ВЫБОР МЕТРИЧЕСКИХ АТТРИБУТОВ РЕДКИХ АНОМАЛЬНЫХ СОБЫТИЙ КОМПЬЮТЕРНОЙ СИСТЕМЫ МЕТОДАМИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

DOI: 10.36724/2072-8735-2021-15-6-40-47

Шелухин Олег Иванович,*Московский технический университет связи и информатики,
Москва, Россия, sheluhin@mail.ru***Раковский Дмитрий Игоревич,***Московский технический университет связи и информатики,
Москва, Россия, dimitor1998@mail.ru***Manuscript received** 02 February 2021;**Accepted** 25 February 2021**Ключевые слова:** маркировка значений атрибутов, корреляционный анализ, k-means, экспериментальные данные, violin plot тренд, difference stationary series

Рассмотрен процесс маркировки многоатрибутных экспериментальных данных для последующего использования средствами интеллектуального анализа данных в задачах обнаружения и классификации редких аномальных событий компьютерных систем (КС). Процесс маркировки выполнен при помощи трех методов: ручная предобработка, статистический анализ и кластерный анализ. Среди атрибутов метрического типа ручным методом выделены две макрогруппы: "интегральные" и "импульсные" атрибуты. Показано, что комбинирование методов статистического и кластерного анализа повышает точность выявления аномальных событий КС, а также позволяет производить отбор атрибутов по их информационной значимости. На примере разделения атрибутов на макрогруппы, анализа распределения плотности значений и удаления трендовой составляющей методом интегрированных временных рядов (DS-рядов, Difference Stationary) показана целесообразность ручной предобработки данных перед проведением кластеризации. При помощи построения скрипичных диаграмм (Violin plot) для атрибута "интегральной" макрогруппы показано распределение состояний КС. Показано, что удаление трендовой составляющей методом DS-рядов, нормализация и приведение к абсолютным значениям позволяет с большей точностью маркировать аномальные выбросы, однако это не всегда допустимо. Интерпретация результатов кластеризации, выполненная для каждого нормализованного атрибута, показывает, что нормальные значения для всех атрибутов сконцентрированы около нулевых значений. Результатом маркировки экспериментальных данных являются поатрибутно маркированные данные, где каждому атрибуту в текущий момент времени присваивается одно из двух состояний: аномальное или нормальное.

Информация об авторах:**Шелухин Олег Иванович,** *Московский технический университет связи и информатики, д.т.н., профессор, заведующий кафедрой "Информационная безопасность", Москва, Россия***Раковский Дмитрий Игоревич,** *Московский технический университет связи и информатики, магистр кафедры "Информационная безопасность", Москва, Россия***Для цитирования:***Шелухин О.И., Раковский Д.И. Выбор метрических атрибутов редких аномальных событий компьютерной системы методами интеллектуального анализа данных // T-Comm: Телекоммуникации и транспорт. 2021. Том 15. №6. С. 40-47.***For citation:***Sheluhin O.I., Rakovsky D.I. (2021) Selection of metric and categorical attributes of rare anomalous events in a computer system using data mining methods. T-Comm, vol. 15, no. 6, pp. 40-47. (in Russian)*

Постановка задачи

Каждая компьютерная система генерирует большое количество данных, которые необходимо анализировать на предмет наличия редких аномальных событий. Состояния параметров КС считываются специальными системными датчиками и представляют собой цифровой сигнал, характеризующий параметр (атрибут) системы. Совокупность значений всех атрибутов в текущий момент времени будет отражать состояние системы в этот момент времени. Иногда в КС возникают аномалии – отклонение какого-либо атрибута от нормального значения, от закономерности [1]. Аномалия может заключаться как в скорости изменения значения атрибута во времени, так и выход значения атрибута за диапазон допустимых значений.

Некоторая, как правило, редкая, труднопредсказуемая и возникающая одновременно совокупность аномальных состояний КС может привести к поломке или аварии. Совокупность аномальных состояний КС, приводящих к аварии, будем называть редким событием [2, 3].

Стандартным подходом к обнаружению аномалий является классификация значений анализируемых атрибутов по двум категориям: «нормальные» и «аномальные». Как правило, это осуществляется либо статистическими методами, либо методами интеллектуального анализа данных.

Для корректного обнаружения редких аномальных состояний в КС методами интеллектуального анализа данных необходимо (разметить) обучающую выборку, используя исторические данные, соответствующие поведению анализируемой КС «в прошлом» [4]. В работе рассмотрен процесс анализа и последующей маркировки экспериментальных данных, снятых с сетевой инфраструктуры, состоящей из 6 хостов, образующих кластер под управлением Rancher (рис. 1).

Рис. 1. Схема исследуемой сетевой инфраструктуры

Данные сняты с локальной КС в период с 2019-09-24 05:35:06 по 2019-09-26 23:39:00 с шагом 1 секунда и содержат в себе 57 атрибутов метрического типа и 10 атрибутов категориального типов [5].

Набор экспериментальных данных сведен в csv-таблицу, в которой каждая строка (запись) отражает состояние КС в соответствующий момент времени.

Целью работы является отбор и последующая бинарная маркировка экспериментальных данных КС метрического и категориального типов методами интеллектуального анализа данных.

