дослідженні рішення є практичними і ними можна користуватися вже зараз. Систематизація рішень дозволяє легко зорієнтуватися у виборі в залежності від об'єкта і вимог.

Література

[1] Шваб К., "Четвертая промышленная революция", Москва, Россия: Эксмо, 2016.
[2] Industrial internet reference architecture [Online]. Available: http://www.iiconsortium.org/IIRA.htm.
[3] "Implementation Strategy Plattform Industrie 4.0 Results Report", Bitkom e.V., VDMA e.V., ZVEI e.V., Berlin, January 2016.
[4] D. Barry. Azure/iot-edge [Online]. Available: https://github.com/Azure/iot-edge.

References

[1] Shvab K., "Chetvertaya promyishlennaya revolyutsiya", Moskva, Rossiya: Eksmo, 2016.
[2] Industrial internet reference architecture [Online]. Available: http://www.iiconsortium.org/IIRA.htm.
[3] "Implementation Strategy Plattform Industrie 4.0 Results Report", Bitkom e.V., VDMA e.V., ZVEI e.V., Berlin, Januar 2016.
[4] D. Barry. Azure/iot-edge [Online]. Available: https://github.com/Azure/iot-edge.

УДК 519.7/007/004:681.5.015

КЛАСТЕРНИЙ АНАЛІЗ ДАНИХ В АВТОМАТИЗОВАНИХ СИСТЕМАХ ПРОСТЕЖУВАНОСТІ

Жигайло О.М.[1], Борис В.В.[2]

[1,2]Одеська національна академія харчових технологій, м. Одеса, Украина
ORCID: [1]0000-0001-6986-4673
E-mail: [1]dr_jam2006@ukr.net, [2]vitaliy_boris@ukr.net

Анотація: Важливою особливістю технологічних процесів харчових виробництв є істотний вплив характеристик сировини, що переробляється, на показники якості готової продукції. Тому при виділенні об'єкта управління пропонується розглядати разом: певний етап технологічного процесу, що реалізується одним агрегатом або якоюсь їх групою, та ті бізнес-процеси, які впливають на особливості його протікання і на одержуваний кінцевий результат. Для управління такими складними об'єктами використовуються різні автоматизовані системи, які накопичують у своїх базах даних великі обсяги інформації. Розробка і реалізація нових алгоритмів на основі методів інтелектуального аналізу даних, які з урахуванням цілей управління і даних про об'єкт управління могли б забезпечувати вибір найбільш ефективного варіанту управлінського рішення, є дуже актуальним завданням. Прийняті управлінські рішення, як результати використання розробленого алгоритму, повинні в подальшому забезпечувати еволюцію технологій управління розглянутими об'єктами.

Широке вивчення предметної області підтвердило доцільність вибору методу кластерного аналізу як основи для розроблюваного алгоритму. Це сприяло створенню авторської класифікації різних методів і алгоритмів кластеризації. Результат їх порівняльного аналізу привів до постановки завдання реалізації процедури вдосконаленою кластеризації даних на основі методу k-means, яка б забезпечувала визначення положень початкових центрів кластерів і автоматичний розрахунок їх кількості. Розроблений новий програмний модуль кластерного аналізу «Zhy & Bor» був апробований на різних тестових прикладах набору даних і при наявності "спірних" об'єктів показав свою перевагу перед результатами використання методу k-means в таких програмних інструментах, як Deductor Studio Academic, Statistica StatSoft, SPSS Modeler IBM.

Abstract: An important feature of the technological processes of food production is the significant influence of the characteristics of the raw materials on the quality indicators of the finished product. Therefore, when selecting a control object, it is proposed to consider a certain stage of the technological process realized by one aggregate or some group of them, as well as those business processes that affect the features of its flow and the receiving of final result. To control such complex objects, various automated systems are used, which accumulate large volumes of information in their databases. A very urgent task is the development and implementation of new algorithms based on methods of data mining, which, taking into account the management objectives and data on the management object, could provide the selection of the most effective version of the management solution. Accepted management decisions, as the results of using the developed algorithm, should subsequently ensure the evolution of management technologies for the objects in question.

A wide study of this topic confirmed the expediency of choosing the method of cluster analysis as the basis for the algorithm being developed. This has helped to create the author's classification of various methods and algorithms of clustering. The result of their comparative analysis led to the formulation of the task of implementing the procedure for improved data clustering based on the k-means method, which would provide the determination of the positions of the initial cluster centers and the automatic calculation of their number. The newly developed "Zhy & Bor" cluster analysis software module was tested on different test cases of the data set and, in the presence of "controversial" objects, showed its advantage over the results of using the k-means method in software tools such as Deductor Studio Academic, Statistica StatSoft, SPSS Modeler IBM.

