

МЕТОДИКА ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

А. М. Новіков, к. т. н. П. С. Сафонов

Одеський національний політехнічний університет

Україна, м. Одеса

farysms@gmail.com; sps@onu.ua

Розкрито актуальність використання інтелектуального аналізу даних для обробки великих та надвеликих обсягів метеорологічних даних. Запропоновано методику інтелектуального аналізу метеорологічних даних, яка за рахунок кластеризації даних, послідовного аналізу та аналізу відхилень дозволяє інтерпретувати велику кількість даних, які динамічно змінюються в часі, попереджати критичні ситуації та ефективно витягати корисні дані з багатовимірних, різномірних масивів даних великих обсягів. Сформульовано основні задачі програмного застосування, що використовує запропоновану методику.

Ключові слова: інтелектуальний аналіз даних, кластеризація, сегментація, аналіз послідовностей, аналіз відхилень, метеорологічні дані.

Сьогодні ми є свідками активного розвитку технології інтелектуального аналізу даних, що пов'язано, в першу чергу, з необхідністю аналітичної обробки надвеликих обсягів інформації, що накопичується в сучасних сховищах даних та виявлення у цих даних тенденцій, закономірностей тощо. Можливість використання добре відомих методів математичної статистики та машинного навчання для вирішення завдань подібного роду відкрила нові перспективи перед аналітиками, дослідниками, прогнозістами. У цей же час спостерігається стримування розвитку виробництва у зв'язку з обмеженим доступом до сучасних технологій та використання мережевих можливостей. На сьогодні основну роль у всіх процесах виробництва відіграє людина [1].

У зв'язку з цим представляють інтерес дослідження, спрямовані на розробку, використання та інтегрування застосунків з використанням інтелектуального аналізу даних, що дозволять мінімізувати фактор людської помилки та значно економити кошти на виробничих процесах. Для реалізації такого підходу потрібно розробити систему інтелектуального аналізу даних, яка буде включати в себе можливість обробки даних, пошуку закономірностей та інтерпретування їх у консультаційні довідки [2, 3].

Мета роботи полягає у розробці методики інтелектуального аналізу даних для використання у спеціалізованому програмному застосунку обробки метеорологічних даних.

Інтелектуальний аналіз даних (ІАД) зазвичай визначають як метод підтримки прийняття рішень, заснований на аналізі залежностей між даними. Метою ІАД є виявлення неявних закономірностей в наборах даних. У нашому випадку цими даними будуть виступати метеорологічні відомості на поточний час та архівні записи. З огляду на різноманітність форм представлення даних, використовуваних алгоритмів і сфер застосування, ІАД може проводитися за допомогою програмних продуктів наступних класів: спеціалізованих програмних продуктів для інтелектуального аналізу, математичних пакетів; електронних таблиць (і різного роду надбудов над ними), засобів, які інтегровані в системи управління базами даних, інших програмних продуктів [4].

У випадку, що розглядається, буде реалізовано спеціалізований програмний застосунок для інтелектуального аналізу метеорологічних даних. В ході проведення ІАД проводиться дослідження показників атмосферного тиску, температури, швидкості вітру тощо. Їх можна представити у вигляді таблиці, кожен рядок якої відповідає одному з варіантів, а в стовпці містять значення параметрів, які його характеризують. Залежна змінна — параметр, значення якого розглядається як залежне від інших параметрів (незалежних змінних). Власне, основною задачею застосування є визначення цієї залежності з використанням методів інтелектуального аналізу даних та наступна інтерпретація цих даних у формі зрозумілої довідки [5].