Анализ значимости атрибутов метрического типа

Перечень анализируемых атрибутов КС метрического типа представлен в таблице 1, в которой каждому значимому атрибуту метрического типа присвоен соответствующий идентификатор. Атрибуты, имеющие постоянное значение на протяжении всего эксперимента (выделены курсивом), являются незначимыми и исключены из последующего анализа. В результате исключения незначимых атрибутов, общее количество атрибутов метрического типа может быть сокращено с 57 до 51.

Таблица 1

Перечень анализируемых атрибутов КС метрического типа

№	Тип атрибута	Идентификатор атрибута
1	Данные по использованию центрального процессора	idle ⁽¹⁾ , iowait ⁽²⁾ , irq, nice ⁽³⁾ , softirq ⁽⁴⁾ , steal, system ⁽⁵⁾ , total ⁽⁶⁾ , user ⁽⁷⁾
2	Данные по использованию памяти	actualfree ⁽⁸⁾ , actualusedbytes ⁽⁹⁾ , actualusedpct ⁽¹⁰⁾ , free ⁽¹¹⁾ , swapfree ⁽¹²⁾ , swaptotal, swapusedbytes ⁽¹³⁾ , swapusedpct ⁽¹⁴⁾ , total
3	Данные по использованию диска	await ⁽¹⁵⁾ , busy ⁽¹⁶⁾ , ioreadawait ⁽¹⁷⁾ , ioreadmergespersec ⁽¹⁸⁾ , ioreadrequestsperssec ⁽¹⁹⁾ , iostatqueueavgsize ⁽²⁰⁾ , iostatrequestavgsize ⁽²¹⁾ , iowriteawait ⁽²²⁾ , iowritemergespersec ⁽²³⁾ , iowriterequestsperssec ⁽²⁴⁾ , readbytes ⁽²⁵⁾ , readcount ⁽²⁶⁾ , readtime ⁽²⁷⁾ , writebytes ⁽²⁸⁾ , writecount ⁽²⁹⁾ , writetime ⁽³⁰⁾
4	Данные по средней загрузке центрального процессора	oneminute ⁽³¹⁾ , fiveminutes ⁽³²⁾ , fifteenminutes ⁽³³⁾
5	Данные по использованию сети	inbytes ⁽³⁴⁾ , indropped, inerrors, inpackets ⁽³⁵⁾ , outbytes ⁽³⁶⁾ , outdropped ⁽³⁷⁾ , outerrors, outpackets ⁽³⁸⁾
6	Данные по DNS запросам	networkbytes ⁽³⁹⁾ , answerscount ⁽⁴⁰⁾
7	Данные по HTTP запросам	networkbytes ⁽⁴¹⁾ , requestbytes ⁽⁴²⁾ , responsebytes ⁽⁴³⁾
8	Данные PING запроса для определения задержки сигнала в сети	avg ⁽⁴⁴⁾ , max ⁽⁴⁵⁾ , min ⁽⁴⁶⁾
9	Данные CURL запроса для определения времени реакции сервера	timeconnect ⁽⁴⁷⁾ , timenamelookup ⁽⁴⁸⁾ , timepretransfer ⁽⁴⁹⁾ , timestarttransfer ⁽⁵⁰⁾ , timetotal ⁽⁵¹⁾

Каждый метрический атрибут можно условно разделить на три макрогруппы на основании анализа их временных диаграмм:

«Интегральные» атрибуты, для которых характерен ярко выраженный тренд (например, атрибут, характеризующий общее количество успешно прочитанных байт с накопителя – read bytes⁽²⁵⁾).

Последующая маркировка «интегральных атрибутов» средствами интеллектуального анализа данных может потребовать удаления трендовой составляющей из численных значений атрибутов [6].

«Импульсные» атрибуты, значения которых характеризуются выбросами, значительно отклоняющимися от нормального поведения (например, атрибут, характеризующий количество запросов на чтение, выданных устройству в секунду – $\text{ioreadrequestspersec}^{(19)}$).

«Иные» атрибуты, к которым относятся все атрибуты, не вошедшие в указанные выше категории.

Разделенные по макрогруппам атрибуты приведены в таблице 2.