Ключові слова: якість сировини, автоматизована система, інтелектуальний аналіз даних, кластеризація.
Keywords: raw material quality, automated system, intelligent data analysis, clustering.

Вступ

Для успішного представлення своєї продукції на різних ринках українські підприємства харчової промисловості повинні відстежувати історію створення готового продукту та передбачати існування автоматизованої системи простежуванності. Великі об'єми інформації, які може в собі містити база даних цієї системи, повинні використовуватися більш раціонально. Це повністю відповідає стандарту системи управління якістю [1], який включає в себе розробку політики та цілей у сфері якості, планування якості, забезпечення якості та поліпшення якості. Отже поліпшення якості має розглядатися як основна частина системи менеджменту, що спрямована на збільшення здатності виконання вимог до якості.

Відомо, що на якість готової продукції великий вплив мають якісні показники сировини, а на них в свою чергу впливають умови її вирощування та особливості проведення прибирання, які насамперед обумовлені метеорологічними показниками. Наприклад, у виробництві соняшникової олії її якість залежить від якості насіння соняшнику відразу після збору врожаю та від умов його зберігання, так як збирають насіння на протязі сезону, а його переробка в соняшникову олію може здійснюватися цілий рік. Тому показники якості насіння (кислотне число, вологість, олійстість, лузжистість, маса 1000 сім'янок) підлягають визначенню, аналізу та дослідженню їх впливу на різні показники (якісні та економічні) виробленої олії [2].

Те саме можна відзначити і для борошномельних виробництв. Серед показників якості зерна пшениці, що характеризують її борошномельну цінність виділяють: натура, маса 1000 зерен, крупність, форма і розміри зерна, скловидність, зольність, співвідношення оболонок і ендосперму зерна, глибина і форма борозенки, міцність зерна. Вони впливають на вихід борошна та її якість. Це, в свою чергу, може визначати її конкретне подальше використання у виготовленні різних виробів: хліба, бубликів, сухарів, макаронів, рулетів, вафель, тістечка та інших, а також в громадському харчуванні. Тобто доцільно підтримувати зв'язок між потребою ринку в борошні відповідної якості і тим зерном, що перероблюється та закупається, а також робити його продовольчу переробку економічно вигідною. Якщо вихід борошна прогнозується низький, то зерно необхідно вчасно згодувати худобі.

Ще одним цікавим прикладом є основна ділянка хлібопекарного виробництва, де відбувається процес замісу тіста. З неї починається забезпечення якості хлібобулочних виробів. Головним агрегатом цього процесу є тістомісильна машина. Для замісу тіста на такій машині може використовуватися багато програм. Вони враховують різні властивості сировини, серед якої основною є мука з показниками: вимірювач деформації клейковини (ВДК), кількість сирої клейковини, число падіння, вологість, білизна. Обрана програма повинна сприяти досягненню найвищої якості структурних властивостей тіста для забезпечення стабільно високої якості хлібобулочних виробів та мінімізації витрат сировини. При прийнятті рішення про вибір необхідної програми, обов'язково присутній людський фактор. А це впливає на його правильність та створює залежність економічних результатів підприємства від дій одного працівника. Отже виникає потреба інтелектуальної підтримки або повної автоматизації процесу вибору програми тістомісильної машини.

В наведених прикладах розглядаються певні етапи технологічного процесу, що реалізується одним агрегатом або якоюсь їх групою, а також ті бізнес-процеси, які впливають на особливості його протікання і на отримуваний кінцевий результат. Такий об'єкт дослідження відповідає всім ознакам складних об'єктів.

Мета і завдання дослідження

Довести ефективність застосування методів інтелектуального аналізу даних (Data Mining), зокрема – кластеризації, для забезпечення стабільного вибору найбільш ефективних варіантів управлінських рішень при функціонуванні складних об'єктів. Для досягнення цієї мети треба:

1) проаналізувати методи кластерного аналізу та результати їх використовування різними програмними інструментами;

2) провести розробку алгоритму кластерного аналізу даних на базі методу k-means та реалізувати на його основі програмний модуль.

