать факты злонамеренных действий. Определение аномалий часто применяют на этапе предварительной обработки для исключения из набора аномальных данных. При использовании методов управляемого обучения исключение аномальных данных из обрабатываемого набора приводит к статистически значимому улучшению точности результатов.

Аномалия (или выброс) определяется как элемент, явно выделяющийся из набора данных, к которому он принадлежит, нарушает статистические свойства распределения набора данных, и существенно отличается от других элементов выборки. Неформально задача определения аномалий в наборах временных рядов ставится следующим образом. Существует коллекция временных рядов, описывающих некоторые процессы. Требуется на основании имеющихся данных разработать алгоритм, который позволит различать нормальные и аномальные значения наблюдаемых процессов в реальном времени.

Алгоритм процесса обнаружения аномалий в данных:

1) считывание и первичная обработка полученных данных (удаление пустых значений);

2) визуальный анализ данных с целью получения первичной информации о временном ряде и возможных методах его анализа и обработки (построение графиков исходного временного ряда и гистограмм разброса данных);

3) переход к рассмотрению разностного ряда, эквивалентного исходному по интересующим нас характеристикам, для улучшения статистических свойств исходного временного ряда.

4) имея собственную обучающую выборку, построить ее разностный аналог с минимальным необходимым порядком дифференцирования для придания ряду свойства стационарности;

5) проведение дальнейшего анализа в режиме реального времени (последовательного поступления и обработки новых значений ряда) с применением статистического правила « ». С добавлением каждой следующей точки разностный ряд пополняется новой разностью и к нему применяется процедура анализа целостности данных по выбранному правилу. Если новое обрабатываемое значение выходит за пределы доверительного интервала, то оно считается выбросом.

Разностный ряд обладает свойством стационарности, гистограмма разброса данных разностного ряда демонстрирует приближение его распределения к гауссовому, что весьма расширяет границы допустимого к использованию статистического аппарата.

Поскольку анализ данных должен производиться в режиме реального времени, то возникает потребность в некоторой статистической базе, на основании которой возможно сформулировать представление о нормальном поведении наблюдаемого процессу, с дальнейшей целью оперативного распознавания значений ряда, которые не характерны данному процессу, т.е. являются аномальными.

Часто может проявляться еще одна проблема. Выбросы могут присутствовать и среди первых наблюдаемых значений изучаемого процесса. Для их распознавания предлагается использовать модификацию критерия Ирвина, адаптированную для работы с несортированными массивами данных, не всегда обладающими свойством нормально распределенной случайной величины, с более высокой эффективностью распознавания серий выбросов, чем у исходного алгоритма.

С целью обнаружения аномалий в наборах значений временных рядов разработан программный продукт с использованием мощного функционального языка программирования и среды анализа наборов статистических данных – R [1]. Результаты работы программы представляются в виде графиков временных рядов, с выделенными цветом на них точек выбросов с указанием их индекса и значения. Откровенные выбросы (нули или многократно завышенные значения) гарантированно обнаруживаются. Использование правила « » также позволяет обнаруживать и те значения, в которых наблюдается относительно большой скачок по отношению к предыдущим значениям ряда (но это не ошибка ввода, а скорее особенности наблюдаемой величины).

Данная программа адаптирована к легкому изменению, есть возможность расширить доверительный интервал (тем самым можно избежать ложных срабатываний в ситуациях скачка наблюдаемой величины) или настроить процедуру непрерывного обучения с целью более качественного анализа аномалий.

Описанная процедура обнаружения аномалий в данных может быть использована для мониторинга различного рода данных, поступающих в реальном времени. Кроме того, важным достоинством является возможность применимости алгоритма к данным, не имеющим предыстории, а также к данным, для которых отсутствует обучающие наборы.

Разработанный программный продукт может быть использован на практике для анализа интернет трафика и банковских транзакций, контроля тенденций спроса на товары, анализа финансовых временных рядов, рядов данных полученных с медицинских приборов и в других важных сферах жизни современного человека.

Список использованных источников

1. Роберт И. Кабаков. R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.