КЛАСТЕРИЗАЦИЯ ДАННЫХ В ЗАДАЧАХ ОБНАРУЖЕНИЯ АНОМАЛИЙ НА ОСНОВЕ ОРТОГОНАЛЬНЫХ ФИЛЬТРОВ

А.В. Скатков¹, Ю.Е. Шишкин^{1,2}

¹ ФГАОУ ВО «Севастопольский государственный университет»,
РФ, г. Севастополь, ул. Университетская, 33

² Институт природно-технических систем,
РФ, г. Севастополь, ул. Ленина, 28
E-mail: yourockpro@gmail.com

Предлагается подход к решению задачи оперативного обнаружения аномалий в данных мониторинга с использованием адаптивной цифровой фильтрации на основе ортогонального фильтра, использующего разложение БПФ. Выполнено математическое моделирование и построена процедура кластеризации в режиме реального времени. Проиллюстрирована работа предложенного подхода на примере выборки данных зондирования акватории вод г. Севастополя с использованием гидрологического SVP зонда.

Ключевые слова: мониторинг, математическое моделирование, Большие Данные, цифровая фильтрация, дискретное преобразование Фурье, детектирование аномалий, кластеризация, критические системы, интеллектуальный анализ данных.

Введение. В настоящее время автоматизированный контроль качества данных систем контроля окружающей среды, детектирование выбросов ключевых параметров и аномалий, характеризующих состояние объекта мониторинга с целью предупреждения аварийных состояний критических инфраструктур, является актуальной научной задачей [1].

Задачи обнаружения аномалий и кластеризации имеют схожую формулировку и метрики качества, например, чувствительность алгоритма классификации TPR , AUC , ROC , PR , AUC и др. [2]. При этом в задачах обнаружения аномалий имеет место значительный дисбаланс классов, когда аварийные состояния относительно редки, вследствие чего сбор статистики ключевых параметров в таких состояниях становится затруднительным [3].

Постановка задачи. Проверяется гипотеза зависимости чувствительности системы бинарного классификатора к возникающим аномалиям (управляемым возмущениям) во входном потоке объектов с использованием ортогонального фильтра в зависимости от порога α и размера обучающей выборки n . В качестве метрики качества используется соотношение между долей объектов от общего количества носителей признака,

верно классифицированных как несущих признак.

Материалы и методы. Для иллюстрации работы предложенного подхода в качестве исходных данных используется выборка данных зондирования, выполненного SVP зондом в акватории вод г. Севастополя [4], при этом заданы наборы данных: A – заведомо содержащий аномальные значения и D – содержащий немаркированные как аномальные, так и корректные данные.

Базовый подход системы поддержки принятия решений (СППР) заключается в подмене реального объекта наблюдения его математической моделью, адаптивно подстраивающейся под изменения состояния объекта, проявляющиеся в статистически значимом изменении параметров наблюдений с последующим сравнением ожидаемого и фактически поступившего потоков данных [5]. Значимые отклонения данных имитационной модели от реальных наблюдений могут быть классифицированы как аномалии, при интерпретации результатов следует учитывать природу задачи, функционал качества и др.

Во множестве задач обнаружения аномалий принято выделять два подмножества:

- выявление выбросов данных;
- выявление новых данных.

Каждая запись в наборе данных, в общем случае, представляет собой объект в многомерном пространстве поля наблюдений, которое может характеризоваться координатой и временем. Например, при анализе значения температуры во временном ряду выбросами можно считать, как аномально большие и малые значения, так и области с нехарактерными изменениями температурного градиента во времени. В то время как новизна обычно характеризуется принципиально новым поведением объекта и оценивается как мера соответствия очередного поступившего в СППР объекта множеству уже имеющихся объектов [6].

Информационная метрика оценки параметров адаптивного фильтра. Для расчета меры соответствия могут использоваться различные методы статистического анализа, такие как метод χ^2 , регрессионный анализ, пороговый анализ и др. Одним из наиболее часто используемых является метод среднеквадратических отклонений, в котором на основе подмножества предыдущих значений $[d_{i-n}; d_i]$ размером n множества данных $[d_0; d_{\max}] \in D$ строится доверительный интервал $[m - k \cdot s; m + k \cdot s]$ для каждой $d_i \in D$, где m – математическое ожидание; s – среднеквадратическое отклонение; k – коэффициент ширины интервала [7]. За счет использования скользящего окна, ширина которого – n определяет скорость перестроения СППР при естественных изменениях в объекте наблюдения, проявляется адаптационная способность метода, позволяющая игнорировать данные о новом состоянии объекта, не носящие системный характер.

