УДК 004.62:004.942
DOI https://doi.org/10.32838/2663-5941/2019.5-1/17

Крісілов В.А.
Одеський національний політехнічний університет

Комлева Н.О.
Одеський національний політехнічний університет

Бурдейний Є.І.
Одеський національний політехнічний університет

ПРОГРАМНА СИСТЕМА АНАЛІЗУ ЯКОСТІ ДЖЕРЕЛ МЕДИЧНОЇ СТАТИСТИЧНОЇ ІНФОРМАЦІЇ

Сфера застосування статистичних методів завдяки їхній потужній теоретичній базі та практичному інструментарію поширюється на багато видів наук. Впровадження статистики в медичні дослідження супроводжується низкою труднощів, зумовлених винятковою складністю, мінливістю і наявністю індивідуальних особливостей досліджуваних об'єктів. При цьому необхідно забезпечувати певну якість даних, що описують стан досліджуваного об'єкта. Метою роботи є підвищення якості обробки медичної статистичної інформації завдяки попередньому аналізу якості даних, що надають джерела інформації. Порівняно з відомими роботами, в яких акценти робилися на виборі методів аналізу стану досліджуваного об'єкта на базі вхідної інформації, запропонований у роботі підхід дає змогу виконати попередній аналіз якості вхідної інформації для оцінювання можливості її подальшого використання. Для досягнення мети було наведено в загальному вигляді роботу системи обробки та аналізу даних з урахуванням перевірки якості вхідних даних. У перелік аномалій даних включено перевірки на невідповідність за типами даних, за діапазонами припустимих значень, за кореляцією значень у пов'язаних між собою залежностях даних. Як практичний приклад виконано проєктування програмної системи, призначеної для аналізу якості медичної статистичної інформації щодо ситуації, пов'язаної з туберкульозом. Інформацію для аналізу, що містить більш ніж сотню аналізованих ознак, взято на сайті Міністерства охорони здоров'я України. Програмна система дає змогу виконувати базові операції щодо аналізу даних, визначати та у разі можливості коригувати знайдені аномалії. Розроблювана система реалізує клієнт-серверну архітектуру. Доступ до розширеного функціоналу, що містить роботу з аномаліями та отримання прогностичних значень аналізованих ознак як даних часового ряду, має аналітик. Деталізовано інтерфейс системи, обрано мову програмування, супутні менеджери та фреймворки. Використання програмної системи дасть змогу забезпечити потрібну якість даних, що надають джерела медичної статистичної інформації.

Ключові слова: *якість інформації, аналіз даних, джерело інформації, аномалії вхідних даних, програмна система.*

Постановка проблеми. Сучасні тенденції застосування математико-статистичних методів у медичних дослідженнях показали можливості їх широкого використання. Під час інформаційного планування та практичної організації досліджень у медичній галузі, крім визначення і формалізації методів дослідження, потрібно звернути пильну увагу на відбір і забезпечення якості реєстрованих ознак. Окремі спостереження реєструються згідно з цілями дослідження відповідно до виділених ознак. Ці ознаки повинні бути істотними і релевантними меті дослідження, кількість ознак має бути мінімальною для досягнення поставленої перед дослідником мети. Важливою властивістю таких ознак є можливість їх комбінування з метою доповнення та взаємного контролю.

Необхідно пред'являти серйозні вимоги до якості таких джерел даних, що дають змогу визначати реєстровані ознаки. Від якості джерел медичної інформації, визначеної з використанням набору ознак якості, суттєво залежать ефективність функціонування систем обробки та аналізу даних і задоволення визначених відповідно до їхнього призначення практичних потреб [1, с. 91; 2, с. 90].

Аналіз останніх досліджень і публікацій. Недотримання об'єктивних вимог, що пред'являються до вхідної інформації, обмежує сферу використання результатів рішення завдання або ж робить їх зовсім непридатними. У дослідженні [3, с. 2] показано, що наявність пропусків у вихідних даних призводить до зниження якості

вхідної інформації і, як наслідок, до зниження якості роботи системи, що її використовує. Іншою причиною зниження якості може стати суперечливість вихідних даних [4, с. 270; 5, с. 62]. Крім того, слід відстежувати ситуації невідповідності форматів даних, помилок введення даних, дублювання вхідної інформації та інше [6, с. 2397].

