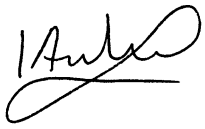
На правах рукописи



АНДРЭА Клитон

Методы и алгоритмы разведочного анализа данных, основанные на робастных модификациях боксплотов

Специальность 05.13.18 — Математическое моделирование, численные методы и комплексы программ

Автореферат

диссертации на соискание учёной степени

кандидата технических наук

Санкт-Петербург — 2013

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Санкт-Петербургский государственный политехнический университет»

Научный руководитель:	**ШЕВЛЯКОВ Георгий Леонидович** доктор физико-математических наук, профессор
Официальные оппоненты:	**ФИЛИМОНОВ Руслан Петрович** доктор физико-математических наук, профессор ФГБОУ ВПО «Санкт-Петербургский государственный университет кино и телевидения», профессор кафедры научной и прикладной фотографии
	КОКОРИН Сергей Владимирович кандидат технических наук ФГБУН Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН) младший научный сотрудник
Ведущая организация:	ФГБОУ ВПО «Санкт-Петербургский государственный университет»

Защита состоится 26 марта 2014 г. в 16 ч. 00 мин. на заседании диссертационного совета Д 212.229.13 при ФГБОУ ВПО «Санкт-Петербургский государственный политехнический университет», расположенном по адресу: 195251, Санкт-Петербург, Политехническая ул., д. 29, I уч. корп., ауд. 41.

С диссертацией можно ознакомиться в фундаментальной библиотеке ФГБОУ ВПО «Санкт-Петербургский государственный политехнический университет» по адресу: 195251, Санкт-Петербург, ул. Политехническая, д. 29, главный учебный корпус. Автореферат диссертации доступен на официальном сайте СПбГПУ (http://www.spbstu.ru/).

Автореферат разослан «____» февраля 2014 г.

Ученый секретарь
диссертационного совета Д 212.229.13
доктор технических наук, профессор  Григорьев Борис Семёнович

Общая характеристика работы

Актуальность темы. Разведочный анализ данных (РАД; Exploratory data analysis) – относительно новый раздел статистики, появление которого связано с развитием вычислительной аппаратуры и автоматизацией вычислений, сделавших возможным графическое представление больших объемов данных. Многие методы, лежащие в основе разведочного анализа данных, были известны задолго до появления работы[1] Дж. Тьюки (J. W. Tukey) в 1977 году, по которой и был назван этот раздел статистики. Вместе с Дж. Тьюки свой вклад в развитие и формирование РАД внесли Ф. Мостеллер[2] (F. Mosteller), Д. Хоаглин (D. Hoaglin), П. Веллеман (P. Velleman)[3]. В российской литературе этот раздел статистики был дополнен трудами[4] С. А. Айвазяна, В. М. Бухштабера, И. С. Енюкова и Л. Д. Мешалкина. Хотя и не существует строгого (точного) определения термина «разведочный анализ данных», основное назначение РАД заключается в следующем:

- Максимальное «проникновение в данные»;
- Выявление основных структур данных;
- Обнаружение отклонений и аномалий в данных;
- Проверка основных гипотез о распределении данных;
- Разработка начальных моделей распределений данных.

Задача обнаружения отклонений и аномалий является одной из целей разведочного анализа данных. В литературе представлены несколько трактовок понятия отклонений (выбросов), что сильно расширяет область исследования задачи выявления аномалий в данных.

Возможность сбора и хранения больших объемов информации в настоящее время требует применения эффективных методов первичного анализа и подготовки данных для дальнейшего изучения. Наше исследование направлено на разработку новых и улучшение существующих методов по обнаружению и отбраковке аномалий в данных. Классические методы обнаружения аномалий построены на статистических оценках, недостаточно устойчивых к выбросам. Предложенные нами методы основываются на новых робастных высокоэффективных оценках параметра масштаба.

В задачах статистической классификации оценка качества классификации связана со значениями критерия мощности и вероятности ложной тревоги согласно подходу Неймана-Пирсона. Проведение сравнения качества классификации исследуемых методов по двум параметрам затруднительно. Согласно подходу

[1] Tukey J. W. Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977.

[2] Mosteller F., Tukey J. W. Data Analysis and Regression. Addison-Wesley, 1977.

[3] Velleman P., Hoaglin D. The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis. Duxbury Press, 1981. P. 354.

[4] Прикладная статистика: Основы моделирования и первичная обработка данных / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков [и др.]. Москва: Финансы и статистика, 1983.

Неймана-Пирсона, для сравнения качества классификации разных методов необходимо обеспечить стабильно низкий уровень вероятности ложной тревоги. Такое требование автоматически позволяет сравнивать лишь оценки критерия мощности для того, чтобы интерпретировать полученные результаты, однако на практике по разным причинам не удается обеспечить одинаково стабильный уровень ложной тревоги одновременно для всех исследуемых методов. В задачах информационного поиска (Information Retrieval) одним из критериев оценки качества классификации является F-мера, комбинирующая оценку полноты (recall) и точности (precision). Но в литературе до сих пор нет исследований статистических методов классификации, оценка качества классификации которых являлась бы комбинацией критерия мощности и вероятности ложной тревоги. В данной работе вводится новая мера качества классификации H-мера, с помощью которой проводится сравнение улучшенных и новых предложенных методов для одномерных, двумерных и многомерных данных.

