УДК 681.3 DOI: 10.33075/2220-5861-2019-3-35-38

МЕРА КУЛЬБАКА В ЗАДАЧАХ ДИНАМИЧЕСКОЙ КЛАСТЕРИЗАЦИИ НАБЛЮДЕНИЙ СОСТОЯНИЯ ОКРУЖАЮЩЕЙ СРЕДЫ

А.В. Скатков[1], А.А. Брюховецкий[1], Д.В. Моисеев[1], Ю.Е. Шишкин[1,2]

[1]Севастопольский государственный университет,РФ, г. Севастополь, ул. Университетская, 33
E-mail: dmitriymoiseev@mail.ru
[2]Институт природно-технических систем, РФ, г. Севастополь, ул. Ленина, 28

Данная работа посвящена исследованию методов обнаружения аномалий экосистем акватории на основе кластеризации данных. Рассматриваются типовые операции над кластерами и основные метрики, базирующиеся на информационной мере Кульбака, а также некоторые важные применения – предсказание, распознавание аномалий. В работе предлагается новый подход, который в значительной степени позволяет преодолеть требования к реализации известных методов контроля состояния экосистем.
Ключевые слова: анализ, аномалии, большие данные, кластеризация, мера Кульбака, мониторинг, окружающая среда, прогнозирование, экосистемы.

Поступила в редакцию: 24.07.2019. После доработки: 06.09.2019.

Введение. С ростом городов, развитием производства техногенная трансформация среды приобретает глобальный характер. Настоящая работа посвящена развитию методов обнаружения аномалий экосистем акватории на основе кластеризации данных. Решение задач, связанных с обнаружением аномалий водной среды, приобретает особую актуальность [1–3].

Предлагаемый подход широко используется в задачах распознавания образов, интеллектуального анализа данных. Особенный интерес к методам анализа данных возник в связи с развитием средств сбора и хранения большого объема данных с многочисленных датчиков. В условиях появления новых технологий «Интернет вещей» и «Big Data» перед специалистами из разных областей человеческой деятельности возникает задача обработки большого объема данных в реальном времени и с высоким уровнем достоверности [4–10].

Материалы и методы. В настоящее время процесс оперативного выявления аномалий данных мониторинга окружающей среды и критических объектов инфраструктуры является комплексной, трудоемкой и трудно формализуемой задачей [3, 11]. Совместное использование средств оперативного мониторинга, имитационного моделирования и вероятностных моделей позволяет прогнозировать динамику изменения состояния экосистемы, предупреждать о возможных аномалиях и превентивно выполнять корректирующие действия, предотвращая тем самым возникновение критических ситуаций [4]. Использование модели распределенных или облачных вычислений с применением агентного подхода, используемого для обработки Больших Данных, позволит решить задачи анализа и репрезентативной визуализации срезов измерений в многомерных базах данных [6–10], производить анализ цикличности изменений эколого-геологических изменений, одновременно используя статистику более чем за 100 лет мониторинга окружающей среды.

С целью развития известных подходов к решению указанных задач предлагается новый подход, который в значительной степени позволяет преодолеть требования к реализации известных методов мониторинга состояния экосистем, в частности:

- обработки больших объемов априорной информации для настройки параметров системы мониторинга;
- принятия достоверного решения о состоянии экосистемы в результате обработки большого объема текущих данных;
- трудно реализуемую стратегию обработки данных в реальном времени.

В данной работе предлагается модель обнаружения аномалий на основе

динамической кластеризации, в которой рассматриваются следующие ситуации:

- образование новых кластеров;
- слияние кластеров;
- расщепление кластеров;

Полученные результаты. Основными преимуществами использования меры Кульбака в задачах динамической кластеризации наблюдений состояния окружающей среды (ОС) являются: чувствительность к изменениям состояния ОС, малая вычислительная трудоемкость, адаптивность к внешним воздействиям.