Чимало систем обробки даних використовують як вхідну інформацію часові ряди, які містять вимірювання, проведені в певні (впорядковані) моменти часу. На відміну від використання випадкових вибірок, під час роботи з такими даними можна виконати формальний опис моделі часового ряду і використовувати для вирішення поставленого завдання один або кілька методів математичної статистики [7, с. 140; 8, с. 80]. Зазвичай це не становить труднощів, однак під час переходу до реальних завдань великої розмірності доцільно враховувати та оцінювати зашумленість і нелінійність джерел даних [9, с. 171; 10, с. 161].

Постановка завдання. Метою статті є підвищення якості обробки медичної статистичної інформації завдяки попередньому аналізу якості даних, що надають джерела інформації.

Виклад основного матеріалу дослідження. З погляду природи властивостей об'єкта, а також характеру процедур, що дають змогу визначити їхні конкретні значення, всі властивості можна розділити на три види. Найбільш бажаними є властивості першого виду – вимірні властивості, їхні значення можуть бути отримані шляхом прямих вимірювань і спостережень. До другого виду слід віднести властивості вищого рівня спільності, які можуть бути обчислені на підставі вимірюваних властивостей. Фактично ці властивості є агрегованими згортками первинних властивостей. Третій вид властивостей – це властивості, значення яких не можуть бути безпосередньо виміряні й обчислені, або трудомісткість і складність цих процесів дуже висока. Ці властивості мають характер експертних оцінок, отриманих унаслідок тієї чи іншої процедури, що забезпечує формалізацію і об'єктивізацію «ручного» оцінювання фахівцем-експертом [11, с. 102].

Найбільш простими способами реєстрації значень ознак є спостереження і підрахунок [12, с. 65]. Однак і такий спосіб передбачає наявність певних категорій помилок, а саме:

- помилки органів спостереження;
- помилки, які надає об'єкт спостерігання;
- інструментальні помилки;
- випадкові помилки.

Розглянемо в загальному вигляді роботу системи обробки та аналізу даних з урахуванням перевірки якості вхідних даних (рис. 1). Початковим етапом є формалізація мети дослідження, що визначає кінцевий прогнозований результат роботи. Потім відповідно до обраного методу отримання вхідної інформації проводиться збір релевантних даних і перевірка їх на якість.

Кожна з реєстрованих ознак належить до певного типу даних, який визначає множину припустимих значень цієї ознаки і операцій над цими значеннями. З кожним типом даних пов'язана вимірювальна шкала (кількісна, якісна та їхні різновиди) [13, с. 13]. Невідповідність обраної шкали типу вимірюваної ознаки приводить до аномалій – помилкових значень, які не повинні використовуватись методами аналізу стану досліджуваного об'єкта.

Під час перевірки даних на якість послідовно перевіряється наявність і можливість усунення таких аномалій: за типами даних, за припустимими значеннями, за пропущеними значеннями (рис. 2). Крім того, часто в реальних системах є кореляція між значеннями декількох пов'язаних між собою залежностей даних [14, с. 17]. Контроль за кореляцією за наявності таких залежностей теж є завданням перевірки якості даних.

Після отримання репрезентативної вибірки обираються відповідні методи для аналізу даних. Якщо існують стандартні програмні засоби для автоматичного чи автоматизованого аналізу даних відповідно до мети дослідження, їх доцільно використовувати, інакше – розроблювати власні програми чи модернізувати наявні. Після отримання результату аналізується можливість його використання, у разі невдачі процес повторюється з можливістю коригування мети дослідження.