Практическое применение новых методов разведочного анализа данных основывается на их эффективной реализации, поэтому разработка программно-алгоритмического обеспечения предложенных методов является весьма актуальной задачей.

Целью данной работы является разработка комплекса методов, алгоритмов и программ реализации новых инструментов визуализации одномерных, двумерных и многомерных данных и отбраковка их аномальных значений на основе высокоэффективных робастных оценок параметров положения, масштаба и корреляции.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Определить и обосновать критерии качества методов классификации, по которым предлагается проводить сравнение эффективности модификаций боксплотов для отбраковки аномальных значений в данных.
2. Исследовать и сравнить качество различных методов отбраковки аномальных значений данных.
3. Построить и исследовать двумерный боксплот на основе быстрых высокоэффективных робастных оценок масштаба и провести сравнение двумерных боксплотов.
4. Применить предложенный двумерный боксплот для обнаружения аномалий в многомерных данных.
5. Разработать программный комплекс, реализующий:
   - алгоритмы вычисления критериев качества отбраковки аномалий в данных;
   - алгоритмы визуализации на основе предложенных новых методов отбраковки аномальных данных;
   - алгоритмы отбраковки аномалий в данных.
6. Применить предложенные методы к отбраковке реальных данных.

Основные положения, выносимые на защиту:

1. Оценка качества отбраковки аномалий в данных в виде H-меры, ее свойства и интерпретация в терминах критериев мощности и вероятности ложной тревоги.
2. Выявление аномалий в данных робастными версиями одномерных боксплотов, основанных на высокоэффективных оценках параметра масштаба по H-мере.
3. Двумерный FQ_n-боксплот: алгоритм построения; подбор параметров с помощью H-меры. Сравнительный анализ воспроизведения эллиптической формы FQ_n-боксплотом и другими двумерными боксплотами. Выявление аномалий на плоскости применением FQ_n-боксплота и сравнение с остальными двумерными методами. Визуализация данных с использованием FQ_n-боксплота.
4. Многомерные методы выявления аномалий в данных и их сравнение по H-мере. Использование двумерных боксплотов для выявления аномалий в многомерных данных.
5. Разработка алгоритмов для обнаружения точек разладки временных рядов.
6. Разработка прикладных программных модулей, реализующих алгоритмы методов классификации и отбраковки аномалий в данных, а также обеспечивающих их визуализацию.

Научная новизна:

1. Предложена новая оценка качества методов отбраковки аномалий в данных на основе H-меры, зависящей от значений мощности метода и вероятности ложной тревоги. Аналитически показано, что высокие значения H-меры гарантируют достаточно высокие значения мощности и низкие значения вероятности ложной тревоги рассматриваемого метода отбраковки.
2. Разработаны и исследованы новые модификации классических одномерных боксплотов Тьюки, основанные на робастных высокоэффективных оценках параметра масштаба.
3. Впервые исследован тип засорения «всплеск» и предложен метод спейсингов для его отбраковки.
4. Предложен новый двумерный боксплот, ориентированный на отбраковку аномалий и визуализацию двумерных данных, распределенных по нормальному закону.
5. Разработаны методы отбраковки аномалий в многомерных данных, основанные на предложенном двумерном FQ_n-боксплоте.

Практическая значимость. Разработан и реализован ряд алгоритмов для выявления аномалий, их отбраковки и визуализации данных для одномерного, двумерного и многомерного случая. Предложены оптимальные коэффициенты внешних границ робастных боксплотов в общем случае. Продемонстрировано применение одномерных боксплотов для решения задачи об определении точки разладки временного ряда для реальных данных.

Методы исследования. В работе использованы методы теории вероятностей, математической статистики, методы оптимизации и статистические методы, а также технологии параллельных и распределенных вычислений. Моделирование данных методом Монте-Карло позволило экспериментально проверить теоретически обоснованные алгоритмы. Для реализации алгоритмов использована статистическая среда программирования *R*.

Достоверность изложенных в работе результатов обеспечивается корректностью постановок рассматриваемых задач и адекватностью алгоритмов и моделирующих программ рассматриваемым математическим моделям.

Апробация работы. Основные результаты работы докладывались и обсуждались на следующих конференциях:

- XII международная научно-практическая конференция "Фундаментальные и прикладные исследования, разработка и применение высоких технологий в промышленности".
- Симпозиум НЕПАРАМЕТРИКА – XIV, Томск, 1 – 3 июля 2012
- Международная конференция по робастной статистике (International Conference on Robust Statistics – ICORS ‘11).
- 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Международная конференция по робастной статистике (International Conference on Robust Statistics – ICORS ‘13).
- 10th International Conference on Computer Data Analysis & Modeling 2013 (CDAM ‘ 13).

Публикации. Основные результаты по теме диссертации изложены в 6 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК, 4 — в тезисах и трудах российских и международных конференций.