Матеріали і методи

Інформаційне забезпечення - є однією з найважливіших складових сучасних автоматизованих систем управління і має постійно вдосконалювати процес надання інформації окремим особам або групам-користувачам цих систем відповідно до їх інформаційних потреб. Основна кількість таких систем призначена для підтримки керівників в управлінні підприємством. Вони базуються на даних, отриманих всередині підприємства за допомогою систем обробки транзакцій, SCADA-систем, а також на даних отриманих за його межами (наприклад в Інтернеті) або наданих діловими партнерами, постачальниками та клієнтами. Всі вони надають істотну допомогу керівникам різних рівнів. Однак інформація, що надається ("в чистому вигляді") є всього лише необхідною умовою, але недостатньою для прийняття правильних рішень. Цю достатність може забезпечити наявність різних експертних оцінок у відповідних предметних галузях знань, застосування сучасних економіко-математичних методів і моделей, що на них базуються. Створення такої можливості аналізу взаємодії всіх факторів у сукупності зможе дати особі, що приймає рішення незаперечну перевагу. Тому найбільш перспективним для майбутнього розвитку будь-якого підприємства можна вважати вдосконалення систем підтримки прийняття рішень (СППР).

Сучасні СППР будуються на основі технологій **Business Intelligence (BI)** і дозволяють користувачеві-непрограмісту легко і оперативно отримувати інформацію з різних джерел, формувати власні звіти, що настроюються, або графічні уявлення, проводити багатовимірний аналіз даних. По суті, в СППР рівня BI вирішуються дві основні задачі: 1) оперативна підготовка даних; 2) підготовка рекомендацій щодо прийняття рішень на базі формування множин альтернатив рішення і критеріїв оцінки цих альтернатив, а також вибір оптимальних альтернатив. При вирішенні першої задачі СППР має надавати аналітику дані у відповідній формі (звіти, таблиці, графіки і т.п.), яка зручна для вивчення та їх аналізу. Серед варіантів аналізу даних виділяють: 1) **інформаційно-пошуковий** - пошук необхідних даних за допомогою заздалегідь визначених запитів (СУБД, SQL); 2) **оперативно-аналітичний** - угруповання, узагальнення, багатовимірне представлення даних у будь-якому вигляді, необхідному аналітикові, без заздалегідь передбачених запитів (OLAP); 3) **інтелектуальний** - пошук функціональних і логічних закономірностей в накопичених даних, побудова моделей і правил, які пояснюють знайдені закономірності і/або прогнозують розвиток деяких процесів (з певною ймовірністю).

Стійку зацікавленість викликає **інтелектуальний аналіз даних** (ІАД, Data Mining, розвідка даних). В цілому він позначає не якусь конкретну технологію або підхід, а процес пошуку кореляцій, тенденцій, взаємозв'язків, асоціацій і закономірностей за допомогою різних математичних і статистичних методів [3]. У процесі розвідки даних багаторазово виконуються різні операції і перетворення над сирими даними (відбір ознак, кластеризація, класифікація, регресія і візуалізація). Мета цього пошуку - представити дані у вигляді, що чітко відображають бізнес-процеси, побудувати модель, за допомогою якої можна прогнозувати процеси для планування бізнесу і проводити історичний аналіз даних для побудови планів і бюджетів. Основна задача Data Mining - **прогнозування** тих чи інших процесів. До інших, більш прикладних, але дуже поширених, можна віднести: аналіз роботи персоналу, аналіз ефективності продажу товарів, оцінка потенційних клієнтів, профілювання клієнтів, аналіз результатів маркетингових досліджень, аналіз роботи регіональних відділень компанії, контроль якості продукції, оцінка дизайну продукції, аналіз конкуруючих фірм.

Так як для управління, в першу чергу, найбільш важливими є роз’яснення отриманих результатів класифікації вибірки досліджуваних об’єктів та проведення прогнозування їх часової динаміки змін у стані, то логічна послідовність і методична доступність кластеризації в задачах агрегування і розбиття на групи роблять її особо цінним інструментом. Ці переваги кластерного аналізу дозволяють вважати його таким, який можна застосовувати для вивчення багатьох нечітко формалізованих систем природного і штучного характеру, а в подальшому для поліпшення процесів в СППР.

Метою кластеризації є пошук існуючих структур. Це описова процедура, яка не робить ніяких статистичних висновків, але дає можливість провести розвідувальний аналіз і вивчити "структуру даних". Процес кластеризації полягає в угрупованні об'єктів, які представлені набором характеристик (ознак). При цьому повинна дотримуватися гіпотеза компактності: всі об'єкти, які найбільш близькі за властивостями, потрапляють в один кластер, а ті, що відрізняються між собою, розподіляються між різними кластерами. В результаті виконання цього процесу можуть вирішуватися задачі зі стиснення даних, виявлення новизни в досліджуваних об'єктах, а також для кращого розуміння даних (обробці інформації та прийняття рішень) завдяки коректному формуванню кластерної структури.