Як практичний приклад виконано проєктування програмної системи, призначеної для аналізу якості інформації щодо захворювання населення України різними формами туберкульозу за 2012–2018 роки [15, с. 253; 16, с. 48]. На сайті Міністерства охорони здоров'я України наведено дані, які включають кількість захворювань серед населення відповідно до вікових діапазонів, територіальних регіонів і соціальних структур, показники профілактичних оглядів, щеплень і хірургічних лікувань, дані щодо лабораторій, ліжкового фонду, укомплектованості лікарями, лікарняної та санаторної допомоги та інше – усього понад сотні таблиць, з яких близько третини містять пов'язані між собою залежності [17, с. 1]. На рисунку 3 наведено UML-діаграму варіантів використання

цієї системи, яка містить дві категорії акторів і формалізований функціонал.

Як можна побачити, під час перевірки даних на якість аналізуються наявність пропусків, аномальних значень і перевірка залежностей. Розроблювана система реалізує клієнт-серверну архітектуру, яка у загальному вигляді представлена на рисунку 4.

Об'єкт класу App – HTTP-сервер, що створюється найпершим у системі. Завданнями App є прийняття запиту та відправка відповіді. Для цього класом App проводиться ініціювання та реєстрація всіх необхідних компонентів фреймворку та передання запиту роутеру (Router).

Роутер є відповідальним за перевірку наявності такого endpoint-а в системі та перевірку дозволеності HTTP-методу. Також роутер застосовує всі фільтри (Middleware), які за ним закріплені. Наприклад, проводиться перевірка повноважень користувача на використання певного роутера. Запити, які не пройшли перевірки Middleware, не будуть делеговані далі до методів контролерів.

Об'єкти-контролери (DataController, AnalysisController) обробляють отримані дані та за необхідності звертаються до об'єктів, через які надається доступ до бази даних (DataAccessObject) або які вміщують у собі логіку системи (TimeSeriesAnalyser).

Дані, що зберігаються в реляційній базі даних, можна отримати, вказавши необхідні критерії (Criteria), наприклад: «Лабораторна діагностика нових випадків туберкульозу легень», «З підтвердженим мазком (M+)», «Одеська область». Отримані дані реалізують інтерфейс часового ряду