Объем и структура работы. Диссертация состоит из введения, шести глав, заключения и приложения. Полный объем диссертации **164** страницы текста с **60** рисунками и 22 таблицами. Список литературы содержит **88** наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена методам разведочного анализа данных (РАД) для их наглядного представления, а также возможностям обнаружения аномальных наблюдений с помощью предложенных инструментов. РАД включает в себя множество визуальных инструментов, но данная работа изучает только боксплоты. Боксплот

– это графическое представление 5-числовой характеристики рассматриваемой выборки. Рассматриваемое графическое представление (боксплот) определяется медианой, нижним и верхним квартилями, а также двумя экстремумами:

$$x_L = \max\{x_{(1)}, LQ - \frac{3}{2}IQR\}, \qquad x_U = \min\{x_{(n)}, UQ + \frac{3}{2}IQR\}, \tag{1}$$

где LQ и UQ соответствуют нижнему и верхнему квартилям, а $IQR = UQ - LQ$ — интерквартильный размах. Все наблюдения, выходящие за пределы экстремумов, являются выбросами (аномальными наблюдениями). Графически (см. Рис. 1) внутренняя часть боксплота представлена как коробчатая конструкция с границами, равными нижнему и верхнему квартилям, содержащая 50% центральных значений выборки, то есть ближайших к выборочной медиане. Ширину коробчатой конструкции принято определять как равную квадратному корню размера выборки. Медиана обозначается линией внутри коробки и делит интерквантильную область на две части. Прямые, исходящие из противоположных сторон коробки, обозначают «хвосты» распределения выборки, их длина определяется экстремумами.

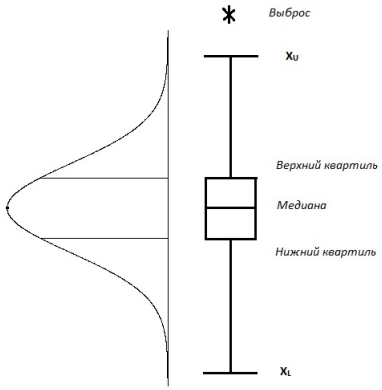


Рис. 1: Построение боксплота

Было предложено множество модификаций боксплота, в основном для представления большего количества характеристик выборки, чем в классическом варианте. Однако в литературе встречается очень мало работ, посвященных улучшению качества робастности одномерного боксплота.

Визуализация информации на плоскости потребовала обобщения понятия параметра масштаба для построения двумерного боксплота. Двумерный боксплот определяется параметром сдвига, внутренней границей (hinge) и внешней границей (fence). Для двумерного нормального закона распределения желательно иметь представление о степени корреляции между исследуемыми параметрами.

Вторая часть первой главы рассматривает существующие в настоящее время понятия об аномалиях в данных и о выбросах, а также методы их отбраковки в одномерном и многомерном случаях.

Вторая глава посвящена критериям качества различных методов отбраковки аномалий в данных. По сути, задача отбраковки аномалий в данных сводится к задаче бинарной классификаю. Рассматриваемые критерии качества классификации берут свое начало в теории обработки сигналов и распознавания образов. По результатам работы алгоритма бинарной классификации, применяемого к тестовой выборке, собирается статистика правильного отнесения наблюдений к соответствующему классу при помощи матрицы сопряженности. Процедура классификации в статистических задачах заключается в проверке нулевой гипотезы H_0 против альтернативной H_1. В результате работы классификационного метода, тестовое наблюдение x заносится в матрицу сопряженности в одну из следующих четырех возможных групп: 1) $x \in H_0$ — правильная классификация; 2) $x \in H_1$ — правильная классификация; 3) $x \in H_0$ — неправильная классификация; 4) $x \in H_1$ — неправильная классификация. В нашей работе проверяется нулевая гипотеза H_0, которая гласит, что наблюдение принадлежит «регулярным» данным, а ее альтернатива состоит в проверке принадлежности наблюдения к «выбросам». Критерий мощности метода определяется вероятностью правильной классификации наблюдения как выброса. Вероятность ложной тревоги относится к случаю, когда «регулярное» наблюдение ошибочно классифицируется как «выброс». С помощью матрицы сопряженности вычисляются значения критерия мощности и вероятности ложной тревоги. В методологии используется кривая ошибок (Receiver Operating Characteristic; ROC-кривая) для полученных значений критерия мощности и вероятности ложной тревоги. Чаще всего, кривая ошибок строится как зависимость критерия мощности от дополнения вероятности ложной тревоги. Оценкой качества классификации с помощью кривой ошибок служит значение площади под этой кривой: чем больше занимаемая площадь под кривой, тем лучше работает алгоритм классификации. В нашей работе вводится оценка качества классификации в виде H-меры как гармонического среднего между критерием мощности P_D и вероятностью ложной тревоги P_F следующим образом:

$$H(\mathsf{P}_D,1-\mathsf{P}_F)=\frac{2\mathsf{P}_D(1-\mathsf{P}_F)}{\mathsf{P}_D+(1-\mathsf{P}_F)} \tag{2}$$

На Рис. 2 представлена геометрически связь между кривой ошибок и значениями H-меры. Максимально возможное значение H-меры достигается в случае, когда кривая ошибок касается изолинии для фиксированного значения H^*.