Так як кластерний аналіз паралельно розвивався в різних дисциплінах (біології, соціології, інформатиці, економіці, маркетингу і др..), то розроблено велику кількість різноманітних методів і алгоритмів, які піддавалися множинним спробам їх класифікувати. Серед найбільш цікавих та відомих можна виділити:

1. **Класична** (найбільш груба): ієрархічні, ітеративні (неієрархічні), масштабовані [4].

2. **За підходами** (не чітка): імовірнісний, теорія графів, ієрархічний, ближнього сусіда, нечіткі алгоритми, штучні нейронні мережі, еволюційний (генетичний), ансамблеві методи [5].

3. **По методам**: за способом обробки даних, за способом аналізу даних, за кількістю застосувань алгоритмів кластеризації, по можливості розширення обсягу оброблюваних даних, за часом виконання кластеризації [6].

4. **У розрізі чотирьох компонент**: тип даних, вид шуканої кластерної структури, критерії оцінки кластерної структури, метод побудови кластерної структури (а - оптимізації загальних функцій, б - комбінаторної локальної оптимізації, в - релаксаційний, г - алгоритми, що імітують природу) [7].

Аналіз та порівняння більшості методів і алгоритмів кластеризації у виділених варіантах класифікації дали можливість визнати їх основний недолік: кількість кластерів є параметром, який задається експертом або визначається як набір можливих значень. В результаті цього він викликає ітераційне проведення досліджень набору даних і може використовуватись для оцінки кожного розбиття на досягнення оптимальної кількості кластерів. Це ускладнює для них процедуру отримання остаточного результату. Тому підвищується цінність других методів і алгоритмів, які дають можливість автоматичних (або напівавтоматичних) розрахунків та оцінок. Саме такий погляд наведений на рис. 1. Запропонована класифікація демонструє вектор розвитку методів і алгоритмів кластеризації та підтверджує актуальність розробок, які б забезпечували ефективне використання вбудованого алгоритму визначення кількості кластерів та корегування результатів кластеризації з урахуванням нових даних, що змінюються у реальному часі.

Серед великого числа методів і алгоритмів кластеризації найбільш широке застосування отримав експертний алгоритм розділення k-means. Він заснований на оптимізації суми квадратів зважених відхилень координат об'єктів від центрів шуканих кластерів (центроїдів - середніх значень координат об'єктів, що входять в кластер). В основі його роботи лежить принцип оптимального в певному сенсі розбиття множини даних на k кластерів. Алгоритм намагається згрупувати дані в кластери таким чином, щоб цільова функція алгоритму розбиття досягала екстремуму.

Серед його достоїнств виділяють: простоту, швидкість, зрозумілість і прозорість роботи. Він використовується в таких відомих програмних інструментах аналізу даних, як **Deductor Studio Academic, Statistica StatSoft, SPSS Modeler IBM**. З огляду на специфіку даних, які беруться із прикладів, описаних у вступі, можна стверджувати, що базові можливості методу k-means повністю задовольняють нашим вимогам до отримуваних кластерів. У задачах, що розглядаються, немає необхідності визначати кластери різної (специфічної) форми або такого типу, при якому один з кластерів розташований всередині іншого.

Серед недоліків алгоритму k-means виділяють: відсутність чітких критеріїв вибору числа кластерів, чутливість до шумів і аномальних значень в даних. Відомо, що вибір числа k може базуватися на теоретичних міркуваннях або інтуїції дослідника, а на правильність результатів кластеризації дуже істотно впливає вибір початкових центроїдів. Так як для k-means у класичному вигляді ця процедура не має жорсткого визначення і початкові центроїди вибираються випадковим чином або по якомусь особливому алгоритму, то це позначається на точності отриманих результатів. Тому робота проводилася у напрямку розвитку алгоритму k-means з алгоритмічними доповненнями по визначенню початкових центроїдів та кількості кластерів.