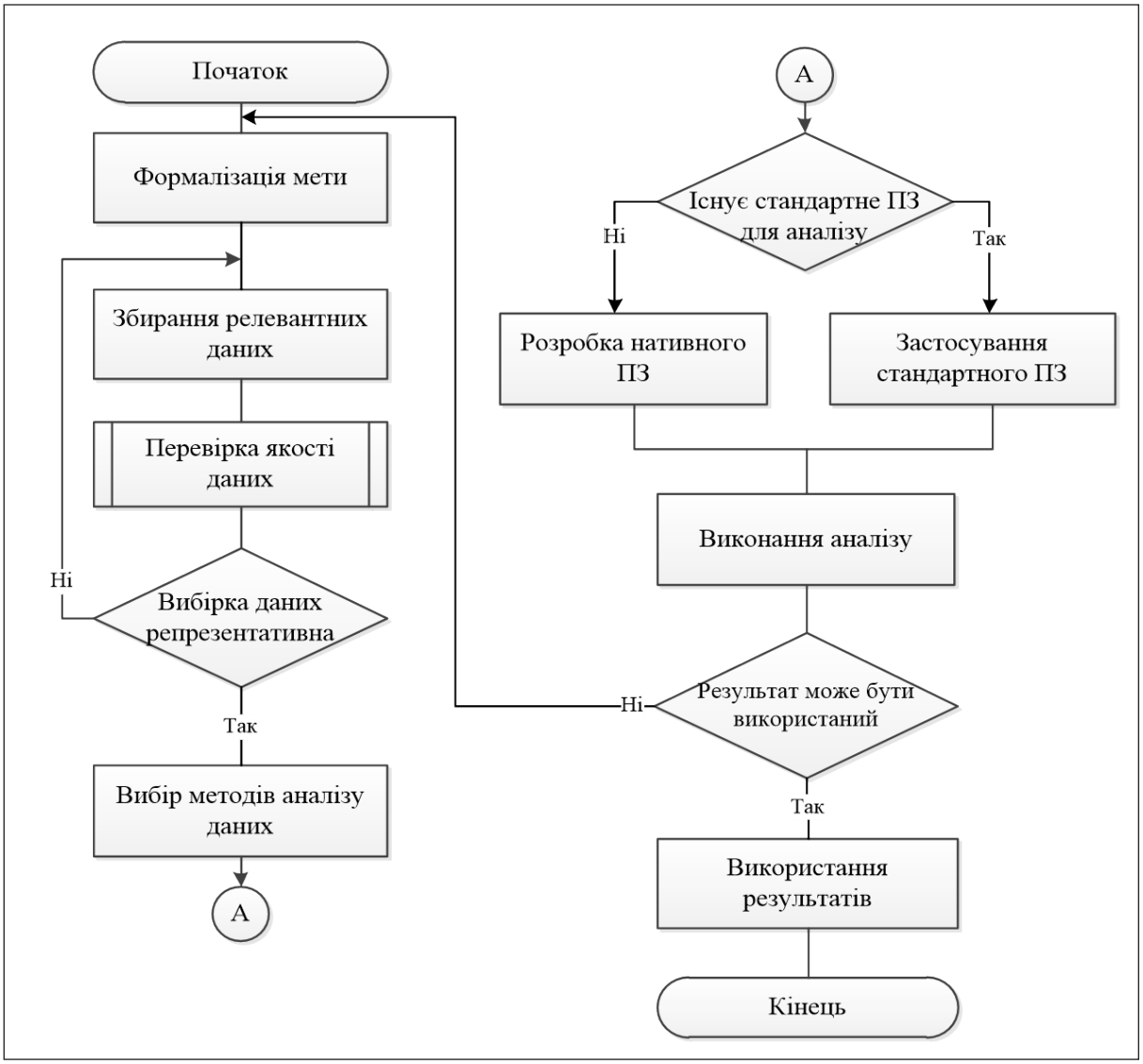

Рис. 1. Схема алгоритму роботи системи обробки та аналізу даних

(TimeSeries), який дає змогу отримати ряд точок даних, проіндексованих у хронологічному порядку (масив DataPoint), або знайти залежності від інших часових рядів (Relation).

Маючи об'єкт інтерфейсу TimeSeries, система може проаналізувати його за допомогою реалізацій інтерфейсу TimeSeriesAnalyser, а саме: передбачити наступне значення часового ряду, розрахувати