Для каждой точки кривой ошибок можно с легкостью вычислить оценку H-меры с помощью уравнения (2). Максимальные значения H-меры достигаются для значений критерия мощности и дополненной вероятности ложной тревоги, одновременно стремящихся к 1. Такое свойство существенно важно для практики оценки классификации методов. Построение кривой ошибок требует большого объема вычислений, в то время как применение H-меры ограничивает такие усилия

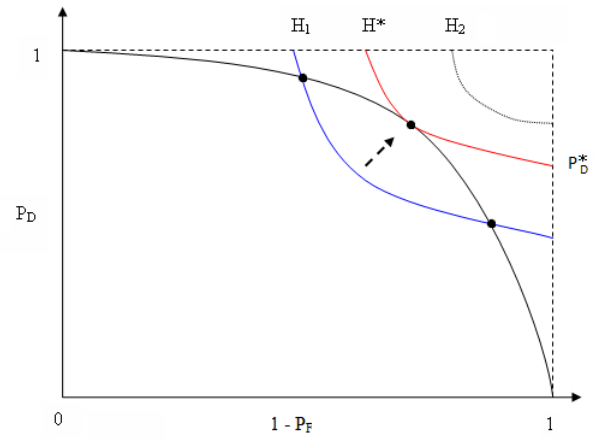


Рис. 2: Интерпретация H-меры и ее связь с графиком ROC-кривой.

до минимально возможного объема вычислений. Сравнение классификационных методов можно провести по оптимизированным значениям H-меры. По значению H-меры возможно определить нижнюю границу критерия мощности метода и верхнюю границу вероятности ложной тревоги.

Теорема 1. *Для фиксированного значения H-меры H значения критерия мощности и вероятности ложной тревоги определяются следующими соотношениями:*

$$\mathsf{P}_D > \mathsf{P}_D^{min} = \frac{H}{2-H}, \quad \mathsf{P}_F < \mathsf{P}_F^{max} = 2\frac{1-H}{2-H}$$

С помощью последней теоремы можем вычислить минимальные значения критерия мощности и максимальные значения ложной тревоги при заданном значении H-меры. В табл. 1 представлены типичные для статистического сравнительного анализа значения.

H	0.5	0.6	0.7	0.8	0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.999
P_D^{min}	0.33	0.43	0.54	0.67	0.82	0.83	0.85	0.87	0.87	0.90	0.92	0.94	0.96	0.98	0.998
P_F^{max}	0.67	0.57	0.46	0.33	0.18	0.17	0.15	0.13	0.11	0.095	0.08	0.058	0.04	0.02	0.002

Таблица 1: Минимальные значения критерия
мощности и максимальные значения ложной тревоги при фиксированных значениях H-меры.

В **третьей главе** рассматривается отбраковка аномалий в одномерных данных как классическими, так и предложенными новыми методами.

Аналитически и экспериментально исследовалась зависимость качества отбраковки по H-мере от параметра сдвига и масштаба для различных параметров в модели распределения Тьюки (gross error model) типа «сдвиг»

$$F(x) = (1-\varepsilon)\Phi + \varepsilon\Phi(x-\mu) \tag{3}$$

и «масштаб»

$$F(x) = (1-\varepsilon)\Phi + \varepsilon\Phi(x/k), \tag{4}$$

где μ — параметр сдвига, k — параметр масштаба, $0 \leq \varepsilon < 1$ — вероятность появления выбросов (аномалий), а $\Phi(x)$ — функция Лапласа.

На практике, при исследовании различных процессов информация о параметрах заложенного в модели распределения процесса отсутствует. В этом случае параметры закона распределения оцениваются статистическими методами. В нашей работе особое внимание уделяется робастным высокоэффективным оценкам положения и масштаба, таким как выборочная медиана, $MAD = med|x - medx|$ и FQ_n[5].

Сравнительное исследование качества отбраковки по классическому методу отбраковки Граббса[6] и его робастной модификации с заменой выборочного среднего на выборочную медиану и среднеквадратичного отклонения на робастную оценку масштаба FQ_n показывает, что метод Граббса сильно проигрывает его робастной модификации как по H-мере (см. табл. 2), так и по мощности и вероятности ложной тревоги.

Параметры масштаба k и сдвига μ	Уровень значимости (α)	Засорение типа «масштаб»		Засорение типа «сдвиг»	
		Робастный метод	Метод Граббса	Робастный метод	Метод Граббса
2	0.9	0.468	0.394	0.728	0.444
2	0.95	0.298	0.236	0.617	0.288
2	0.99	0.104	0.072	0.394	0.142
5	0.9	0.997	0.784	0.999	0.799
5	0.95	0.994	0.674	0.998	0.625
5	0.99	0.986	0.465	0.965	0.283
10	0.9	1.0	0.880	1.0	0.766
10	0.95	0.999	0.822	1.0	0.621
10	0.99	0.999	0.666	1.0	0.325

Таблица 2: Сравнение качества классификации по H-мере для засорения типа «сдвиг» и «масштаб».

Боксплот является не только инструментом визуализации данных, с его помощью также можно выделить выбросы, лежащие за пределами его экстремумов. Оценки параметра положения — выборочная медиана — и внутренней границы — интерквартильный размах (IQR) — являются обоснованными оценками, не подлежащими какой-либо модификации для повышения качества отбраковки аномалий в данных.

В нашей работе внешние границы боксплота задаются следующим образом:

$$x_L = \max\{x_{(1)}, LQ - k_{S_n} S_n\}, \qquad x_U = \min\{x_{(n)}, UQ + k_{S_n} S_n\}, \tag{5}$$

[5]Shevlyakov G. L., Smirnov P. O. On Approximation of the $Q_n - estimate$ of Scale by Fast $M - estimates$ // Int. Conf. on Robust Statistics. Parma, Italy: 2010.

[6]Grubbs F. E., Beck G. Extension of sample sizes and percentage points for significance tests of outlying observations // Technometrics. 1972. Vol. 14, no. 4. P. 847–854.