Такой подход позволяет детектировать широкий класс аномальных состояний объекта мониторинга, проявляющихся в виде резких перепадов наблюдаемых параметров относительно их стационарного положения. При этом в случае постепенного изменения параметров во время перехода объекта наблюдения в предаварийное, а затем в аварийное состояние или их целенаправленном изменении со стороны злоумышленника возникает ситуация пере-

обучения СППР, характеризующаяся ее функциональной неспособностью к детектированию аномалий [8]. Устранение указанной уязвимости метода возможно следующими способами [9]:

- увеличить размер n обучающей выборки, что в свою очередь приведет к снижению адаптационных свойств СППР;

- определить защищенную область репрезентативной выборки данных системы, заведомо находящейся в стационарном состоянии.

Следует отметить, что под новизной понимается абсолютно новое состояние объекта мониторинга, проявляющееся в изменении его реакции на внешние сигналы, несвойственные аномалии наблюдаемых характеристик. Сбор обучающей выборки, содержащей все возможные неисправности, поломки, аварии крайне трудоемкий и зачастую нереализуемый процесс, что обусловлено низкой частотой их возникновения. Поэтому на практике применяется обучение на выборке данных достоверно характеризующей нормальное (штатное) множество состояний объекта наблюдения.

Определим основные причины возникновения аномалий в наборах данных:

- ошибки в исходных данных, погрешности измерений, неверное округление, дискретизация аналогового сигнала, некорректность формата представления данных;

- наличие источника шума, неверно классифицированные объекты;

- наличие объектов сторонних выборок (в частности, записи вышедших из строя средств мониторинга, датчиков).

Аномалиями обычно являются экстремальные значения параметров мониторинга, но это не всегда так. Приведем пример визуализации участка данных зондирования, выполненного водах акватории г. Севастополя (рис. 1). Построение производится в пространстве, характеризуемом двумя признаками, температурой и скоростью звука в воде, для глубины 0–20 м.

Из рис. 1 видно, что данные содержат как небольшие шумы, расположенные в окрестности линии тренда, так и большие возмущения, модуль которых доходит до 10 м/с.

Рис. 1. Точечная диаграмма зависимости скорости звука от температуры по данным 96-го рейса НИС «Профессор Водяницкий», замер выполнен 18.07.2017 в 20:30, температура воздуха +26°C, направление ветра 70°, скорость ветра 6 м/с, волнение 2 б., координаты станции 44°29'15.4" с.ш. 33°30'37.4" в.д.

Если относительно слабые выбросы не сильно искажают границы наблюдаемого кластера и на физическом уровне обуславливаются наличием тонкой микроструктуры, стратификации поверх-

ностного слоя, то сильные выбросы действительно являются аномальными, что легко доказать с использованием термодинамического уравнения состояния морской воды TEOS-10 [10]:

$$S(C, T, P) = \frac{w_1 + w_2 T - w_3 T^2 + w_4 T^3 + w_5 |P| + w_6 P^2 - w_7 |P^3| - w_8 T |P^3| + w_9 T^2 |P| - C}{-w_{10} + w_{11} T - w_{12} T^2 - w_{13} |P|}, \quad (1)$$

где C – скорость звука; T – температура; P – давление; значения коэффициентов w составляют:

$$w_1 = 14025;$$

$$w_2 = 5;$$

$$w_3 = 0.0544;$$

$$w_4 = 2.1 \cdot 10^{-4};$$

$$w_5 = 1.56 \cdot 10^{-2};$$

$$w_6 = 2.55 \cdot 10^{-7};$$

$$w_7 = 7.3 \cdot 10^{-12};$$

$$w_8 = 9.5 \cdot 10^{-13};$$

$$w_9 = 3 \cdot 10^{-7};$$

$$w_{10} = 1.33;$$

$$w_{11} = 1.23 \cdot 10^{-2};$$

$$w_{12} = 8.7 \cdot 10^{-5};$$

$$w_{13} = 1.43 \cdot 10^{-5}.$$

Согласно представленному выше уравнению, область, соответствующая всплескам пары скорости звука и температуры возможна только при уровне солёности, значительно выходящем за рамки океанографического диапазона, доходя в абсолютных значениях до 10‰ и с перепадом до 7‰, что несвойственно для региона. Одним из способов устранения таких аномалий является выполнение линейного преобразования солёности так, чтобы оно лежало в допустимом интервале и восстановления, с использованием симплекс метода, соответствующего ему пространства температуры и скорости звука. Очевидно, что применение такого преобразования целесообразно производить не для всей области данных, что в противном случае приведет к значительному искажению результата, а с использованием порогового фильтра. Учитывая нелинейность возникающих искажений и недетерминированный характер вызывающих их аномалий, оптимальным будет применение ортогонального адаптивного фильтра.

Предположим, что входной поток данных D , задан на интервале $[d_0; d_{\max}]$, аномалии a_i в данных являются случайными не коррелированными с d_i величинами, принимающими значение $a_j = 0$ с вероятностью $p \geq 0.99$ на всей области определения и нулевым средним значением $M[a_j] = 0$.

Ортогонализация предназначенного для обработки потока данных выполняется на основе дискретного преобразования Фурье (ДПФ), но может производиться и на основе других ортогональных преобразований [11]. Предположим, что входящий поток данных может быть описан функцией с ограниченным спектром, другими словами $F(\nu) = 0$, $|\nu| > \nu_0$ при этом выполняется неравенство Найквиста:

$$\Delta_x \leq \frac{1}{2\nu_0},$$

где Δ_x – шаг дискретизации по частоте.

Выбирается длина периодической последовательности n , используемая в качестве обучающей выборки, при этом будет рациональным принять $n = 2^m$. Такое допущение позволит использовать алгоритм быстрого преобразования Фурье, что значительно снизит число необходимых вычислительных операций при выполнении фильтрации длинных последовательностей $n \geq 10^3$.

Разобьем множество входного потока D на два непересекающихся подмножества D_{normal} и $D_{anomaly}$ такие, что:

$$\begin{cases} d_i \in D_{anomaly} \forall i \in [0, n-1], \text{ где } 1 - \frac{f_{filtered}(i)}{d_i} \geq \alpha \\ d_i \in D_{normal} \forall i \in [0, n-1], \text{ где } 1 - \frac{f_{filtered}(i)}{d_i} < \alpha \end{cases}, \quad (4)$$

где α – порог распознавания аномалии.

Высокая скорость сходимости в численных методах, вообще говоря, является преимуществом, вследствие того, что оптимум рабочей функции достигается при минимальном числе итераций. Сле-

дует отметить, что для СППР это справедливо, когда рабочая функция, на которой происходит поиск оптимума, заранее известна, но на практике это не всегда так.

$$F_p(l) = \frac{1}{n} \sum_{j=0}^{n-1} f_p(j) e^{-\frac{2\pi i}{n} jl},$$

где $l = 0, \dots, n-1$, $i = \sqrt{-1}$, $F_p(l)$ имеет период n , каждому l соответствует $\omega_i = \Delta_\omega l$, $\Delta_\omega = \frac{2\pi}{\Delta_x n}$ – шаг дискретизации.

Введем функцию фильтрующего множителя, определяющего ошибку фильтрации, уменьшая коэффициенты ДПФ, обусловленные аномалиями:

$$W_a(\omega) = \frac{1}{1 + \omega^{2p}}, \quad (2)$$

где p – коэффициент крутизны фильтрации для всех $\omega > \omega_0$, ω_0 – граница частотного интервала информационных составляющих входного потока данных.

С использованием функции (2) получим вектор фильтрованных коэффициентов ДПФ:

$$F_{filtered}(l) = W_a(l) F_p(l), \quad l \in [0, n-1],$$

по значениям которого произведем спектральный синтез, выполнив обратное ДПФ:

$$f_{filtered}(j) = \sum_{l=0}^{n-1} F_{filtered}(l) e^{\frac{2\pi i}{n} jl}, \quad (3)$$

где $j \in [0, n-1]$.

Построим точечную диаграмму по отфильтрованным данным D_{normal} (рис. 2).