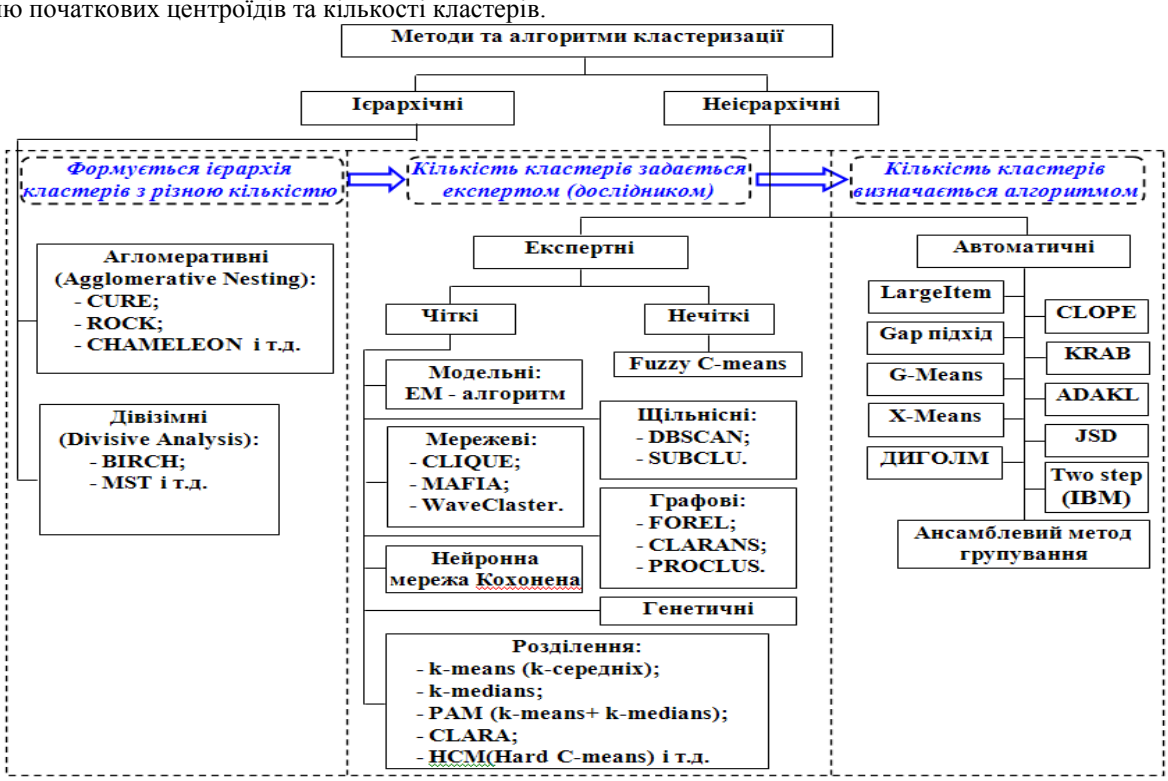

Рис. 1 – Класифікація методів і алгоритмів кластеризації

Результати дослідження та їх обговорення

Використання відомих програмних продуктів (інструментів) показало неоднозначність рішення задачі

кластеризації даних. Проблема виникла у зв'язку з наявністю серед даних таких об'єктів (“спірних”, “суперечливих”), які можуть потрапляти в різні кластера. Причому не важливо як ми назвемо метод, чіткої або нечіткої кластеризації. Важливо отримати такий результат і ті пояснення (факти), які підтвердять розміщення того чи іншого об'єкта у відповідному кластері. Якщо аналізувати розподіл по кластерам об'єктів, які являють собою не абстрактні дані, а реальні характеристики (ознаки) сировини (наприклад, показники якості насіння, зерна пшениці або борошна), то приналежність тієї чи іншої партії в кластері в подальшому може визначати регламент технологічного процесу, режим роботи обладнання і в кінцевому підсумку впливати на якість одержуваної продукції. Це означає, що якщо об'єкт не потрапив у правильний кластер, то якість на виході процесу змінилася в гіршу сторону.

Тому їх розвиток триває в різних напрямках [8], а якість одержуваних результатів дуже сильно залежить від специфіки розглянутої предметної області та від ступеня її вивченості. Продовжують проводитися нові експерименти, розроблятися нові програмні модулі і вдосконалюватися старі. Для забезпечення стабільного вибору найбільш ефективних варіантів управлінських рішень при функціонуванні складних об'єктів **був розроблений новий програмний модуль «Zhy&Bor». Він реалізує процедуру вдосконаленої кластеризації даних на основі методу k-means, а його відмінною рисою стали алгоритми автоматичного розрахунку кількості кластерів та визначення положень початкових центрів кластерів (центроїдів).**

В програмному модулі передбачено сім кроків. На кожному кроці є змога повернутися на попередні налаштування, а у разі необхідності змінити свій вибір. Перед початком роботи треба завантажити данні, які будуть підлягати кластеризації. Для цього необхідно в головному меню навігації перейти у розділ «Додати данні», натиснути команду «Вибрати файл», ввести назву таблиці та провести завантаження файлу (Excel-формату) (рис. 2а). Після переходу в розділ «Головна», ми можемо побачити перелік попередньо завантажених таблиць, який випадає у вигляді списку. Далі вибирається таблиця, яка підлягає аналізу (рис. 2б).