За способом вирішення завдання ІАД запропонована методика відноситься до класу навчання без вчителя, метою якого є виявлення наявних закономірностей в існуючому наборі даних. При цьо-

му не потрібен навчальний набір даних, тобто набір, на якому створюється і навчається модель інтелектуального аналізу даних. Запропонована методика передбачає кластеризацію, завдання якої полягає в розподілі безлічі об'єктів на схожі за параметрами групи (кластери). При цьому, на відміну від класифікації, число кластерів і їхні характеристики можуть бути заздалегідь невідомі і визначатися в ході побудови кластерів, виходячи зі ступеня близькості, об'єднуються об'єкти за сукупністю параметрів, тобто запропонована методика ІАД відноситься до некерованих завдань. Слід зазначити, що некероване завдання дозволяє позбутися етапів навчання та тестування моделі, класифікації та регресії. Завдання визначення взаємозв'язків, а саме пошуку асоціативних правил, полягає у визначенні наборів об'єктів, що часто зустрічаються, серед безлічі подібних наборів.

Крім того, запропонована методика передбачає аналіз послідовностей або сіквенціальний аналіз для виявлення закономірностей в послідовності метеорологічних подій. Подібна інформація дозволяє, наприклад, попередити несподівані погодні умови, отримавши сигнал про настання події. Аналіз відхилень, що також передбачено методикою ІАД, дозволяє відшукати серед безлічі подій ті, які істотно відрізняються від норми. Відхилення може сигналізувати про якусь незвичну подію (несподіваний град, аномальний наступ комах).

У сучасних системах ІАД застосовується надзвичайно широкий спектр математичних, логічних і статистичних методів: від аналізу дерев рішень до нейронних мереж. Поки важко говорити про перспективність застосування або перевагу тих чи інших методів [6]. В області обробки метеорологічних даних технологія ІАД зараз знаходиться на початку шляху, і практичного матеріалу для будь-яких рекомендацій або узагальнень недостатньо, але запропонована методика передбачає можливість додавання методів ІАД за необхідності.

Інтелектуальний аналіз даних являє собою новітній напрям в області інформаційних систем в обробці метеорологічних даних, орієнтований на вирішення завдань підтримки прийняття рішень на основі кількісних та якісних досліджень надвеликих масивів різномірних метеорологічних даних. Запропонована методика дозволяє інтерпретувати велику кількість даних, які динамічно змінюються в часі, попереджати критичні ситуації, ефективно витягати корисні дані з багатовимірних, різномірних, неповних, неточних, суперечливих, непрямих масивів даних великих обсягів. Спеціалізований програмний застосунок для інтелектуального аналізу метеорологічних даних, що використовує запропоновану методику, позбавить людей від інформаційного перевантаження за рахунок переробки оперативних даних в корисну інформацію, щоб потрібні дії могли бути виконані в потрібний час.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Гурін В. К. Метеорологія и климатология в развитии сельского хозяйства.— Новочеркасск: НГМА, 2004.
2. Барсегян А. А., Куприянов М. С. Анализ данных и процессов.— СПб.: БХВ-Петербург, 2009.
3. Das M. Data-driven approaches for meteorological time series prediction: A comparative study of the state-of-the-art computational intelligence 82 techniques. // Machine Learning and Applications in Artificial Intelligence.— 2018.— Vol. 105.— P. 155–164.
4. Arroyo A. Analysis of meteorological conditions in Spain by means of clustering techniques. // Journal of Applied Logic.— 2017.— Vol. 24.— P. 76–89
5. Das M. semBnet: A semantic Bayesian network for multivariate prediction of meteorological time series data. // Pattern Recognition Letters.— 2017.— Vol. 93.— P. 192–201.
6. Badjana H. M. Bayesian trend analysis in annual rainfall total, duration and maximum in the Kara River basin (West Africa). // Journal of Hydrology: Regional Studies.— 2017.— Vol. 13.— P. 255–273.

A. M. Novikov, P. S. Safronov

Data mining technique

The paper reveals the urgency of using data mining for the processing of large volume meteorological data. The proposed technique of meteorological data mining uses data clustering, sequential analysis and deviation analysis in order to interpret a large volume of dynamically changing data, prevent critical situations, and effectively extract useful data from multidimensional, heterogeneous data sets of large volumes. The authors formulate the main tasks of a software application using the proposed data mining technique for the processing of current meteorological information and archival records.

Key words: data mining, clustering, segmentation, sequential analysis, deviation analysis, meteorological data.