ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА АДАПТИВНОГО ВЫБОРА СЦЕНАРИЕВ ПАРАМЕТРИЧЕСКОГО ОБНАРУЖЕНИЯ ДИВЕРГЕНЦИЙ ДАННЫХ МОНИТОРИНГА

Ю.Е. Шишкин^{1,2}, А.В. Скатков¹

¹ Севастопольский государственный университет,
РФ, г. Севастополь, ул. Университетская, 33

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

Предложена интеллектуальная система поддержки принятия решений по выбору статистического критерия различия при анализе данных мониторинга сложных объектов и сред, учитывающая мощность и чувствительность критерия, наличие рисков. Представлена схема выбора параметрического критерия различия в зависимости от объема и количества выборок. Результирующая система направлена на повышение обоснованности при выборе критерия различия в процессе статистического анализа данных мониторинга и позволяет лицу, принимающему решение, выбирать оптимальный сценарий анализа данных с учетом допустимых для себя возможных рисков и потерь.

Ключевые слова: интеллектуализация, мониторинг среды, параметрические критерии, математическое моделирование, сложные системы, интеллектуальный анализ данных, поддержка принятия решений.

Поступила в редакцию: 06.05.2019.

Введение. На сегодняшний день для решения задач статистического анализа данных мониторинга сложных объектов и сред, сравнения репрезентативных выборок измерений, в частности данных первичных измерителей физических показателей окружающей среды, разработан и успешно применяется обширный математический аппарат параметрических или непараметрических критериев обнаружения различия [1–3]. При этом большинство критериев обладают широким спектром приложения несмотря на специфичность изучаемых объектов, а эффективность того или иного критерия зависит от характеристик исходных данных наблюдений, в частности их закона распределения, дисперсионности и объема [4–6]. Если рассматривать эффективность критерия обнаружения различия как совокупную характеристику, учитывающую его мощность и чувствительность, вероятности ошибок первого и второго рода, при анализе конкретной выборки данных наблюдений, а также возникающих при принятии решений на его основе рисков, то вопрос выбора наилучшего критерия, в этом контексте, остается открытым [7, 8].

Постановка задачи. В рамках статьи требуется разработать интеллектуальную информационную технологию, осуществляющую выбор оптимального параметрического критерия обнаружения различия данных мониторинга на основе автоматического выбора сценариев. В рамках информационной технологии рассматриваются три базовых сценария принятия решений в условиях неопределенности, в частности, минимаксный критерий, прямое голосование и использование весовых коэффициентов. В ходе целенаправленного вычислительного эксперимента выбор сценариев осуществляется на основе критериев вероятности ошибок первого и второго рода, размеров зон уверенного и неуверенного распознавания.

Материалы и методы. Основой интеллектуализации системы поддержки принятия решений по выбору критериев различия выборок данных мониторинга сложных объектов и сред является ввод точек принятия решений, дающих возможность лицу, принимающему решения (ЛПР), определять приоритеты: чувствительность или вероятность ложного срабатывания, в случае, когда Парето-

оптимальное решение отсутствует [9, 10]. Эти точки помещаются в обучаемую интеллектуальную технологию, где в соответствии с принятыми пользователями решениями, осуществляется выбор весовых коэффициентов по критериям мощности или чувствительности. По существу, это представляет собой элементы искусственного интеллекта с автоматизацией принятия решения, что является конечной задачей многих проблем искусственного интеллекта [11, 12].

В рамках работы рассматриваются параметрические критерии, поскольку они обладают большей статистической мощностью и меньшей вероятностью совершить ошибку второго рода в сравнении с непараметрическими. Это означает, что в случае существования какой-либо закономерности между выборками с наибольшей вероятностью эта закономерность будет выявлена с помощью параметрических критериев. К числу параметрических критериев относят критерии: Стьюдента, Фишера, Кохрена, Бартлетта, хи-квадрат и др.

Следует отметить, что параметрические методы можно применять только при одновременном соблюдении двух условий: количественном характере вариации и предположении, что данные исходной выборки взяты из генеральной совокупности, имеющей конкретное распределение, как правило, нормальное. Однако, как показывает практика, часть данных, получаемых в результате мониторинга окружающей среды, не распределена нормально, поэтому применение параметрических критериев при анализе результатов таких исследований может привести к ошибкам в статистических выводах. В таком случае непараметрические критерии могут стать более мощными, т.е. способными с большей достоверностью отвергать нулевую гипотезу [13].