Таблица 2

Метрические атрибуты, сгруппированные в соответствии с обозначенными критериями

№ типа	Тип атрибута	Идентификатор атрибута
Макрогруппа 1 – интегральные атрибуты 16 шт. (31%)		
2	Данные по использованию памяти	$\text{actualfree}^{(8)}, \text{actualusedbytes}^{(9)}, \text{actualusedpct}^{(10)}, \text{swapfree}^{(12)}, \text{swapusedbytes}^{(13)}, \text{swapusedpct}^{(14)}$.
3	Данные по использованию диска	$\text{readbytes}^{(25)}, \text{readcount}^{(26)}, \text{readtime}^{(27)}, \text{writebytes}^{(28)}, \text{writecount}^{(29)}, \text{writetime}^{(30)}$.
5	Данные по использованию сети	$\text{inbytes}^{(34)}, \text{inpackets}^{(35)}, \text{outbytes}^{(36)}, \text{outpackets}^{(38)}$.
Макрогруппа 2 – импульсные атрибуты – 21 шт. (41%)		
3	Данные по использованию диска	$\text{await}^{(15)}, \text{ioreadawait}^{(17)}, \text{ioreadmergespersec}^{(18)}, \text{ioreadrequestspersec}^{(19)}, \text{iostatqueueavgsize}^{(20)}, \text{iowriteawait}^{(22)}, \text{iowritemergespersec}^{(23)}, \text{iowriterequestpersec}^{(24)}$.
6	Данные по DNS запросам	$\text{networkbytes}^{(39)}, \text{answerscount}^{(40)}$.
7	Данные по HTTP запросам	$\text{networkbytes}^{(41)}, \text{requestbytes}^{(42)}, \text{responsebytes}^{(43)}$.
8	Данные PING запроса для определения задержки сигнала в сети	$\text{avg}^{(44)}, \text{max}^{(45)}, \text{min}^{(46)}$.
9	Данные CURL запроса для определения времени реакции сервера	$\text{timeconnect}^{(47)}, \text{timenamelookup}^{(48)}, \text{timepretransfer}^{(49)}, \text{timestarttransfer}^{(50)}, \text{timetotal}^{(51)}$.
Макрогруппа 3 – иные атрибуты – 14 шт. (27%)		
1	Данные по использованию центрального процессора	$\text{idle}^{(1)}, \text{iowait}^{(2)}, \text{nice}^{(3)}, \text{softirq}^{(4)}, \text{system}^{(5)}, \text{total}^{(6)}, \text{user}^{(7)}$.
2	Данные по использованию памяти	$\text{free}^{(11)}$.
3	Данные по использованию диска	$\text{busy}^{(16)}, \text{iostatrequestavgsize}^{(21)}$.
4	Данные по средней загрузке центрального процессора	$\text{oneminute}^{(31)}, \text{fiveminutes}^{(32)}, \text{fifteenminutes}^{(33)}$.
5	Данные по использованию сети	$\text{outdropped}^{(37)}$.

Для оценки значимости и оценки количества атрибутов воспользуемся построением корреляционной диаграммы на основе оценок коэффициентов корреляции атрибутов методами Пирсона, Кендалла и Спирмена [7] для каждого из 51 атрибутов.

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между значениями двух выборок одинаковой размерности. Пусть даны две выборки одинаковой размерности m : $S_{k_1}^m = \{x_{k_1,1}, \dots, x_{k_1,m}\}$ и $S_{k_2}^m = \{x_{k_2,1}, \dots, x_{k_2,m}\}$, где k_1 и k_2 – идентификаторы атрибутов КС, по которым снимались выборки; $x_{k_1,i}$ – i -й элемент в первой выборке; $x_{k_2,i}$ – i -й элемент во второй выборке. Тогда коэффициент корреляции между этими двумя выборками определяется как:

$$r_{k_1 k_2} = \frac{\text{cov}(k_1, k_2)}{\sqrt{s_{k_1}^2 s_{k_2}^2}}, \quad (1)$$

где $\text{cov}(k_1, k_2)$ – ковариация между выборками $S_{k_1}^m$ и $S_{k_2}^m$; $s_{k_1}^2$ – выборочная дисперсия выборки $S_{k_1}^m$; $s_{k_2}^2$ – выборочная дисперсия выборки $S_{k_2}^m$.

В отличие от (1) коэффициенты корреляции Кендалла и Спирмена характеризуют линейную ранговую зависимость между двумя выборками $S_{k_1}^m$ и $S_{k_2}^m$. Преобразуем значения выборок $S_{k_1}^m$ и $S_{k_2}^m$ в ранговые шкалы:

$$f: S_k^m \rightarrow d_k^m \quad (2)$$

Далее по тексту будем называть $d_{k_1}^m$ и $d_{k_2}^m$ ранговыми выборками.

Для вычисления коэффициента корреляции Спирмена необходимо провести преобразование выборок $S_{k_1}^m$ и $S_{k_2}^m$ в ранговые выборки в соответствии с формулой (2), после чего вычислить коэффициент корреляции:

$$\rho = 1 - \frac{6}{m(m-1)(m+1)} \sum_{i=1}^m r_{k_1,i} r_{k_2,i} \quad (3)$$

где m – размерность ранговых выборок $d_{k_1}^m$ и $d_{k_2}^m$; $r_{k_1,i}$ – элемент ранговой выборки $d_{k_1}^m$; $r_{k_2,i}$ – элемент ранговой выборки $d_{k_2}^m$.

Вычисление коэффициента корреляции Кендалла проводится с аналогичным преобразованием численных значений выборок $S_{k_1}^m$ и $S_{k_2}^m$ в ранговые шкалы по формуле (2). С этой целью ранговую выборку $d_{k_1}^m$ необходимо дополнительно преобразовать в вариационный ряд, расположив значения ранговой выборки в порядке не убывания: $f: d_{k_1}^m \rightarrow V_{k_1}^m$. Коэффициент корреляции Кендалла, вычисляемый для вариационного ряда $V_{k_1}^m$ и ранговой выборки $d_{k_2}^m$ вычисляется как:

$$\tau = 1 - \frac{4}{m(m-1)} R \quad (4)$$

где m – размерность вариационного ряда $V_{k_1}^m$ и ранговой выборки $d_{k_2}^m$, а R – количество инверсий между ранговой

выборкой $d_{k_2}^m$ и вариационным рядом $V_{k_1}^m$, вычисляемое из соотношения:

$$R = \sum_{i=1}^{m-1} \sum_{j=1}^m \left[\left[v_{k_1,i} < v_{k_1,j} \right] \neq \left[r_{k_2,i} < r_{k_2,j} \right] \right]$$

где $v_{k_1,i}$, $v_{k_1,j}$ ранги вариационного ряда $V_{k_1}^m$; $r_{k_2,i}$, $r_{k_2,j}$ элементы ранговой выборки $S_{k_2}^m$.