При достаточно общей постановке задачи речь идет о необходимости выполнения кластеризации результатов наблюдений над состоянием ОС. Совокупность наблюдений представляет собой набор измерений состояния ОС:

$$X=\{X_1, X_2,\ldots, X_N\},$$

где $X_i=\{ x_1, x_2,\ldots, x_K\}$, N – количество выборок, k – объём выборки, x_i – значение измеряемого параметра ОС. Рассматриваются три типа кластеров, каждый из которых содержит значения x_i, которые могут быть классифицированы как: нормальное, предкритическое, критическое. Вопрос заключается в том, к какому кластеру следует отнести значения рассматриваемой выборки X_i , сравнивая её с выборкой X_j на основе оценки информационной меры Кульбака.

Информационная мера Кульбака (расстояние Кульбака) распределения X_i относительно X_j рассчитывается в соответствии с правилом (1)

$$R(X_i \| X_j) = \sum_{l=1}^{K} p_l \log \frac{p_l}{q_l}, \qquad (1)$$

где p_l и q_l вероятности попадания значений выборок X_i и X_j соотвественно в l-интервал.

Образование новых кластеров в процессе обработки статистической информации показано на рис. 1, где *R1-j* – расстояние Кульбака между первым и *j*-ми кластерами. Для иллюстрации слияния и расщепления кластеров условимся, что в результате динамической кластеризации на основе информационной меры Кульбака построены 4 кластера.

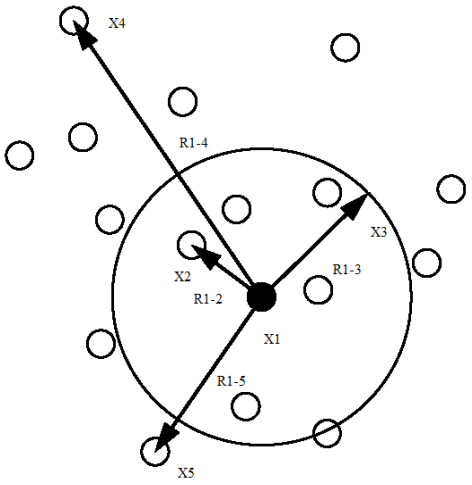

Рис. 1. Формирование новых кластеров на основе информационной меры Кульбака

Тогда расстояния Кульбака между ними можно представить в виде полного графа (рис. 2).

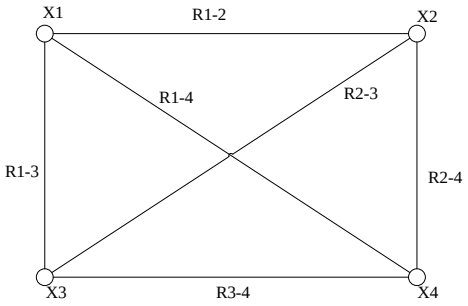

Рис. 2. Полный граф, характеризующий расстояния Кульбака между рассматриваемыми кластерами

В случае если расстояние *Ri-j*, определяемое информационной мерой Кульбака между рассматриваемыми кластерами X_i и X_j, будет больше или меньше некоторого заранее заданного порога $R_{max}(R_{min})$, то необходимо произвести слияние или расщепление данных кластеров.

Как известно, полный граф для n вершин имеет $n(n-1)/2$ рёбер. Рассмотрим случай наличия шести кластеров. Тогда количество рёбер (расстояний Кульбака между кластерами) в полном графе будет равно пятнадцати.

В табл. 1 представлены значения расстояний Кульбака *Ri-j* для данного случая, полученные в результате моделирования.

Отсортировав значения *Ri-j* по убыванию можно построить следующий график (рис. 3), показывающий мощность *R*.

Таблица 1. Результаты моделирования

Номер Кластера	Номер кластера					
	1	2	3	4	5	6
1	0					
2	25	0				
3	24	25	0			
4	14	8	24	0		
5	12	7	16	21	0	
6	17	6	7	6	7	0

Как видно из рис. 3 есть две группы кластеров, выделенные зелёным и розовым цветом с большим расстоянием от соседних и с малым соответственно. Очевидно, что группу кластеров, выделенных красным цветом целесообразно объединить в один увеличенный кластер.