Крок 1. Визначення ознак для кластеризації.

Після завантаження необхідної таблиці першим кроком є вибір ознак (атрибутів) та інформаційного поля. Інформаційне поле може бути тільки одне. Воно виступає в якості назви об’єкта. Ознак кластеризації може бути різна кількість, але не менше одної, і всі вони повинні мати цифрові значення. Якщо в переліку значень стовбців знайдеться поле, в якому є хоча б одне не цифрове значення, то автоматично заблокується відповідна ознака. Нижче форми вибору ознак надається можливість перегляду обраної таблиці.

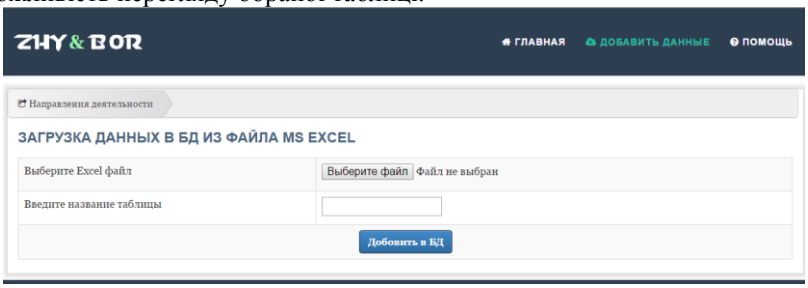

Рис. 2 – Завантаження (а) та вибір (б) даних кластерізації

Крок 2. Нормалізація (стандартизація) даних: а) «без нормалізації»; б) «нормалізація [0,1]»; в) «нормалізація [-1, 1]»; г) «стандартне відхилення»; д) «середньоквадратичне відхилення».

На другому кроці треба ретельно проаналізувати данні та обрати алгоритм нормалізації. Не можна проводити кластеризацію без попередньої їх нормалізації, якщо діапазон зміни значень відрізняється на порядок (одна ознака має діапазон значень [0…5], а інша [0…470]). Програмний модуль «Zhy&Bor» надає набір варіантів у алгоритмах нормалізації. Кожен з них може значно вплинути на результат кластеризації. Наприклад, якщо в діапазоні будь-якої ознаки присутні від'ємні значення, то слід обрати алгоритм «нормалізація [-1, ..., 1]».

Крок 3. Візуалізація даних за ознаками: а) загальна статистика; б) наведення унікальних значень з сортуванням по зростанню; в) ступінь відмінності (абсолютна різниця) між унікальними значеннями в нормованому вигляді; г) полігон розподілу частоти появи об'єкта з нерівномірними інтервалами; д) візуалізація об'єктів в 3D без кластеризації.

На третьому кроці користувачеві надається допомога по прийняттю рішення, щодо вибору кількості кластерів (у разі ручного введення кількості) або корегування результатів їх автоматичного розрахунку. В інтерфейсі є три основні вкладки. Перша «Інформація про данні» - дає можливість переглянути інформацію про статистичні данні кожної ознаки (наприклад, мінімальне та максимальне значення, кількість унікальних значень та ін..). Друга вкладка «Полігони». Їх побудова відбувається на основі відмінностей в значеннях ознаки. Тому в поточній вкладці ми маємо можливість переглянути інформацію по кожній з ознак щодо таких відмінностей, а саме кількість об’єктів які змінюються на однакову величину і її значення, як абсолютну різницю. Далі ми маємо можливість проаналізувати інтервали, які будуються по кожній з ознак, починаючи з ознаки, яка має найбільше відхилення (найбільше значення абсолютної різниці). Інтерфейс програми надає по парі графіків. На першому відображено гістограму відхилень (абсолютних різниць) (рис. 3), де по осі абсцис ідуть абсолютні значення ознаки кожного об’єкта у зростаючому порядку в форматі «номер об’єкту – значення ознаки», а по осі ординат – величина абсолютної різниці ознаки між сосідніми об’єктами.