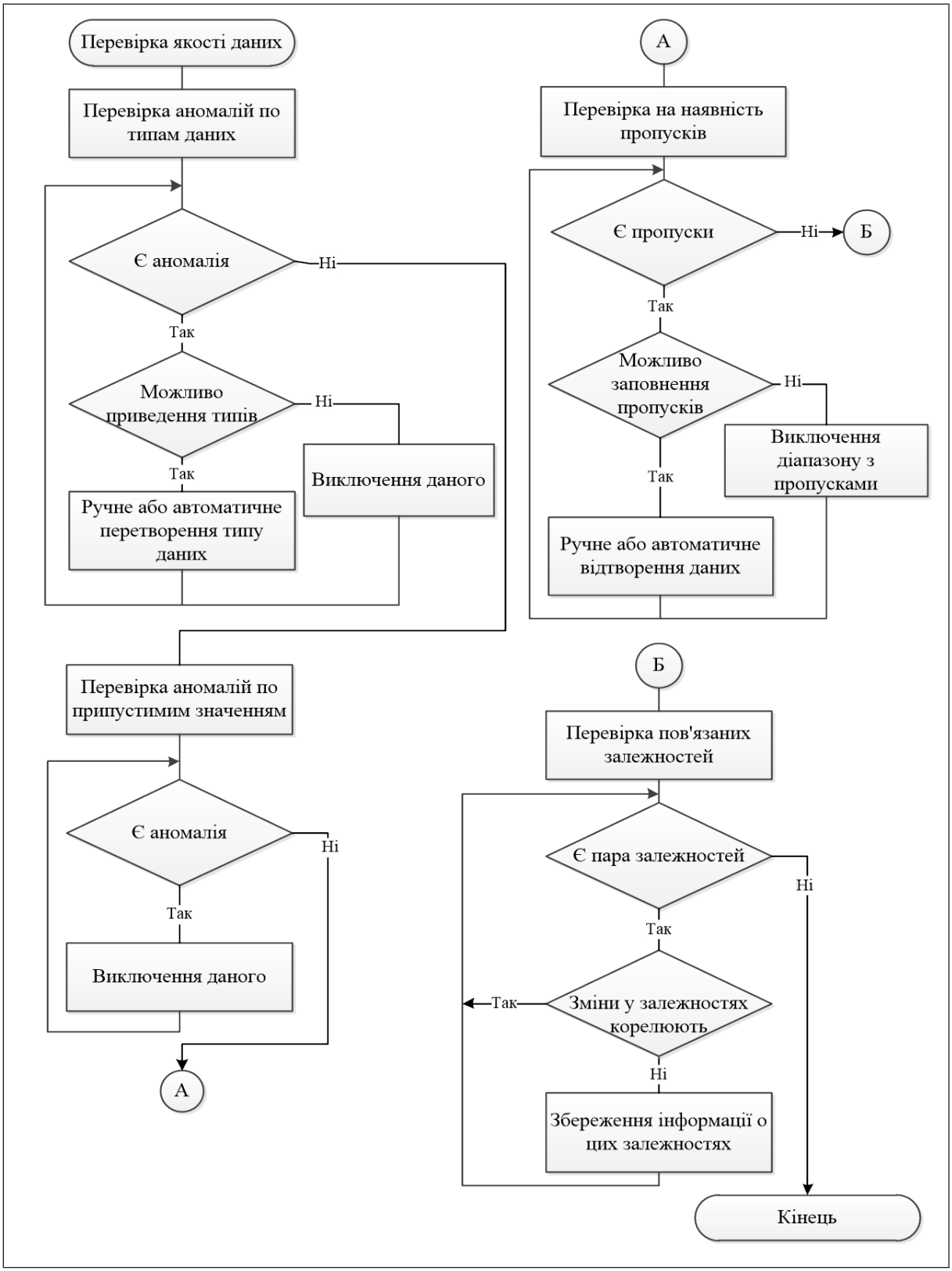

Рис. 2. Схема алгоритму підсистеми перевірки якості даних

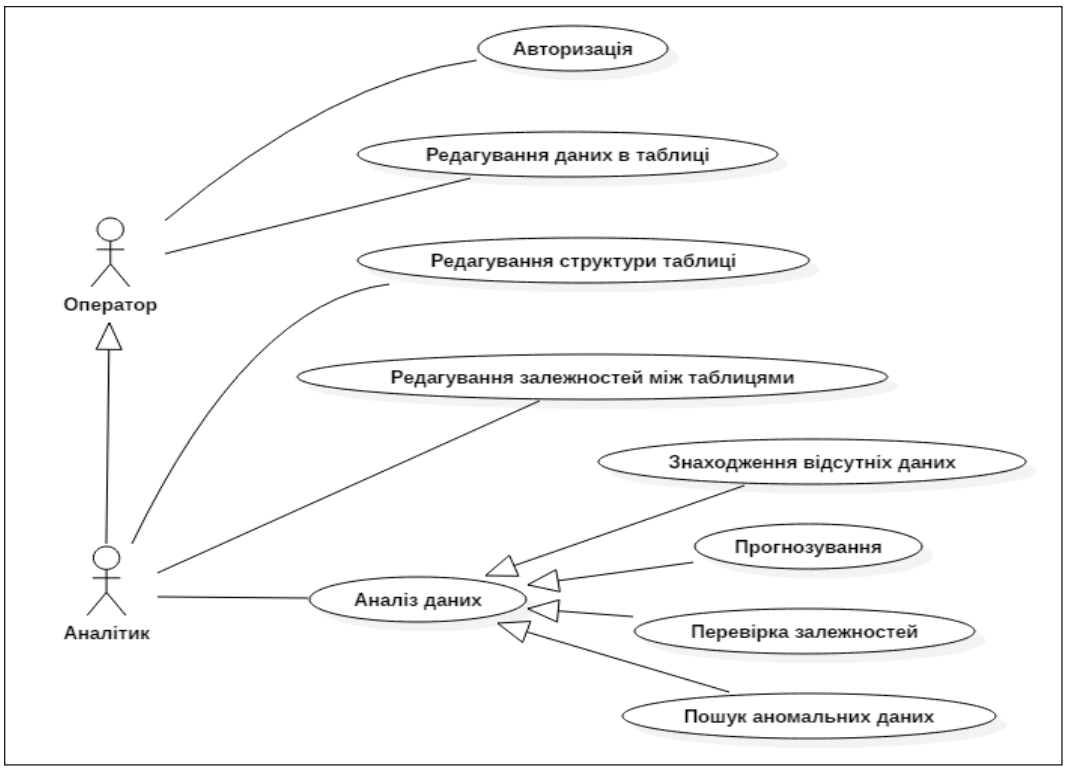

Рис. 3. Діаграма варіантів використання програмної системи

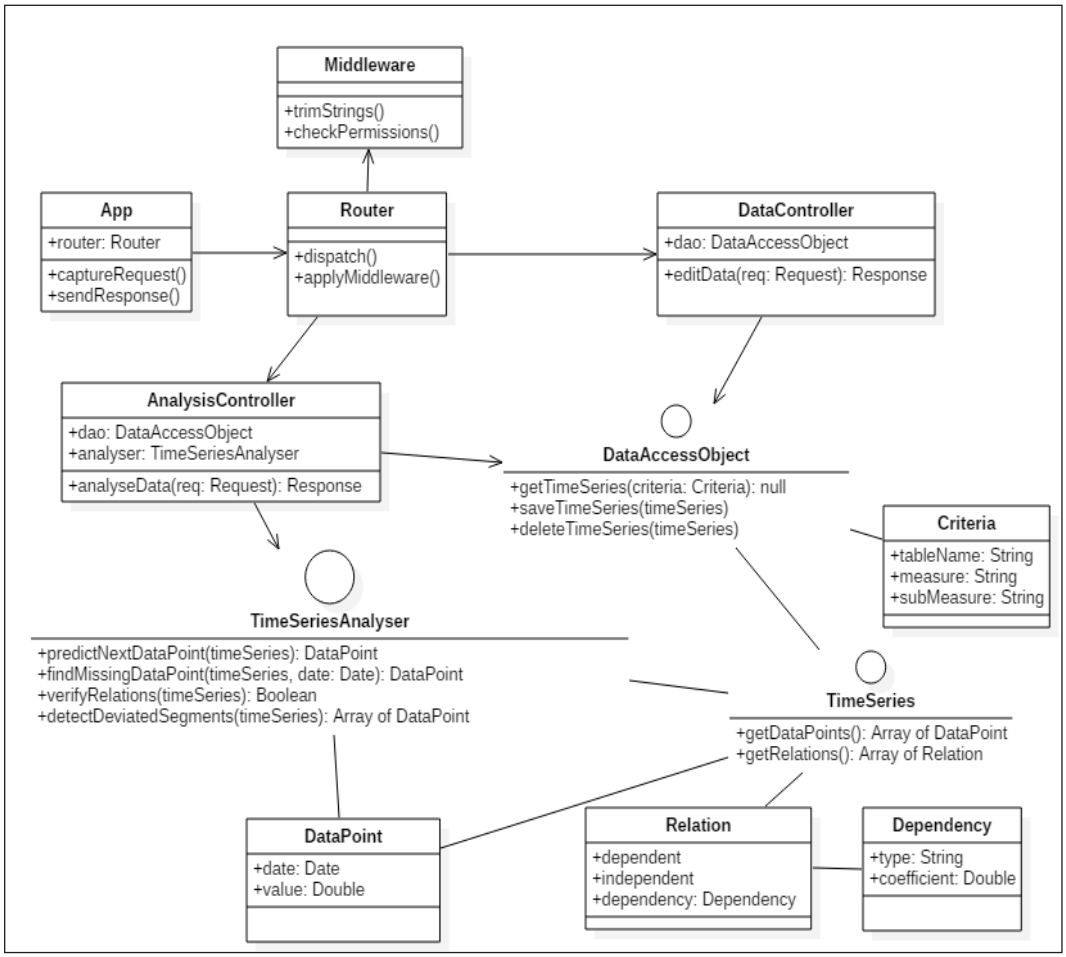

Рис. 4. Архітектура програмної системи

значення часового ряду в заданий проміжок часу, перевірити дотримання залежностей або знайти різкі коливання значень.

Під час реалізації системи була обрана мова програмування JavaScript (серверна частина виконується в середовищі Node.js). Для полегшення процесу розроблення і написання коду, що функціонує швидко, масштабується та не викликає труднощів у підтримці, було використано такі фреймворки:

– Express – бібліотека, яка дає змогу створювати HTTP-сервери;

– React.js – для створення динамічного та інтерактивного інтерфейсу користувача.

Також використовуються пакетний менеджер npm, Webpack для збірки JavaScript та CSS-файлів, TensorFlow.js – відкрита програмна бібліотека для машинного навчання. Розроблювана програмна система перебуває на стадії тестування функціоналу.