Каждый из критериев различия имеет свою специфику, различаясь между собой по накладываемым ограничениям на входные наборы данных: тип измерительной шкалы, максимальный объем выборки, количество выборок, качество выборок. Разнообразие критериев обнаружения различия позволяет:

- выбирать критерий, адекватный типу шкалы, в которой получены экспериментальные данные;

- производить сравнение данных, характеризующихся связанными и несвязанными выборками;

- производить сравнение данных с неравными по объему выборками;

- в зависимости от целей исследования выбирать подходящие по мощности критерии.

Математическая модель интеллектуальной информационной системы. В общем виде модель интеллектуальной системы обнаружения дивергенции данных мониторинга сложных объектов и сред, при условии $\forall C \in \Omega$, может быть представлена выражением:

$$M : C_i, S(G, m), \alpha, F(L) \rightarrow R, \quad (1)$$

где Ω – весь допустимый набор критериев различия, C_i – i -й критерий различия, $S(G, N)$ – выборка из генеральной совокупности G объемом N , α – допустимая для данной задачи вероятность ошибки первого рода (на практике часто полагают $\alpha = 0,05$), $R = \langle \text{true}, \text{false} \rangle$ – бинарный результат применения критерия различия C , $F(L)$ – оператор выбора алгоритма.

Выбор алгоритма $F(L)$ осуществляется на основе подхода информационных сценариев и описывается выражением:

$$F : L \langle SR, O(S), W, N \rangle \rightarrow A, \quad (2)$$

где $L = \langle SR, S, W \rangle$ – приоритеты, заданные пользователем, SR – выбор приоритета чувствительности или надежности, O – оператор выбора сценариев, S – список сценариев использования информационной системы, W – весовые коэффициенты критериев различия, N – адаптивное правило, A – выбранный алгоритм. Таким образом, интеллектуальное правило выбора приоритетов $F(L)$ осуществляет работу с учетом соотношения априорной и апостериорной информации. В настоящее время технологии сценариев активно используются для решения задач в условиях высокого дефицита априорной информации и дает возможность компенсировать отсутствие первоначальных знаний у персонала,

предлагая ряд альтернативных типовых стандартных решений, где каждому такому решению соответствует некоторая алгоритмическая схема.

Результатом применения критерия различия C для выборок $S_1, \dots, S_m$ является численная оценка статистической однородности выборок, для некоторых критериев (Стьюдента, Фишера, хи-квадрат) существует ограничение $m=2$. Для удобства i -ю выборку $S(G, N)$ далее будем обозначать S_i .

При исследовании мощности критериев C предлагается сформулировать и экспериментально проверить гипотезу H_0 – сравниваемые выборки однородны и извлечены из одной и той же генеральной совокупности. Таким образом, основная гипотеза состоит в том, что для любых выборок S_i , их функции распределения равны между собой $F(S_a) = F(S_b) \mid a \neq b$. В случае принятия гипотезы H_0 следует вывод об отсутствии изменений в состоянии объекта

мониторинга или о незначительности этих изменений.

Конкурирующая гипотеза H_1 – выборки $S_a(G_1, N), S_b(G_2, N)$ извлечены из разных генеральных совокупностей, описываемых различными законами распределения $F(G_1) \neq F(G_2)$. Конкурирующая гипотеза H_1 принимается в случае потенциально опасного изменения параметров, характеризующих состояние сложного объекта или среды. В рассматриваемом случае H_0 и H_1 – простые параметрические гипотезы, т.е. полностью определяют все множество параметров распределения случайной величины.

На рис. 1 представлена структура интеллектуальной системы, которая в зависимости от характеристик выборок $S(G, N)$ позволяет выбрать подмножество критериев различия $C \in \Omega$, применение которых будет наиболее целесообразно с имеющимся входным набором данных.

Рис. 1. Структура интеллектуальной системы выбора статистического критерия обнаружения дивергенции

Функция блока оценки качества мониторинга – подсчет частоты использования сценариев, отрицательных и положительных эффектов их использования. При адаптации выбора сценария изначально выбор равновероятностный, затем предпочтение отдается тому сценарию, который дает наибольший относительный эффект. Подсчитываются веса, критерии эффективности, и выбирается тот сценарий, который дает положительный результат чаще.

Для того чтобы отсеять заведомо бесперспективные варианты S для каждого конкретного случая следует учитывать объемы выборок, число выборок (две или более), равные или не равные они по объему, подчинены ли известному закону распределения. Выбор конкретной реализации S недетерминирован и осуществляется в соответствии с предпочтениями пользователя $L = \langle SR, S, W \rangle$ и интеллектуальной системы M .