Рис. 2. Точечная диаграмма зависимости скорости звука от температуры после фильтрации по данным 96-го рейса НИС «Профессор Водяницкий», замер выполнен 18.07.2017 в 20:30, температура воздуха +26°C, направление ветра 70°, скорость ветра 6 м/с, волнение 2 б., координаты станции 44°29'15.4" с.ш., 33°30'37.4" в.д.

Функция вычисляется приближенно на основе входного потока данных и постоянно дополняется и адаптируется при поступлении новых данных. В таком случае, при снижении скорости сходимости имеет место фильтрация входного сигнала, которая снижает шумовую составляющую.

Произведем анализ зависимости размера обучающей выборки n на характер поверхности функции отклонения фактических данных от прогнозируемых (функция аномалии f), в силу (3) и (4) для $n = 20$ получим (рис. 3).

Рис. 3. График поверхности функции аномалии f в пространстве скорости звука и температуры при размере обучающей выборки $n = 20$, где $f = 0$ соответствует минимальной вероятности аномалии в данной области и $f = 1$ – максимальной вероятности

При малых значениях n достаточно быстро наблюдается эффект переобучения, проявляющийся в наличии пиковых значений $f = 1$ в областях, где уже были обнаружены аномалии. С другой стороны, при увеличении объема обучающей выборки (рис. 4), функция f «сглаживается», тем самым затрудняя процесс де-

тектирования данных бинарным классификатором и повышая требования к обоснованности выбора порогового значения детектора α . При этом функция f становится значительно более чувствительной к отклонениям потока поступающих данных, и уменьшаются пиковые значения $\max(f) \leq 0.6$.

Рис. 4. График поверхности функции аномалии f в пространстве скорости звука и температуры при размере обучающей выборки $n = 200$

Выбор порогового уровня отклонения α крайне важный процесс, по причине того что он определяет какие именно данные будут признаны аномальными. Выбор основывается на экспертных решениях и имеющейся априорной информации о аномалиях. В соответствии с выбранной моделью аномалий, была принята вероятность того, что очередной поступивший в СППР объект не является

ся аномалией с вероятностью $p \geq 0.99$. Это позволяет построить управляющую матрицу бинарного классификатора при низком $\alpha = 0.2$ (рис. 5) и высоком $\alpha = 0.6$ (рис. 6) порогах. В результате решения задачи оптимизации получим результирующую для СППР диаграмму управляющей матрицы бинарного классификатора (рис. 7).

Рис. 5. Диаграмма управляющей матрицы бинарного классификатора, где синим цветом обозначены области, данные в которых признаются аномальными, при пороге $\alpha = 0.2$

Рис. 6. Диаграмма управляющей матрицы бинарного классификатора, где синим цветом обозначены области, данные в которых признаются аномальными, при пороге $\alpha = 0.6$

Рис. 7. Результирующая диаграмма управляющей матрицы бинарного классификатора, где синим цветом обозначены области, данные в которых признаются аномальными, оранжевым цветом – истинные данные, белым цветом – зона неопределенности

Заключение. Предложен подход к решению задачи оперативного обнаружения аномалий в данных мониторинга с использованием адаптивной ортогональной цифровой фильтрации и математического моделирования в режиме реального времени. Особенности подхода проиллюстрированы на примере разложения с использованием дискретного преобразования Фурье, но этот же подход может быть использован и при других ортогональных преобразованиях.

На основе предложенного подхода разработан бинарный классификатор, осуществляющий сравнение реального объекта наблюдения с его математической моделью, адаптивно подстраивающейся под изменения состояния объекта, проявляющиеся в статистически значимом изменении параметров наблюдений.

Работа бинарного классификатора проиллюстрирована на выборке данных зондирования скорости звука и температуры, полученных с использованием *SVP* зондом в акватории вод г. Севастополя.

Исследование зависимости чувствительности системы бинарного классификатора к возникающим аномалиям (управляемым возмущениям) во входном потоке объектов с использованием ортогонального фильтра в зависимости от размера обучающей выборки n показало, что малые значения $n = 16$ обеспечивают высокую скорость реакции на возникновение аномалий, но при этом быстро возникает эффект переобучения и относительно большая вероятность ложных срабатываний. Размеры обуча-

ющей выборки $n \geq 256$ значительно в меньшей степени подвержены переобучению, но обладают меньшей чувствительностью, что в свою очередь накладывает повышенные требования к точности выбора порога α . Выбор параметра n остается открытым, он должен быть достаточно большим для того, чтобы обеспечить достаточную чувствительность фильтра, обеспечивая возможность адаптивной настройки порога α и во многом определяет вычислительную сложность алгоритма.