Висновки. У роботі обґрунтовано необхідність формалізації вимог до якості джерел даних. Показано в загальному вигляді роботу системи обробки та аналізу даних з урахуванням перевірки якості вхідних даних. Серед етапів перевірки якості виділено аналіз та усунення таких аномалій: за типами даних, за діапазонами припустимих значень, за пропущеними значеннями, за кореляцією значень пов'язаних залежностей даних.

Наведений підхід реалізовано під час проєктування програмної системи, яка призначена для аналізу якості інформації щодо ситуації, пов'язаної з туберкульозом з використанням багатьох аналізованих ознак. Обрано базовий функціонал системи, виконано проєктування її клієнт-серверної архітектури. Деталізовано інтерфейс системи з можливістю використання аналізованих ознак як даних часового ряду. Для реалізації програми обрано мову програмування, супутні менеджери та фреймворки.

Розроблюваний програмний продукт дасть змогу відстежувати ситуації з інформаційними аномаліями, у разі можливості усувати їх автоматичним чи автоматизованим чином і зробити процес роботи з медичними статистичними даними більш легким і наочним.

Список літератури:

1. Крисилов В.А., Комлевая Н.О. Анализ и оценка компетентности источников информации в задачах интеллектуальной обработки данных. *Международная научно-практическая конференция "Электротехнические и компьютерные системы: теория и практика" ELTECS-2019. Problemele energeticii regionale*. 2019. Вып. 1–1 (40). С. 91–104.

2. Крисилов В.А., Тарасенко Р.А. Предварительная оценка качества обучающей выборки для нейронных сетей в задачах прогнозирования временных рядов. *Труды Одесского политехнического университета.* Одесса, 2001. Вып. 1. С. 90–96.

3. Awawdeh S., Edinat A., Sleit A. Enhanced K-means Clustering Algorithm for Multi-attributes Data. *International Journal of Computer Science and Information Security (IJC-SIS)*. 2019. Vol. 17 (2). P. 1–6.

4. Rokach L. A survey of clustering algorithms. *Data mining and knowledge discovery hand-book.* Springer, Boston, MA. 2009. P. 269–298.

5. Firdaus S., Uddin M.A. A Survey on Clustering Algorithms and Complexity Analysis. *International Journal of Computer Science Issues (IJCSI)*. 2015. Vol. 12 (2). P. 62–85.