В качестве критериев для проверки отсутствия дивергенции по средним значениям можно использовать критерии Фишера, Кохрена и Бартлетта. Для проверки отсутствия дивергенции по дисперсии можно использовать критерии Стьюдента, Фишера и приближенный критерий.

Известно, что в теории принятия решений наибольшую сложность вызывает формирование оценочной матрицы, процесс которого до сих пор до конца не формализован и не имеет конкретного алгоритма решения. Поэтому в работе предлагается осуществлять формирование оценочной матрицы на основе технологии информационных сценариев, которые сейчас достаточно успешно применяются в разных областях [14, 15]. У ЛПР имеется некоторая свобода в выборе сценариев, а в рамках сценария эта свобода ограничена, это ограничение можно снять путем расширения списка сценариев. Облегчить процедуру выбора сценариев можно путем введения интеллектуальных решений на основе технологий искусственного интеллекта и механизмов адаптации. Механизм адаптации необходим для компенсации недостаточности априорной информации.

Для иллюстрации этого подхода в данной статье рассматриваются следующие сценарии: минимаксный подход с относительными оценками, прямое голосование и минимизация рисков.

В интеллектуальной системе M принятие решений осуществляется на основе оценочного функционала или управляющей матрицы $P = P(f_{r,j})$, характеризующей выигрыш или проигрыш при принятии решения $r \in R$ если система или среда находится в состоянии j . Для формирования оценочной матрицы в качестве оценочной характеристики ЛПР принимает решение использовать один из перечисленных сценариев.

Сценарий 1. Минимаксный подход с относительными оценками. Элементы оценочной матрицы P_1 задаются как отношение эмпирического значения критерия к его критическому значению $P_1 = C_{emp} / C_{crit}$ тогда если $P_1 < 1$, то наблюдаемое значение критерия принадлежит области принятия гипотезы и гипотеза согласуется с экспериментальными данными, в противном случае будем считать, что наблюдаемое значение критерия принадлежит критической области и нулевую гипотезу отвергают. Применяется минимаксных критерий, в соответствии с которым система M решает задачу поиска такого $S \in \Omega$, который обеспечивает $\min(P_1)$.

Сценарий 2. Прямое голосование. Все критерии различия имеют одинаковые единичные веса $P(f_{r,j}) = 1 \forall j \in J$, в таком случае система M при определении $r \in R$ руководствуется принципом прямого голосования, когда каждый критерий $S \in \Omega$ «голосует» за определенную реализацию R , а система M сообщает ЛПР результат голосования.

Сценарий 3. Минимизация рисков. В качестве оценочной характеристики используется численное выражение степени риска – вероятность ошибки второго рода β (что будет принята неправильная нулевая гипотеза), умноженная на потери, заданные ЛПР. Система M будет находить такое решение, которое

достигается при минимально возможном риске.

Каждый из сценариев имеет свои преимущества и недостатки, которые проявляются или не проявляются в зависимости от конкретной ситуации, заранее определить эффективность какого-либо сценария формальными методами не представляется возможным.

Заключение. Предлагаемая интеллектуальная информационная система адаптивного выбора сценариев параметрического оценивания и обнаружения дивергенции данных мониторинга сложных объектов и сред обладает свойствами обучения и адаптации, в зависимости от эффективности применяемых сценариев, и направлена на повышение обоснованности при выборе критерия различия, что позволяет ЛПП выбирать оптимальный сценарий анализа данных с учетом допустимых рисков и потерь.

Вводится понятие дивергенции данных как совокупности количественных оценок их аномальности, анормальности, сдвигов центров распределений, неоднородности дисперсий, эффектов значительной асимметричности. Дивергенция рассматривается как комплексная оценка, которая в целом определяет качество имеющихся данных с точки зрения наличия в них отмеченных эффектов. В англоязычной литературе этот термин *data divergence* встречается, однако он больше соответствует описанию ошибки вследствие устаревания данных, ошибки отклика, кодирования или ввода данных, различия в кодировании и пр.

Данная система сценариев является открытой для пополнения другими параметрическими методами выявления различий и эвристическими подходами, позволяющими ЛПП более обоснованно осуществлять выбор информационного сценария.

Выбор конкретного сценария изначально осуществляется в соответствии с предпочтениями ЛПП, а затем, по мере накопления опыта использования в данной проблемной области и с учетом частоты успешных и неуспешных использований сценариев, можно построить интеллектуальный регулятор с адаптивным управлением по автоматизации выбора.