Предложен способ построения адаптивной управляющей матрицы бинарного классификатора при заданном пороге α , позволяющей оперативно принимать решения о возникновении аномалии в данных наблюдений, их отсутствии или зоны неопределенности.

Предложенный подход разработан для использования в СППР, реализующийся в виде бинарного классификатора и блоков задания порога α и размера обучающей выборки n , используемых для минимизации ошибки первого и второго, рода. Предназначен для нахождения пограничных зон неуверенного распознавания аномалий, поскольку целью мониторинга является не только сбор информации об объекте, но и обнаружение критических событий для оперативного принятия неотложных мер по ликвидации критических состояний.

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (грант № 15-29-07936/17).

СПИСОК ЛИТЕРАТУРЫ

1. *Анализ данных и процессов: учеб. пособие* / А.А. Барсегян, М.С. Куприянов, И.И. Холод [и др.]. 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.
2. *Зубриенко Г.А., Лапоница О.Р. Методы оптимизации выборки данных для определения аномального трафика* // International Journal of Open Information Technologies. 2016. Т. 4. № 10. С. 1–8.
3. *Скатков А.В., Шишкин Ю.Е. Модель обнаружения аномалий в полях наблюдений с использованием систем параметрического мониторинга* // Системы контроля окружающей среды. 2017. Вып. 10 (30). С. 48–53.
4. *Греков А.Н., Греков Н.А., Шишкин Ю.Е. Исследование характеристик профилографа скорости звука и коррекция результатов измерения* // Системы контроля окружающей среды. 2017. Вып. 10 (30). С. 24–30.
5. *Шишкин Ю.Е. Свидетельство № 2017664038 Российская Федерация. Модуль снижения избыточности данных мониторинга «МОСИДАМ»: свидетельство о государственной регистрации программы для ЭВМ / заявитель и правообладатель Ю.Е. Шишкин; № 2017660984; заявл. 27.10.2017, опубл. 14.12.2017; Бюл. № 12. 1 с.*
6. *Шишкин Ю.Е. Исследование возможностей систем обнаружения взаимодействий в методологии Больших Данных* // Фундаментальные основы инновационного развития науки и образования: монография / под ред. Г.Ю. Гуляева. Пенза: Наука и Просвещение, 2017. С. 55–73.
7. *Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения*. М.: Высшая школа, 2007. 491 с.
8. *Боев В.Д. Концептуальное проектирование систем в Anylogic 7 и GPSS World*. М.: НОИ Интуит, 2016. 556 с. ISBN: 978-5-9556-0161-8.
9. *Shishkin Y.E. Big Data visualization in decision making* // Science in Progress: тез. Всерос. науч.-практ. конф. магистрантов и аспирантов. Новосибирск 20 октября 2016 г. Новосибирск: НГТУ, 2016. С. 203–205.
10. *The International Thermodynamic Equation of Seawater 2010 (TEOS-10): Calculation and Use of Thermodynamic Properties* / T.J. McDougall, Rainer Feistel, F.J. Millero [et al.]. Intergovernmental Oceanographic Commission IOC of Unesco, March 2010. 218 pp.
11. *Девятков В.В. Методология и технология имитационных исследований сложных систем: современное состояние и перспективы развития: монография*. СПб.: Вузовский учебник, 2013. 448 с.

DATA CLUSTERING IN ANOMALIES DETECTION TASKS BASED ON ORTHOGONAL FILTERS

A.V. Skatkov¹, Y.E. Shishkin^{1,2}

¹ Federal State Educational Institution of Higher Education «Sevastopol State University», Russian Federation, Sevastopol, Universitetskaya St., 33

² Institute of Natural and Technical Systems, Russian Federation, Sevastopol, Lenin St., 28

An approach to solve the problem of anomalies operative detection in monitoring data using adaptive digital filtering based on an orthogonal filter using FFT decomposition is proposed. A mathematical simulation and a real-time clustering procedure was performed. The work of the proposed approach is exemplified by the sampling Sevastopol water area sampling data using a hydrological SVP probe.

Keywords: monitoring, mathematical modeling, Big Data, digital filtering, discrete Fourier transform, detection of anomalies, clustering, critical systems, data mining.