Ранговые корреляции Спирмена (3) и Кендалла (4) отличаются различной интерпретацией. Так если коэффициент корреляции Спирмена может рассматриваться как прямой аналог коэффициента корреляции Пирсона (1), вычисленный по рангам, то при вычислении коэффициента корреляции Кендалла проверяется наличие различий между вероятностями порядка (ранга) расположения наблюдаемых данных.

В корреляции методом Спирмена инверсиям придаются дополнительные веса. В результате коэффициент корреляции Спирмена значительно реагирует на несогласие выборок, чем коэффициент корреляции Кендалла. В большинстве случаев $|\rho| > |\tau|$.

На рисунке 2 приведены фрагменты корреляционных диаграмм, исследуемых экспериментальных данных, выполненные для группы из шести атрибутов из макрогруппы 1, приведенных в таблице 2. (Данные по использованию диска): «disk_readbytes, disk_readcount, disk_readtime, disk_writebytes, disk_writecount, disk_writetime». Названия атрибутов на диаграммах изменены на их идентификаторы.

На рисунке 2а приведена корреляционная диаграмма, выполненная методом Пирсона, демонстрирующая заметный разброс численных показателей корреляционных связей между коэффициентами. Меж атрибутная корреляция по Спирмену (рис. 2б), наоборот, демонстрирует полную зависимость атрибутов между собой. На рисунке 2в изображена меж атрибутная корреляция по Кендаллу, которая так же демонстрирует практически полную линейную меж атрибутную зависимость.

Рис. 2. Корреляционные диаграммы, выполненные по трем методам для атрибутов №25...№30: а) метод Пирсона; б) метод Спирмена; в) метод Кендалла

Кластеризация атрибутов

Построение корреляционных диаграмм для 51 атрибута всеми тремя методами позволяет выделить несколько кластеров взаимосвязанных атрибутов. Под кластером понимается набор атрибутов, корреляция между элементами которого превышала 0,95.

Совокупность выделенных кластеров приведена в таблице 3. Каждый обнаруженный кластер соотносится с одной из макрогрупп и ей присвоен специальный маркер атрибута: «интегральные» – «1»; «импульсные» – «2»; «иные» – «3». Если внутри одного типа атрибута (например, «данные по использованию памяти») выделено несколько разных кластеров, то каждый такой кластер дополнительно маркирован римской цифрой.

Таблица 3

Сравнение кластеризованных атрибутов по результатам построения корреляционных диаграмм тремя методами

Макро-группа	Тип атрибута	Идентификатор атрибута
Атрибуты, однозначно кластеризуемые методами <i>Пирсона, Кендалла и Спирмена</i>		
3	Данные по использованию центрального процессора	idle ⁽¹⁾ , total ⁽⁶⁾ , user ⁽⁷⁾ .
1	Данные по использованию памяти I	actualfree ⁽⁸⁾ , actualusedbytes ⁽⁹⁾ , actualusedpct ⁽¹⁰⁾ .
1	Данные по использованию памяти II	swapfree ⁽¹²⁾ , swapusedbytes ⁽¹³⁾ , swapusedpct ⁽¹⁴⁾ .
1	Данные по использованию диска	readbytes ⁽²⁵⁾ , readcount ⁽²⁶⁾ , readtime ⁽²⁷⁾ .
2	Данные по HTTP запросам	networkbytes ⁽⁴¹⁾ , responsebytes ⁽⁴³⁾ .
2	Данные CURL запроса для определения времени реакции сервера	timeconnect ⁽⁴⁷⁾ , timenamelookup ⁽⁴⁸⁾ , timepretransfer ⁽⁴⁹⁾ .
Уникальный кластер, определенный методами <i>Кендалла и Спирмена</i>		
2	Данные по использованию диска	readbytes ⁽²⁵⁾ , readcount ⁽²⁶⁾ , readtime ⁽²⁷⁾ , writebytes ⁽²⁸⁾ , writecount ⁽²⁹⁾ , writetime ⁽³⁰⁾ .
Уникальный кластер, определенный <i>методом Спирмена</i>		
1	Данные по использованию сети	inpackets ⁽³⁵⁾ , outpackets ⁽³⁸⁾ .
Уникальный кластер, определенный <i>методом Пирсона</i>		
3	Данные по использованию диска	await ⁽¹⁵⁾ , ioreadawait ⁽¹⁷⁾ , iostatqueueavgsize ⁽²⁰⁾ .