Как направление для дальнейших исследований стоит отметить, что параметрические методы эффективны для выборки большой размерности с хорошей информационной поддержкой, обладают большей статистической мощностью и меньшей вероятностью совершить ошибку второго рода в сравнении с непараметрическими. Однако как показывает практика, часть данных, получаемых в результате мониторинга окружающей среды, не распределена нормально, поэтому применение параметрических критериев при анализе результатов таких исследований может привести к ошибкам в статистических выводах, что ограничивает область применения такой системы. Особенно это актуально в случае, когда наблюдений мало, выборки небольшого объема вследствие высокой стоимости их получения. В таком случае рациональным будет дополнить представленную в работе интеллектуальную информационную систему классом непараметрических критериев выявления различий, что значительно расширит область ее применения в контексте обнаружения дивергенции данных мониторинга сложных объектов и сред.

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

СПИСОК ЛИТЕРАТУРЫ

1. Чимитова Е.В., Ведерникова М.А., Галанова Н.С. Непараметрические критерии согласия в задачах проверки адекватности моделей надежности по цензурированным данным // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 4 (25). С. 115–124.
2. Кудлаев Э.М., Орлов А.И. Вероятностно-статистические методы исследования в работах А.Н. Колмогорова // Заводская лаборатория. Диагностика материалов. 2003. Т. 69. № 5. С. 55–61.
3. 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.

4. Гржибовский А.М. Анализ трех и более независимых групп количественных данных // *Экология человека*. 2008. № 3. С. 50–58.

5. Новиков Д.А. Статистические методы в педагогических исследованиях (типовые случаи). М.: МЗ-Пресс, 2004. 67 с.

6. Шишкин Ю.Е., Скатков А.В. Метрики качества для оценки и прогнозирования критических состояний // *Качество и жизнь*. 2019. № 1 (21). С. 61–66.

7. Ларичев О.И. Свойства методов принятия решений в многокритериальных задачах индивидуального выбора // *Автоматика и телемеханика*. 2002. № 2. С. 146–158.

8. Шишкин Ю.Е., Скатков А.В. Повышение достоверности оценок рисков в процессах мониторинга при распределениях общего вида // *Системы контроля окружающей среды*. 2019. № 1 (35). С. 41–47. DOI: 10.33075/2220-5861-2019-1-41-47

9. Подиновский В.В., Ногин В.Д. Парето-оптимальные решения многокритериальных задач. М., 2007. 256 с.

10. Гришко А.К., Кочегаров И.И., Лысенко А.В. Многокритериальный выбор оптимального варианта сложной технической системы на основе интер-

вального анализа слабоструктурированной информации // *Измерение. Мониторинг. Управление. Контроль*. 2017. № 3 (21). С. 97–107.

11. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Информатика в техническом университете. М., 2005. 302 с.

12. Скатков А.В., Брюховецкий А.А., Шишкин Ю.Е. Разработка интеллектуальной технологии обнаружения аномалий экосистем акватории г. Севастополя // *Системы контроля окружающей среды*. 2019. № 1 (35). С. 27–34. DOI: 10.33075/2220-5861-2019-1-27-34

13. Лемешко Б.Ю., Лемешко С.Б. Об устойчивости и мощности критериев проверки однородности средних // *Измерительная техника*. 2008. № 9. С. 23–28.

14. Архипова Н.И., Кононов Д.А., Кульба В.В. Проблема выбора сценария управления мониторингом безопасного функционирования сложных систем // *Проблемы управления безопасностью сложных систем: труды XVII Междунар. конф. Российская академия наук / под редакцией Н.И. Архиповой, В.В. Кульбы*. 2009. С. 148–151.

15. Массель Л.В., Гальперов В.И. Разработка многоагентных систем распределенного решения энергетических задач с использованием агентных сценариев // *Известия Томского политехнического университета. Инжиниринг георесурсов*. 2015. Т. 326. № 5. С. 45–53.

INTELLIGENT SYSTEM FOR ADAPTIVE SELECTION OF SCENARIOS FOR PARAMETRIC MONITORING DATA DIVERGENCE DETECTION

Y.E. Shishkin^{1,2}, A.V. Skatkov¹

¹ Sevastopol State University, RF, Sevastopol, Universitetskaya St., 33

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

An intelligent decision support system for the selection of a statistical criterion for differences in the analysis of monitoring data of complex objects and environments, taking into account the power and sensitivity of the criterion, the presence of risks, is proposed. A scheme for selecting the parametric difference criterion depending on the size and number of samples is presented. The resulting system is aimed at increasing the validity of the choice of the criterion for differences in the statistical analysis of monitoring data and allows the decision maker to choose the optimal scenario for analyzing data, taking into account the possible risks and losses that are permissible for themselves.

Keywords: intellectualization, environment monitoring, parametric criteria, mathematical modeling, complex systems, data mining, decision support.