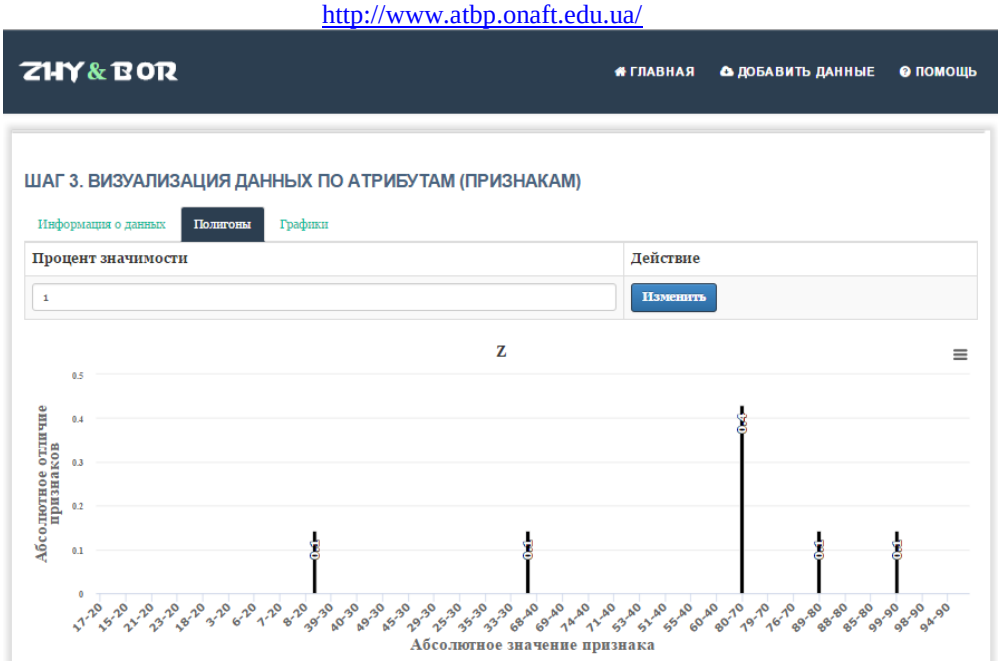

Рис. 3 – Гістограмма відхилень (абсолютних різниць)

Полігони служать для автоматичного розрахунку кількості кластерів та визначення положень їх початкових центрів. На другому графіку відображено розбиття ознаки на інтервали. Можна впливати на кількість інтервалів, якщо корегувати значення настройки поточного інтерфейсу «Відсоток значимості». Третя вкладка «Графіки». В цій вкладці є додаткові вкладки по кожній ознаці і це дає можливість побачити зміни відповідної ознаки для всіх об'єктів на графіку унікальності ознаки (рис. 4), а нижче надається графік відхилень (абсолютних різниць).

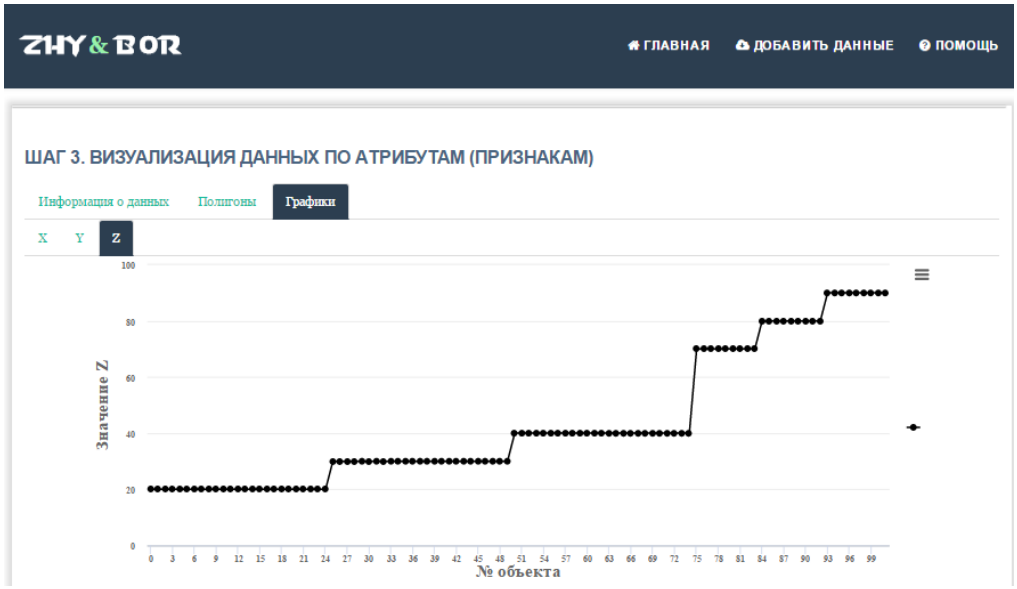

Рис. 4 – Графік унікальності ознаки для всіх об'єктів

На графіку унікальності відображено як змінює свое значення ознака для всього набору об'єктів. По осі абсцис відображено «інформаційне поле об'єкту», а по осі ординат – абсолютне значення ознаки.

Крок 4. Визначення кількості кластерів: а) ручне введення; б) автоматичний розрахунок.