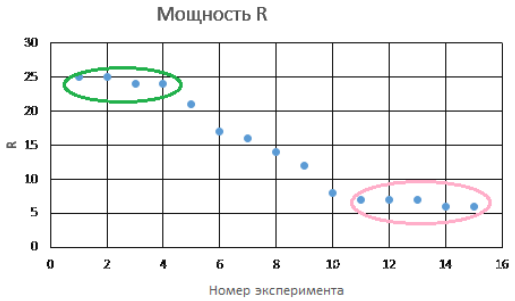

Рис. 3. Распределение мощности R

Для определения целесообразности слияния кластеров рассмотрим график расстояния Кульбака $R_{6\text{-}i}$, где $i = \overline{1,(n-1)}$.

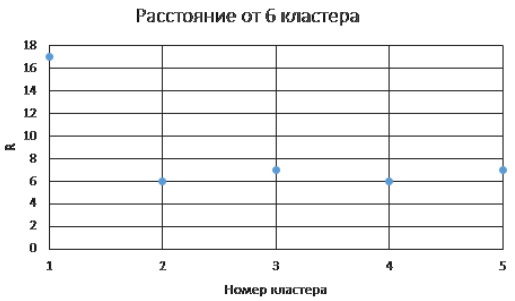

Рис. 4. Расстояния Кульбака $R_{6\text{-}i}$

Из рис. 4 становится очевидным слияние шестого кластера со вторым и четвёртым, а в случае увеличения R_{min} то и с третьим и пятым кластером.

Аналогичным образом рассматривается ситуация расщепления кластеров.

Для оценки однородности двух случайных выборок мы используем понятие дивергенции Кульбака-Лейблера (*J*-эффекта). Расчет дивергенции Кульбака-Лейблера выполнялся в соответствии с формулой [12, 13]:

$$J = \sum_{i=1}^{k} \left(\frac{f_i}{m_1} - \frac{q_i}{m_2} \right) \ln \frac{f_i m_2}{q_i m_1}, \qquad (2)$$

где *fi*, *qi* – числа попадания в *i*-й интервал примеров сравниваемых выборок, *m*1, *m*2 – численность обеих выборок.

С учетом изложенного, общий алгоритм содержит следующую последовательность действий:

1) организуется мониторинг за совокупностью исследуемых объектов путем фиксации признаков этих объектов во временных окнах; объекты, зафиксированные в первом временном окне кластеризуются с использованием *J*-метрики. При этом начальное число кластеров подбирается итеративно или задается из каких-либо априорных соображений;

2) во втором и последующих временных окнах осуществляется проверка динамических изменений кластерной структуры (выявление новых кластеров, сливающихся, расщепляющихся, исчезающих кластеров, а также смещение центров существующих кластеров);

3) при необходимости проводится коррекция структуры кластеров в соответствии с выявленными изменениями;

4) после выявления изменений и соответствующей коррекции состояния текущей кластерной структуры может выполняться прогнозирование состояния экосистемы.

Заключение. Разработка интеллектуальной технологии обнаружения аномалий экосистем акватории, базирующейся на применении новых подходов и методов, приведет к повышению обоснованности, достоверности и оперативности процессов поддержки принятия решений о наличии аномальных значений с учетом ошибок первого и второго рода.

Адаптивные методы принятия решений в условиях неопределенности позволят устранить недостатки и ограничения свойственные классическим подходам в случае зашумленных данных и неполной информации. На основе технологии больших данных и специального моделирующего стенда повышается качество оценивания принимаемых решений.

Практическая значимость результатов работы приведёт к снижению уровня негативного воздействия природных и антропогенных факторов на состояние экосистем акватории.

Исследование выполнено при финансовой поддержке РФФИ и города Севастополь в рамках научных проектов № 18-47-920005\19, № 18-47-920007\19.