Методы Кендалла и Спирмена определяют линейно-зависимый кластер из 6 атрибутов №25...№30. На рисунке 2 видно, что при построении диаграммы методом Пирсона (рис. 2а) атрибут №29 не является линейно зависимым от других атрибутов, а для остальных методов (рис. 2б, 2в) – является. Метод Пирсона определяет уникальных кластер из трех атрибутов: await, ioreadawait, iostatqueueavgsize (№15, 17, 20). Его уникальность заключается в том, что другие методы построения корреляционной диаграммы его не обнаруживают.

В качестве иллюстрации, на рисунке 3 приведено сравнение корреляционных диаграмм, построенных тремя методами, для атрибутов №15, 17, 20. По аналогии, метод Спирмена также определяет собственный уникальный кластер, состоящий из двух атрибутов: «network_inpackets» и «network_outpackets» (№35, 38).

Рис. 3. Корреляционные диаграммы, выполненные для атрибутов №15, №17, №20 методом: а) Пирсона; б) Спирмена; в) Кендалла

Линейная зависимость атрибутов свидетельствует о том, что выборке имеются излишние атрибуты [8], размерность которых можно понизить методом их отбора. Отбор признаков позволяет снизить штраф за избыточность, вызванную присутствием этого атрибута в других выбранных признаках [9]. Проведем отбор признаков по критерию наибольшей независимости каждого атрибута от других. С этой целью введем в рассмотрение «вес» атрибута:

$$W_n = \frac{|r_{n,x_1}| + |r_{n,x_2}| + \dots + |r_{n,x_i}| + \dots + |r_{n,x_m}|}{m},$$

где W_n – «вес» n -го атрибута; $|r_{n,x_i}|$ – коэффициент корреляции между n -м и x_i -м атрибутами по модулю; m – количество атрибутов (в рассматриваемом случае $m = 51$).

Определив «вес» каждого из 51 атрибутов по формуле (6), выполним операцию снижения размерности исходной выборки. Для этого в каждом кластере линейно зависимых атрибутов, представленных в таблице 3, определим атрибут, обладающий наименьшим весом, а все остальные исключим из дальнейшего рассмотрения.

В результате отбора в выборке оставлены атрибуты, имеющие идентификаторы, приведенные в таблице 3 и имеющие надстрочные индексы: 7, 8, 14, 15, 28, 35, 43, 48. Общее количество исключенных атрибутов составило 17.

Корреляционная диаграмма для оставшихся 34 атрибутов, построенная методом Пирсона, приведена на рисунке 4.

Диаграмма дополнительно маркирована цветовым спектром. Наиболее темные участки диаграммы характеризуются наиболее сильной корреляционной связью. Понизив порог максимально допустимой корреляции между атрибутами по шкале Чеддока [10], например, до «высокой» ($0,7 < r < 0,9$) возможно дальнейшее снижение размерности вектора наиболее значимых информативных атрибутов.

Рис. 4. Общий план корреляционной диаграммы для 34 атрибутов при максимально допустимой корреляции между атрибутами $r=0,95$

Бинарная кластеризация

Для задач обнаружения аномальных событий часто достаточно ограничиться двумя кластерами: «аномальный» и «нормальный». В этом случае процесс образования кластеров на примере алгоритма «k средних» реализуется следующим образом:

1. значения каждого атрибута нормализуются и удаляется тренд;
2. для каждого атрибута выполняется алгоритм «k средних» с разделением выборки на два кластера;
3. выполняется интерпретация результатов: набор из двух кластеров для каждого атрибута вручную маркируются на «аномальный» и «нормальный» кластеры.

Результатом кластеризации являются поатрибутно маркированные данные, в которых каждому атрибуту в текущий момент времени присваивается одно из двух состояний: аномальное или нормальное.

Перед кластеризацией необходимо удалить трендовую составляющую из значений атрибутов интегральной группы. Удаление тренда при помощи преобразования временных рядов атрибутов интегральной группы к виду интегрированных временных рядов (DS-ряд, Difference Stationary), [19] позволяет обнаруживать выбросы и повышает разделимость данных при последующей кластеризации. Удаление тренда и последующая нормализация значений атрибута приводит к появлению отрицательных значений атрибута, лежащих в пределах $[-1; 1]$. Интерес представляют значительные выбросы (отклонения) от нулевых значений подобных атрибутов. Для этого целесообразно преобразовать значения данных атрибутов в абсолютные величины.

В качестве иллюстрации этого, на рисунке 5 приведены временные диаграммы атрибута «memory_actualfree» до и после удаления тренда, а также соответствующие им результаты кластеризации (число кластеров – 2). Диаграммы совмещены с распределения плотности значений атрибута при помощи построения скрипичных диаграмм. «Центры» кластеров (в данном случае под центром следует рассматривать некоторый порог по оси ординат графиков распределения плотностей) маркированы черными кружками. На рисунке 5а приведена временная диаграмма атрибута «memory_actualfree» до удаления трендовой составляющей.