где k_{S_n} — константа, а S_n — робастная высокоэффективная оценка масштаба. Значения k_{S_n} определены с помощью оптимизации качества отбраковки робастных модификаций боксплотов по H-мере. Рекомендуется использовать оптимальные по H-мере значения $k_{MAD} = 1.44$ для MAD-боксплота и $k_{FQ_n} = 0.97$ для одномерного FQ_n-боксплота. В результате сравнительного анализа по H-мере наблюдается улучшение качества отбраковки с помощью предложенных модификаций классического боксплота.

Данные, подчиняющиеся асимметрическому закону распределения, создают определенные трудности для выявления и отбраковки аномалий. Трудность отбраковки аномалий в этом случае связана с асимметричностью распределения данных на хвостах. Как правило, на стороне более тяжелого хвоста статистические методы отбраковки аномалий не в состоянии отличить наблюдения, принадлежащие «регулярной» группе данных, от выбросов, вследствие чего сильно растет вероятность ложной тревоги. В литературе встречаются работы, нацелены на разработку боксплотов, учитывающих асимметрию в данных, таких как SIQR-боксплот[7] и настраиваемый боксплот[8] (adjusted boxplot). Мы провели сравнительное исследование классического боксплота, SIQR-боксплота и настраиваемого боксплота по качеству отбраковки аномалий по H-мере для данных, распределенных по асимметрическому закону. В результате проведения 1000 испытаний на выборках объемом 1000 с засорением типа «сдвиг» при $\varepsilon = 0.05$ для пяти групп асимметричных распределений самые высокие показатели H-меры принадлежат SIQR-боксплоту.

В **четвертой главе** приведено описание простого в построении двумерного боксплота, определение коэффициента корреляции которого является устойчивым к выбросам. На основе приведенного двумерного боксплота предлагается новый алгоритм построения двумерного боксплота, ориентированного на модели двумерного закона распределения. В предлагаемом алгоритме двумерного FQ_n-боксплота применяется высокоэффективная робастная оценка параметра масштаба FQ_n, значение которой влияет на определение внешней границы двумерного боксплота, а также с ее помощью вычисляется робастный коэффициент корреляции.

Было проведено исследование параметров FQ_n-боксплота для выбора параметра положения и для выбора коэффициента внешней границы. В результате в качестве параметра положения выбрана пространственная медиана. Для подбора коэффициента внешней границы проводили численное моделирование выборок объемами 50, 100 и 1000.

[7] Kimber A. C. Exploratory data analysis for possibly censored data from skewed distributions // Applied Statistics. 1990. Vol. 11, no. 1. P. 21–30.

[8] Hubert M., Vandervieren E. An adjusted boxplot for skewed distributions // Computational Statistics & Data Analysis. 2008. Vol. 52, no. 12. P. 5186–5201.

По аналогии с параметрическими и непараметрическими методами можем выделить боксплоты, ориентированные на данные, и модельно-ориентированные боксплоты. В нашем случае, предложенный FQ_n-боксплот наряду с relplot-ом и quelplot-ом[9] являются модельно-ориентированными боксплотами, предназначенными для обнаружения двумерной нормально распределенной структуры данных. В качестве двумерного боксплота, ориентированного на данные, можем привести bagplot[10]. Нами было проведено исследование способности воспроизведения эллиптической формы для засоренной двумерной выборки, закон распределения которой подчиняется нормальному закону, для четырех двумерных боксплотов: FQ_n-боксплот, bagplot, quelplot и relplot. Способность воспроизведения эллиптической формы оценивается с помощью вычисления среднеквадратичной относительной ошибки путем интегрирования квадрата относительной ошибки по всем возможным направлениям $(0;2\pi)$

$$I = \frac{1}{2\pi}\int_0^{2\pi}\left[\frac{R(\phi)-R_e(\phi)}{R_e(\phi)}\right]^2 d\phi$$

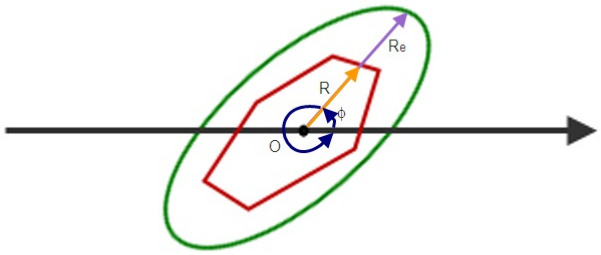


Рис. 3: Вычисления меры отклонения внешней области боксплота от эллиптической формы.

В результате исследования по воспроизведению эллиптической формы установлено, что в большинстве случаев предложенный нами боксплот обладает более низкими показателями среднеквадратичной относительной ошибки по сравнению с bagplot-ом, quelplot-ом и relplot-ом.

Нами был проведен сравнительный анализ качества отбраковки выбросов по H-мере для FQ_n-боксплота и для bagplot-а, в случае засорения типа «сдвиг» и «масштаб». Численное моделирование состояло в воспроизведении выборок объемом 50, 100 и 1000 по 10000 повторений. Результаты качества отбраковки по H-мере для предложенного FQ_n-боксплота достигают значения H=0.84 в случае малых засорений, в то время как для bagplot-а такие показатели находятся на уровне H=0.72.

Необходимо еще раз подчеркнуть, что боксплоты являются средствами визуализации данных. В разведочном анализе данных боксплоты в первую очередь дают

[9]Goldberg K. M., Iglewicz B. Bivariate extensions of the boxplot // Technometrics. 1992. Vol. 34. P. 307–320.