СПИСОК ЛИТЕРАТУРЫ

1. *Об утверждении* Концепции построения и развития АПК «Безопасный город», утв. Распоряжением Правительства Российской Федерации 03.12.2014, №2446-р. [Электронный ресурс] – Режим доступа: http://14.mchs.gov.ru/document /2632303 (дата обращения: 12.01.2018).

2. *Современное* состояние загрязнения вод Черного моря / под ред. А.И. Симонова, А.И. Рябинина // Гидрометеорология и гидрохимия морей. Т. IV. Черное море. Вып. 3. Севастополь: ЭКОСИ-Гидрофизика, 1996. 230 с.

3. *Основные* источники загрязнения морской среды Севастопольского региона / Е.И. Овсяный, А.С. Романов, Р.Я. Миньковская [и др.] // Экологическая безопасность прибрежной и шельфовой зон и комплексное использование ресурсов шельфа. Севастополь: ЭКОСИ-Гидрофизика, 2001. С. 138–152.

4. *Скатков А.В., Брюховецкий А.А., Моисеев Д.В.* Методология организации мониторинговых процессов при решении крупномасштабных задач в облачных вычислительных средах // Информационные технологии и информационная безопасность в науке, технике и образовании "ИНФОТЕХ - 2017": сборник статей Всерос. науч.-техн. конф. Севастопольский государственный университет, Институт «Информационные технологии и управление в технических системах». Севастополь: СевГУ, 2017. С. 78–80.

5. *Бондур В.Г.* Аэрокосмический мониторинг объектов нефтегазового комплекса. М.: Научный мир, 2012. 558 с.

6. Достоверный и правдоподобный вывод в интеллектуальных системах / В.Н. Вагин, Е.Ю. Головина, А.А. Загорянская [и др.]. М.: Физматлит, 2008. 712 с.

7. *Анализ* данных и процессов / А.А. Барсегян, И.И. Холод, М.Д. Тесс [и др.]. СПб: БХВ-Петербург, 2009. 512 с.

8. *Барсегян и др.* Методы и модели анализа данных: OLAP и Data Mining. СПб., 2004. 478 с.

9. *Ганти В., Герке Й., Рамакришнан Р.* Добыча данных в сверхбольших базах данных // Открытые системы, 1999. № 9–10.

10. *Гимаров В.А., Дли М.И., Круглов В.В.* Временная изменчивость образов // Вестник МЭИ. 2003. № 2. С. 91-95.

11. *Скатков А.В., Шишкин Ю.Е.* Модель обнаружения аномалий в наблюдениях параметров полей окружающей среды с использованием систем мониторинга // Системы контроля окружающей среды. Севастополь: ИПТС. 2017. № 10 (30). С. 48–53.

12. *Кульбак С.* Теория информации и статистика. М.: Наука, 1967. 408 с.

13. *Hurvich, C.M., & Tsai, C.L.* Regression and time series model selection in small samples. Biometrika. 1989. Vol. 76(2), 297–307. DOI:10.1093/biomet/76.2.297.

MEASURE OF KULBAK IN PROBLEMS OF DYNAMIC CLUSTERING OF OBSERVATIONS OF THE ENVIRONMENTAL STATE

A.V. Skatkov, A.A. Bryukhovetskiy, D.V. Moiseev, Iu.E. Shishkin

[1]Sevastopol State University, RF, Sevastopol, Universitetskaya St., 33
[2]Institute of Natural and Technical Systems, RF, Sevastopol, Lenin St., 28

This work is devoted to the study of methods for detecting ecosystem anomalies in the water area on the basis of data clustering. Typical operations on clusters and basic metrics based on the Kullback information measure are considered, as well as some important applications - prediction, recognition of anomalies. The paper proposes a new scientific approach, which to a large extent overcomes the requirements for the implementation of well-known methods for monitoring the state of ecosystems.
Keywords: analysis, anomalies, big data, clustering, Kullback measure, monitoring, environment, forecasting, ecosystems.