Соответствующее распределение плотности численных значений атрибута по оси ординат приведено на рисунке 5б. Данный график позволяет визуализировать наиболее плотные скопления численных значений атрибута. Разделение данных на два кластера («0» – значения, близкие к максимальным, «1» – остальные значения) показано при помощи цветовой маркировки, а также при помощи разделения распределений по разные стороны от вспомогательной вертикальной числовой оси. Удаление тренда (рис. 5в) оставляет только значительные изменения численных переменных за единицу времени (выбросы). Соответствующее распределение плотностей (рис. 5г) однозначно выделяет два кластера: «0», сконцентрированный вблизи нулевых значений атрибута (нормальные значения) и «1», в который включены почти все выбросы, происходящие в значениях атрибута.

Удаление трендовой составляющей методом DS-рядов, нормализация и приведение величин к абсолютным значениям позволяет с большей точностью маркировать аномальные выбросы. Однако не всегда подобное удаление тренда

является допустимым, поскольку непрерывное возрастание или убывание некоторых атрибутов может свидетельствовать о приближающейся аварии.

Рис. 5. Временные диаграммы и границы кластеров для атрибута «memory_actualfree»: а) до удаления тренда; б) после удаления тренда; и соответствующие им графики плотности распределения значений: в) до удаления тренда; г) после удаления тренда

Сведение всех участков стагнации атрибута в область нулевых значений удаляет информацию о часто наблюдаемых стабильных состояниях КС по данному атрибуту. Из графика 5б видно, внутри кластера «1» выделяются два подкластера почти разноточной плотности, а также несколько подкластеров с меньшей плотностью: участков стагнации. Наличие данных участков сигнализирует о более сложном распределении «нормальных» и «аномальных» режимов работы КС, но их рассмотрение и анализ выходят за рамки данной работы.

После удаления трендовой составляющей в численных значениях атрибутов интегральной макрогруппы, их нормализации и приведения к абсолютным значениям возможно проведение операции кластеризации.

Результаты кластеризации рассмотренных выше 34 атрибутов приведены на рисунке 6. По оси абсцисс отложены идентификаторы атрибутов, по оси ординат – их нормированное значение $[0, 1]$. Ромбами отмечены получившиеся центры кластеров. Цветом фона маркированы области допустимых и аномальных значений по каждому атрибуту.

Рис. 6. Результат работы алгоритма «k-means» для 34 атрибутов

Интерпретация результатов кластеризации, выполненная вручную для каждого нормализованного атрибута, показывает, что нормальные значения для каждого атрибута сконцентрированы около нулевых значений. В отличие от статистического анализа, где при помощи построения корреляционных диаграмм возможно выделение зависимости между значениями разных атрибутов, кластерный анализ при по-

мощи методов интеллектуального анализа данных позволяет исследовать схожесть как значений одного атрибута, так и многоатрибутных наблюдений.

Однако кластерный анализ, не предоставляя статистических выводов, тем не менее, позволяет обнаруживать выбросы (аномалии) в структуре данных [20, 21]. Результаты бинарной классификации, при условии хорошей разделимости данных, легко интерпретируются человеком. Комбинирование методов статистического и кластерного анализа повышает точность выявления аномалий в данных [22] за счет извлечения большего количества полезной информации (информация о межатрибутой зависимости и о внутриатрибутовой схожести), а также позволяет производить отбор атрибутов по их информационной значимости.

Выводы

Комбинирование методов статистического и кластерного анализа повышает точность выявления аномалий в данных за счет извлечения большего количества полезной информации, а также позволяет производить отбор атрибутов по их информационной значимости.

Показано, что комбинирование трех методов корреляционного анализа при отборе атрибутов по их информационной значимости позволяет отбросить большее количество атрибутов за счет выявления новых межатрибутных связей (межатрибутных кластеров). Разные методы корреляционного анализа выделяют разные кластеры среди атрибутов, что доказывает целесообразность их использования в совокупности. Отбор атрибутов, проведенный в ходе корреляционного анализа, позволил сократить исходную выборку экспериментальных данных на 41%.

Показана необходимость ручной предобработки данных перед проведением кластеризации на примере ручного разделения атрибутов на макрогруппы, анализа распределения плотности при помощи скрипичных диаграмм и удаления трендовой составляющей методом DS-рядов.

Показано, что удаление трендовой составляющей методом DS-рядов, нормализация и приведение величины к абсолютным значениям позволяет с большей точностью маркировать аномальные выбросы. Показано, что удаление тренда не всегда допустимо, т.к. наличие непрерывно возрастающей или убывающей величины в некоторых атрибутах может нести в себе информацию о приближающейся аварии.

Проведенный кластерный анализ позволил получить по-атрибутно маркированные данные, в которых каждому атрибуту в текущий момент времени присваивается одно из двух состояний: аномальное или нормальное.

Показано, что результаты бинарной классификации с дополнительной цветовой маркировкой, при условии хорошей разделимости данных, легко интерпретируются человеком.

Предварительный анализ временных диаграмм по каждому атрибуту позволил выделить две атрибутные макрогруппы: «интегральные атрибуты» и «импульсные атрибуты». Корреляционный анализ, проведенный в последствии, не выявил коллизий с выделенными макрогруппами, но выделил кластеры внутри каждой макрогруппы. Удаление тренда и последующая нормализация выборок позволили обнаружить аномальные выбросы в численных значениях атрибутов «интегральной» макрогруппы. Интерпретация результатов кла-

стеризации, выполненная вручную для каждого нормализованного атрибута, показывает, что нормальные значения для всех атрибутов сконцентрированы около нулевых значений.