[10]Rousseeuw P. J., Ruts I., Tukey J. W. The bagplot, a bivariate boxplot // The American Statistician. 1999. Vol. 53. P. 382–387.

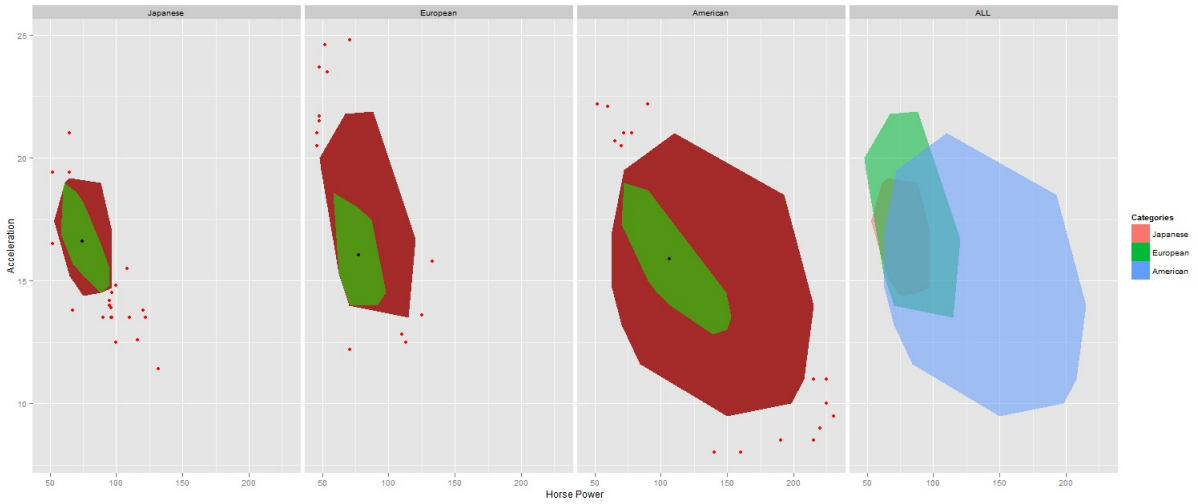


Рис. 4: Визуализация двумерных FQ_n-боксплотов.
Представление зависимости значений разгона от мощности автомобилей,
сгруппированных по разным представителям автопромышленности на разных континентах:
Северная Америка(Buick Dodge, Chrysler, GM, Ford, Oldsmobile, Chevrolet), Европа (Citroen, VW,
Peugot, Audi, Saab, BMW, Fiat, Volvo, Renault, Mercedes) и Азия (Datsun, Toyota, Mazda, Subaru, Honda)

представление о статистических характеристиках данных. С помощью одномерного боксплота возможны исследование и сравнение нескольких выборок. По аналогии с одномерными боксплотами, мы предлагаем следующий вариант визуализации FQ$_n$-боксплотов (см. Рис. 4). Предложенный нами вариант упорядочивает двумерные боксплоты горизонтально по пространственным медианам, сохраняя общий масштаб для оси ординат. В последнем разделе отображаются внешние выпуклые оболочки всех боксплотов, что позволяет получить представление об их взаимном расположении на плоскости. Для визуализации многомерных данных высоких размерностей рекомендуется построить графическую матрицу, составленную из двумерных боксплотов по всем возможным плоскостям, при этом по главной диагонали будут располагаться одномерные боксплоты, а вне главной диагонали — двумерные боксплоты.

Пятая глава посвящена отбраковке аномалий в многомерных данных, размерность которых от трех и более. На практике широко известен подход к отбраковке в этом случае, основанный на мере отдаленности многомерных наблюдений от выборочного параметра положения с помощью расстояния Махаланобиса. Однако, с увеличением размерности пространства исследование данных при помощи расстояния Махаланобиса для выявления аномалий усложняется. Переход в пространство большей размерности сопровождается нарушением структуры связей между случайными величинами, что не позволяет судить о принадлежности наблюдения только по оценке расстояния. Для устранения проблем, возникающих при высоких значениях размерности, предложены различные методы, группирующиеся по следующим категориям: методы, основанные на оценках расстояния,

методы, использующие главные компоненты для снижения размерности, методы кластеризации и пространственные методы.

Согласно методам, основанным на анализе расстояний, наблюдения, расстояние которых от остальной части данных больше, чем заранее определенное пороговое значение, обозначаются аномальными (выбросами). Мерой отдаленности наблюдений служит расстояние Махаланобиса или его робастные модификации. В случае исследования данных не очень высоких размерностей эти методы дают хорошие результаты.

Применение метода главных компонент помогает снизить размерность пространства за счет подбора более информативных признаков. Главные компоненты представляют собой направления, которые максимизируют дисперсию вдоль каждой компоненты при условии ортогональности. Наличие выбросов в данных способствует увеличению дисперсии в своем направлении, остальные направления являются малоинформативными. В нашем исследовании был выбран для исследования эффективный метод **PCOUT**, предложенный Филцмозером[11] (P. Filzmoser). Алгоритм этого метода состоит из двух частей: первый шаг специализируется на выбросах типа «сдвиг», второй шаг эффективен для выбросов типа «масштаб». Для снижения размерности применяется метод главных компонент и в первой, и во второй части алгоритма.