6. Goswami J.A Comparative Study on Clustering and Classification Algorithms. *International Journal of Scientific engineering and Applied Science (IJSEAS)*. 2015. Vol. 1 (3). P. 2395–3470.

7. Ishaq R., Nasim R. Enhancing information extraction techniques from structured database using artificial intelligence. *International Journal of Computer Science and Information Security*. 2018. Vol. 16 (11). P. 140–143.

8. Крисилов В.А., Побережник С.М. Аппроксимация сложных зависимостей структурногибкими полиномиальными и гармоническими рядами. *Управляющие системы и машины*. 2003. № 2. С. 80–86.

9. Hoifung P., Domingos P. Joint Inference in Information Extraction. *Association for the Advancement of Artificial Intelligence.* USA, 2015. Vol. 34 (5). P. 171–176.

10. Tagasovska N., Andritsos P. Distributed clustering of categorical data using the information bottleneck framework. *Information Systems*. 2017. Vol. 72. P. 161–178.

11. Крисилов В.А. Оценка сложных объектов – основной механизм при решении задач количественного обоснования решений. *Труды Одесского политехнического университета*. 2003. Вып. 1 (19). С. 102–106.

12. Cherneha K.S., Tymchenko B.I., Komleva N.O. Decision support System for Automated Medical Diagnostics. *Electrotechnic and Computer Systems*. 2016. No. 23 (99). P. 65–72.

13. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск : Изд-во Ин-та математики, 1999. 270 с.

14. Юнкеров В.І., Григор'єв С.Г. Математико-статистична обробка даних медичних досліджень. Санкт-Петербург, 2002. 266 с.

15. Komlevaya N.O., Komlevoy A.N., Chernega K.S. Designing of the specialized computer system for making pulmonology diagnosis. *CEUR Workshop Proceedings, 9th International Conference of Programming, UkrPROG*. Kyiv, 2014. Vol. 1843. P. 253–263.

16. Komleva N.O., Cherneha K.S., Tymchenko B.I., Komlevoy O.M. Intellectual approach application for pulmonary diagnosis. *Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing, DSMP*. 2016. Art. № 7583505. P. 48–52.

17. URL: https://old.phc.org.ua/pages/diseases/tuberculosis/surveillance/ statistical-information (дата звернення: 27.09.2019).

Krisilov V.A., Komleva N.O., Burdeinyi E.I. SOFTWARE FOR QUALITY ANALYSIS OF SOURCES OF MEDICAL STATISTICAL INFORMATION

The scope of statistical methods, due to their powerful theoretical base and practical tools, extends to many kinds of sciences. The introduction of statistics into medical research is accompanied by a number of difficulties due to the exceptional complexity, variability and availability of individual features of the studied objects. At the same time it is necessary to provide certain quality of the data describing the condition of the investigated object. The aim of the paper is to improve the quality of medical statistical information processing through preliminary analysis of the quality of data provided by sources of information. In comparison with the known works, in which the emphasis was on the choice of methods of analysis of the condition of the object under investigation based on input information, the approach proposed in this work allows to perform a preliminary analysis of the quality of the input information to evaluate the possibility of its further use. In order to achieve this aim, the work of the data processing and analysis system was considered in general, taking into account the quality of the input data. The list of data anomalies includes checks for discrepancies by data types, ranges of acceptable values, correlation in values in related data dependencies. As a practical example, the design of the software system was performed. It is intended to analyze the quality of medical statistical information regarding a tuberculosis situation. Information for analysis which containing more than one hundred analyzed features is taken from the website of the Ministry of Health of Ukraine. The software system allows user to perform basic data analysis operations, identify and, if possible, correct any found anomalies. The developed system implements client-server architecture. The analyst has access to advanced functionality that contains the work with anomalies and obtains the predicted values of the analyzed features as time series data. System interface is detailed, programming language, related managers and frameworks are selected. Using a software system will provide the required quality of data given by sources of medical statistical information.

Key words: *quality of information, data analysis, source of information, anomalies of input data, software system.*