Литература

1. Шелухин О.И. Сетевые аномалии. Обнаружение, локализация, прогнозирование. 4 изд. М.: Научно-техническое издательство «Горячая линия – Телеком», 2019. 448 с.
2. Talleb N. The Black Swan. London: Allen Lane, Penguin Books Ltd. 2011. 480 с.
3. Бородин А.В., Тищенко В.А. Имитационное моделирование неоднородного процесса деградации в системе с постепенными и внезапными отказами // Труды карельского научного центра российской академии наук. 2018. №7. С. 3-13.
4. Шелухин О.И., Осин А.В., Костин Д.В. Мониторинг и диагностика аномальных состояний компьютерной сети на основе изучения "исторических данных" // T-Comm: Телекоммуникации и транспорт. 2020. Том 14. №4. С. 23-30.
5. Шелухин О.И., Костин Д.В., Резник И.Ю. Мониторинг и структура аномальных паттернов системных журналов компьютерных систем // REDS: Телекоммуникационные устройства и системы. №2. 2020. С. 3-8.
6. Komleva N.O., Zinovatnaya S.L., Liubchenko V.V. Methodology of information monitoring and diagnostics of objects represented by quantitative estimates based on cluster analysis // Applied Aspects Of Information Technology. 2020. №1. С. 376-392.
7. Гмурман В. Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. 10-е издание, стереотипное. Москва: Высшая школа, 2004. 479 с.
8. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. 2003. Т. 3.
9. Peng H. C., Long F., Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005. Т. 27, вып. 8. DOI:10.1109/TPAMI.2005.159
10. Сизова Т.М. Статистика: Учебное пособие. СПб.: СПб ГУ ИТМО, 2005. 190 с.
11. sklearn.cluster.KMeans // scikit-learn Machine Learning in Python URL: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> (дата обращения: 14.11.2020).
12. Gueniche T., Fournier-Viger P., Raman R., Tseng V. S. CPT+: Decreasing the Time/Space Complexity of the Compact Prediction Tree // Conference: Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2015. pp. 1-12.
13. Daw S., Finney C., Tracy E. R. A Review of Symbolic Analysis of Experimental Data // Review of Scientific Instruments. 2003. №2. С. 915-930. DOI: 10.1063/1.1531823
14. Boriah S., Chandola V., Kumar V. Similarity measures for categorical data: A comparative evaluation // Proc. 8th SIAM International Conference on Data Mining. 2008. P. 243-254.
15. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. СПб: БХВ-Петербург, 2003. 654 с.
16. Christos B. Detection and Prediction of Rare Events in Transaction Databases // International Journal on Artificial Intelligence Tools. 2007. №1-20.
17. Witten I.H., Frank E., Hall M.A. Data mining practical machine learning tools and techniques, 3rd edition. 2011. pp. 217-221.
18. Daw C. S., Kennel, M. B., Finney C.E.A., Connolly F. T. Observing and modeling nonlinear dynamics in an internal combustion engine // Physical Review E. 1998. С. 2811.
19. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело, 2007. 504 с. ISBN 978-5-7749-0473-0.
20. Яценков К.Г., Дымко К.С., Уханов Н.О., Якунин Ю.Ю., Хныкин А.В. Использование алгоритма поиска аномалий в задаче повышения качества метео-данных // В сборнике: интеллектуальные информационные системы: теория и практика. Сборник научных статей по материалам I Всероссийской конференции. Курск, 2020. С. 40-48.
21. Катенко Ю.В. Применение алгоритмов кластеризации для выявления аномалий во временных рядах при анализе данных банковских и телекоммуникационных систем // Охрана, безопасность, связь. 2018. Т. 3. № 3(3). С. 88-94.
22. Попов Е.Ф., Тюкова А.А., Фучко М.М., Захаров А.А. Выявление нетипичных событий средствами статистического анализа // Вестник УрФО. Безопасность в информационной сфере. 2015. № 1 (15). С. 44-47.

SELECTION OF METRIC AND CATEGORICAL ATTRIBUTES OF RARE ANOMALOUS EVENTS IN A COMPUTER SYSTEM USING DATA MINING METHODS

Oleg I. Sheluhin, Moscow Technical University of Communication and Informatics, Moscow, Russia, sheluhin@mail.ru

Dmitry I. Rakovsky, Moscow Technical University of Communication and Informatics, Moscow, Russia, dimitor1998@mail.ru

Abstract

The process of marking multi-attribute experimental data for subsequent use by means of data mining in problems of detection and classification of rare anomalous events of computer systems (CS) is considered. The labeling process is carried out using three methods: manual preprocessing, statistical analysis and cluster analysis. Among the attributes of the metric type, the authors identified two macrogroups: "integral attributes" and "impulse attributes". It is shown that the combination of statistical and cluster analysis methods increases the accuracy of detecting anomalous events in the CS, and also allows the selection of attributes according to their information significance. The expediency of manual preprocessing of data before clustering is shown by the example of dividing attributes into macrogroups, analyzing the density distribution using violin plot and removing the trend component using the method difference stationary series. With the help of construction of violin diagrams (Violin plot) for the attribute of the "integral" macrogroup, the distribution of states of the CS is shown. It is shown that the removal of the trend component by the DS-series method, normalization and reduction to absolute values allows more accurate marking of anomalous outliers, but this is not always acceptable. The interpretation of the clustering results performed for each normalized attribute shows that the normal values for all attributes are concentrated around zero values. The result of labeling experimental data is attribute-labeled data, where each attribute at the current time is assigned one of two states: abnormal or normal.