Методы кластеризации относятся к разделу статистической обработки данных для выявления структур наблюдений в пространстве любой размерности. Применение таких методов для выявления аномалий основывается на предположении, что аномальное наблюдение принадлежит кластеру с единственным представителем. Для нашего исследования мы выбрали метод кластеризации DBSCAN (density reachability and connectivity clustering). В отличие от других методов кластеризации, выбранный нами метод не ограничен сферической формой кластера. Алгоритм DBSCAN основан на понятии наличия плотности определенного значения внутри кластеров. Наблюдения, выходящие за пределы окрестности, в которой не нарушается установленное значение плотности в кластере, создают новый кластер.

Предложенные нами методы выявления аномалий в многомерных данных основываются на исследовании двумерных проекций по главным осям координат исследуемой выборки. Если обрабатываемые данные имеют размерность p, то число полученных проекций равно числу сочетаний из p по 2: C_p^2. Для каждой проекции отбраковываем выбросы с помощью двумерного FQ_n-боксплота. По сути данная процедура отбраковки выбросов в многомерном пространстве заключается в построении графической $p \times p$ матрицы, элементы которой являются двумерными FQ_n-боксплотами.

[11] Filzmoser P., Maronna R., Werner M. Outlier identification in high dimensions // Computational Statistics & Data Analysis. 2008. Vol. 52, no. 3. P. 1694–1711.

Модификация вышеприведенного алгоритма выявления выбросов в многомерных данных состоит в добавлении уточняющего шага (второй итерации). Во второй итерации предполагаем, что степень засорения — ε — заранее известна. Если количество выбросов, полученных в результате алгоритма, меньше предполагаемой степени засорения $\varepsilon \cdot N$, где N – размер выборки, тогда результат остается без изменений. В противном случае выбросы упорядочиваются по расстоянию Махаланобиса в убывающем порядке и выбираются первые $\varepsilon \cdot N$ элементов.

Экспериментально сравнивались пять методов выявления аномалий в многомерных данных: два предложенных нами метода, **PCOUT**, DBSCAN и настраиваемый метод квантилей (adjusted quantile method). Для оценки времени выполнения и значений H-меры использовались сгенерированные данные размерностей 3, 4 и 5 для двух типов засорения: «сдвиг» и «масштаб». Был проведен эксперимент для многомерных данных размерностью 10, но полученные результаты еще раз подтвердили ухудшение качества отбраковки аномалий при очень больших размерностях.

Самым быстрым в вычислительном смысле алгоритмом является **PCOUT**. Метод адаптивного квантиля или кластерный метод DBSCAN с высокой точностью справляются с засорением типа «сдвиг». Предложенный нами итерационный метод дает лучшие результаты и может быть рекомендован для использования в случае засорения типа «масштаб» при размерностях 3–5.

В **шестой главе** рассматривается метод решения реальной задачи с помощью одномерных боксплотов.

В нефтедобывающей промышленности для эффективной эксплуатации месторождений необходимо обеспечить плавный бесперебойный режим работы откачивания нефти. Одним из показателей этого режима является скорость потока жидкости, откачиваемой из скважины. Скорость откачиваемой жидкости может варьироваться в зависимости от изменения физических свойств жидкости (от более густого до газообразного), а также в результате таких физических явлений, как постоянно появляющиеся турбулентные потоки.

По скорости откачиваемой жидкости в нефтедобывающей промышленности выделяются следующие основные режимы: 1) высокоамплитудный колебательный режим (High Amplitude Oscillation — HAO), 2) низкоамплитудный колебательный режим (Low Amplitude Oscillation — LAO) и 3) нормальный режим без колебаний.

Для обеспечения эффективной работы нефтедобывающего комплекса требуется определить за минимальное время переход из одного режима в другой. Результаты измерений скорости откачиваемой жидкости представляют собой временной ряд. На языке теории временных рядов здесь требуется решить задачу определения точки разладки.

Временной ряд скорости потока откачиваемой жидкости подвергается множеству воздействий, поэтому необходимо предварительная обработка. Быстрый

просмотр реализации временного ряда выявляет нестационарность исследуемого процесса, а также выбросы в виде отдельных элементов, сильно отличающихся от остальных близких по времени результатов наблюдений. С помощью сглаживающего медианного фильтра и последующего центрирования результатов наблюдений устранены одиночные выбросы и временной ряд преобразован к приближению стационарной модели.

Описанная выше задача определения точки разладки по результатам измерений скорости откачиваемой жидкости до сих пор не имеет эффективного решения и остается предметом для исследования. Было предложено несколько методов классификации режимов, но для их правильной классификации необходимо собрать достаточно большую выборку, что приводит к задержке определения точки разладки.

В нашей работе для определения точки разладки предлагается решение, основывающееся на алгоритмах отбраковки аномалий. Рассматривается фиксированный по размеру набор ближайших к исследуемой точке результатов наблюдений. После предварительной обработки данных желательно сохранить значения параметров положения, масштаба и размаха текущей выборки. Часто применяемый на практике метод скользящей медианы не в состоянии справиться с решением этой задачи. Предложенный нами метод скользящего боксплота представляет особый интерес, так как он является не только средством отбраковки аномалий, но также дает большую информацию о характеристиках выборки.

С помощью метода скользящего боксплота возможно исследование различных правил для выявления перехода из одного режима в другой. Критерием оценки для выявления точки разладки в виде аномалии служит матрица сопряженности и задержка в ее определении. Лучшим правилом в отношении критерия матрицы сопряженности и задержки определения режима является алгоритм скользящих боксплотов с использованием размеров ящика боксплота (нижнего и верхнего квантилей) и выбросов с размером окна в 31 отсчет.

Метод скользящих боксплотов рекомендуется использовать совместно с правилами классификации на основе оценок спектров данных.

В **заключении** приведены следующие основные результаты работы:

1. Разработан комплекс методов, алгоритмов и программ разведочного анализа данных для визуализации и выявления аномалий. Эта цель была достигнута модификацией классических и созданием новых инструментов визуализации одномерных, двумерных и многомерных данных и отбраковки их аномальных значений на основе высокоэффективных робастных оценок параметров положения, масштаба и корреляции.
2. Предложена оценка качества метода отбраковки аномалий в данных в виде гармонического среднего, зависящая от критерия мощности и вероятности ложной тревоги.

3. На основе H-меры впервые исследовалось качество отбраковки аномалий в данных для одномерных, двумерных и многомерных методов выявления аномалий.
4. Математическое моделирование показало преимущество применения робастной высокоэффективной FQ_n оценки параметра масштаба для робастных модификаций традиционных методов выявления аномалий в данных и построения новых алгоритмов. На основе FQ_n оценки параметра масштаба предложена робастная модификация одномерного боксплота, введен одномерный метод «λ сигм», вычисляется робастный коэффициент корреляции, а также на ее основе задается алгоритм построения двумерного FQ_n-боксплота.
5. Для предложенного одномерного метода «λ сигм» установлены пороговые значения λ, для которых достигаются максимальные оценки по H-мере.
6. Предложены робастные модификации одномерного боксплота Тьюки на основе робастных высокоэффективных оценок параметра масштаба: MAD и FQ_n.
7. Было проведено исследование качества отбраковки аномалий по H-мере асимметричных одномерных боксплотов: SIQR-боксплота, классического боксплота Тьюки и настраиваемого боксплота. Для сравнения асимметричных боксплотов смоделированы различные по виду распределения выборки (пять групп асимметричных распределений).
8. Впервые было исследовано засорение типа «всплеск» и предложен метод спейсингов для выявления такого вида аномалий. Для метода спейсингов оценка качества выявления аномальных наблюдений типа «всплеск» по H-мере гарантировано больше, чем 0.9, когда размер выборки превышает 200.
9. Предложен модельно-ориентированный двумерный FQ_n-боксплот. Проведено исследование на способность воспроизведения эллиптической формы для двумерного нормального закона распределения. Исследование качества отбраковки аномалий в данных и последующее сравнение с bagplot-ом подтверждает преимущество предложенного двумерного FQ_n-боксплота.
10. Предложены два алгоритма выявления выбросов для многомерных данных: метод проекций и его итеративная модификация (итеративный метод). Проведен сравнительный анализ предложенных и существующих методов. Для размерностей данных 3-5 предложенные методы дают хорошие результаты. Лучшие результаты итеративного метода по сравнению с проекционным объясняются тем, что на последующих итерациях уменьшаются ошибки проекционного метода.

СПИСОК РАБОТ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Андрэа, К.* Визуализация данных двумерными FQn-боксплотами [Текст] / К. Андреа, Г. М. Лаврентьева, П. О. Смирнов, Г. Л. Шевляков // Высокие технологии, фундаментальные исследования, экономика. — Т. 1. — Санкт-Петербург, Россия : Изд-во Политехн. ун-та, 2011. — С. 59 - 66.
2. *Андрэа, К.* Двумерный боксплот на основе высокоэффективных робастных оценок масштаба и корреляции [Текст] /К. Андреа, П. О. Смирнов, Г. Л. Шевляков // Вестник Томского Государственного Университета. Управление. Вычислительная техника и информатика. — 2013. — Т. 22, № 1. — С. 25 - 31.
3. *Андрэа, К.* Обнаружение выбросов с помощью боксплотов, основанных на новых высокоэффективных робастных оценках масштаба [Текст] / К. Андреа, Г. Л. Шевляков // Научно-технические ведомости Санкт-Петербургского Государственного Политехнического Университета. Информатика. Телекоммуникации. Управление. — 2013. — Т. 5, № 181. — С. 39 - 45.
4. *Andrea, K.* Fast low-complexity bivariate boxplots based on highly efficient and robust estimates of dispersion and correlation [Text] / G. Shevlyakov, K. Andrea, G. Lavrentyeva, P. Smirnov // Book of Absracts: International Conference on Robust Statistics (ICORS 2011). — Valladolid, Spain : University of Valladolid, 2011. — P. 72.
5. *Andrea, K.* Robust versions of the Tukey boxplots with their application to detection of outliers [Text] / Georgy L. Shevlyakov, Kliton Andrea, Lakshminarayan Choudur [et al.] // IEEE International Conference on Acoustics, Speech, and Signal Processing. — Vancouver, Canada : IEEE, 2013. — P. 6506 – 6510.
6. *Andrea, K.* Detection of outliers with boxplots [Text] / K. Andrea, G. L. Shevlyakov, P. O. Smirnov // Proceedings of the 11th International Conference on Computer Data Analysis and Modeling. — Minsk, Belarus : Publishing center of BSU, 2013. — P. 141 - 144.