Keywords: attribute value markup, correlation analysis, k-means, experimental data, trend, violin plot, difference stationary series.

References

1. Sheluhin O.I. (2019) Network anomalies. Detection, localization, forecasting. 4th ed. Scientific and technical publishing house "Hot line - Telecom", Moscow, 448 p. (in Russian)
2. Talleb N. (2011) The Black Swan. Allen Lane, Penguin Books Ltd, London, 480 p.
3. Borodina A.V., Tishchenko V.A. (2018) Simulation modeling of a non-uniform degradation process in a system with gradual and sudden failures. *Proceedings of the Karelian Scientific Center of the Russian Academy of Sciences*, no. 7, pp. 3-13. (in Russian)
4. Sheluhin O.I., Osin A.V., Kostin D.V. (2020) Health monitoring of a computer network based on sequential analysis of serial pattern. *T-Comm*, vol. 14, no. 2, pp. 9-16. (in Russian)
5. Sheluhin O.I., Kostin D.V., Reznik I.Yu. (2020) Monitoring and structure of abnormal patterns of system logs of computer systems. *REDS: Telecommunication devices and systems*, vol. 2, pp. 3-8. (in Russian)
6. Komleva N.O., Zinovatnaya S.L., Liubchenko V.V. (2020) Methodology of information monitoring and diagnostics of objects represented by quantitative estimates based on cluster analysis. *Applied Aspects of Information Technology*, no. 1, pp. 376-392.
7. Gmurman V.E. (2004) Probability theory and mathematical statistics: a textbook for universities. 10th edition, stereotyped. Higher school, Moscow, 479 p. (in Russian)
8. Guyon I., Elisseeff A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, Vol. 3.
9. Peng H.C., Long F., Ding C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8. DOI: 10.1109 / TPAMI.2005.159
10. Sizova T.M. (2005) Statistics: Study Guide. SPb GU ITMO, Saint Petersburg, 190 p. (in Russian)
11. sklearn.cluster.KMeans // scikit-learn Machine Learning in Python URL: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> (accessed 11/14/2020).
12. Gueniche T., Fournier-Viger P., Raman R., Tseng V.S. (2015) CPT +: Decreasing the Time / Space Complexity of the Compact Prediction Tree. *Conference: Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 1-12.
13. Daw S., Finney C., Tracy E.R. (2003) A Review of Symbolic Analysis of Experimental Data. *Review of Scientific Instruments*, no. 2, pp. 915-930.
14. Boriah S., Chandola V., Kumar V. (2008) Similarity measures for categorical data: A comparative evaluation. *Proc. 8th SIAM International Conference on Data Mining*, pp. 243-254.
15. Gasfield D. Strings, (2003) Trees and Sequences in Algorithms. SPb: BHV-Petersburg, Saint Petersburg, 654 p. (in Russian)
16. Christos B. (2007) Detection and Prediction of Rare Events in Transaction Databases. *International Journal on Artificial Intelligence Tools*, no. 5. P. 829-848.
17. Witten I.H., Frank E., Hall M.A. (2011) Data mining practical machine learning tools and techniques, 3rd edition, pp. 217-221.
18. Daw C. S., Kennel, M. B., Finney C.E.A., Connolly F. T. (1998). Observing and modeling nonlinear dynamics in an internal combustion engine. *Physical Review E*. P. 2811-2819.
19. Magnus Ya.R., Katyshev P.K., Perebeckij A.A. (2007). Ekonometrika. Nachal'nyj kurs. Moscow: Delo. 504 p. (in Russian)
20. Yashchenkov K.G., Dymko K.S., Uhanov N.O., Yakunin Yu.Yu., Hnykin A.V. (2020). Ispol'zovanie algoritma poiska anomalij v zadache povysheniya kachestva meteorodannyh. In the collection: intellektual'nye informacionnye sistemy: teoriya i praktika, a collection of scientific articles based on the materials of the I All-Russian scientific-practical conference. Kursk. pp. 40-48. (in Russian)
21. Katenko Yu.V. (2018). Primenenie algoritmov klasterizacii dlya vyyavleniya anomalij vo vremennyh ryadah pri analize dannyh bankovskih i telekommunikacionnyh system. *Ohrana, bezopasnost', svyaz'*. no 3. pp. 88-94. (in Russian)
22. Popov E.F., Tyukova A.A., Fuchko M.M., Zaharov A.A. (2015). Vyyavlenie netipichnyh sobytij sredstvami statisticheskogo analiza. *Vestnik UrFO. Bezopasnost' v informacionnoj sfere*. No 1. P. 44-47. (in Russian)

Information about authors:

Oleg I. Sheluhin, Moscow Technical University of Communication and Informatics, Professor at the Department of Information Security, Moscow, Russia

Dmitry I. Rakovsky, Moscow Technical University of Communication and Informatics, Moscow, Russia