На четвертому кроці необхідно визначитися з кількістю кластерів. Модуль «Zhy&Bor» надає можливість автоматичного розрахунку кількості кластерів, а також ручного вводу (згідно стандартам методу кластеризації k-means). Якщо на цьому кроці кількість кластерів призначається, то на наступному кроці вже не буде можливості для автоматичного розрахунку положень початкових центроїдів і кількості кластерів. Тому далі алгоритм зведеться до використання класичного «k-means».

Крок 5. Визначення положень початкових центроїдів: а) «випадково»; б) «по максимальній відстані»; в) «підсумовування координат»; г) «автоматично по полігонах».

На п'ятому кроці вибирається варіант алгоритму по визначенню розташування початкових центроїдів. Варіант «випадково» є класичним у неієрархічних алгоритмах кластеризації. Але він є не надійним, тому що може призвести

до неякісних результатів. Варіант «по максимальній відстані» дає якісний результат у випадках, коли об’єкт має не більше двох ознак. Алгоритм «автоматично по полігонах» базується на правилах визначення інтервалів (крок 3) та за результатами проведених досліджень є найбільш ефективним.

Крок 6. Вибір метрики для розрахунку відстані між об'єктами: а) «евклідова відстань»; б) «квадрат евклідової відстані»; в) «"зважена" евклідова відстань»; г) «відстань міських кварталів» або «Манхеттенське відстань»; д) «відстань Чебишева»; ж) «статечне відстань».

На шостому кроці вибирається варіант визначення відстані (метрики) між об'єктами, яка потрібна для відображення міри подібності, близькості об'єктів між собою по всій сукупності використовуваних ознак. За допомогою метрики кожний об’єкт зв’язується з відповідним центроїдом для подальшого формування кластерів. Більшість програм, які реалізують метод k-means, обирають варіант «евклідова відстань». Модуль «Zhy&Bor» надає більш широкий вибір метрик. Вони значно різняться між собою, тому це може впливати на результат. Вибір того чи іншого варіанту визначення відстані (метрики) не залежить від сутності даних, що обробляються, та повністю пов’язаний з інтуїцією і досвідом дослідника.

Крок 7. Вибір варіантів оцінки якості кластеризації: а) «індекс сумарного відстані»; б) «індекс Дана» в) «індекс оцінки силуету»; г) «індекс щільності CDbw»; д) індекс «VNND»; е) «індекс MB»; ж) «індекс Score Function».

На сьомому кроці можуть вибиратися різні показники якості і залежить це від умов порівняльного аналізу. Наприклад, якщо менше «індекс сумарного відстані», то менше відстань між об’єктами, тобто краще результат кластеризації. Але порівняння за даним показником є коректним лише для однакової кількості кластерів. Поясненням такої умови є те, що «індекс сумарного відстані» буде рівним нулю, якщо кількість кластерів буде рівною кількості об’єктів. «Індекс Дана» порівнює відстань між кластерами з діаметром кластера. Вважається, що якщо діаметр кластера малий у порівнянні з відстанню між кластерами, то кластери отриманої структури досить компактні і віддільні. Отже чим більше значення індексу, тим краще кластеризація. «Індекс оцінки силуету» відноситься до засобу інтерпретації та перевірки узгодженості даних в кластерах. Ця методика забезпечує таким графічним представленням, яке пояснює наскільки добре кожен об'єкт знаходиться всередині свого кластера, чим більше значення показника, ти краще. «Індекс VNND» вимірює однорідність кластерів. Чим нижче його значення, тим більше однорідність і, відповідно, краще структура кластерів. Однак індекс абсолютно не враховує віддільність і тому не зможе впізнати випадок, коли два компактних, добре відокремлених кластера виявилися об'єднаними в один. «Score Function» - чим більше значення SF, тем краще структура кластерів. «Індекс щільності CDbw» - складається з трьох складових: компактність структури кластерів, зв'язаність кластерів і віддільність кластерів. Чим більше значення індексу - тем краще.

Крок 8. Вибір варіантів заповнення кластерів: а) метод k-means з автоматизацією; б) метод k-means з оптимізаційною процедурою.

Крок 9. Візуалізація отриманих кластерів: 3D-модель, дані кластерів, гістограми центроїдів, результати оцінки якості.

Після проходження усіх кроків алгоритму кластеризації, переходимо до результату, який надається у декількох варіантах:

1. **3D модель** призначена для візуального сприйняття результату кластеризації (рис. 5). Кожний кластер виділено своїм кольором. Мається можливість повертати модель навколо своєї осі для більш детального та зручного розгляду структури кластерів. Її можна експортувати в PDF-файл.

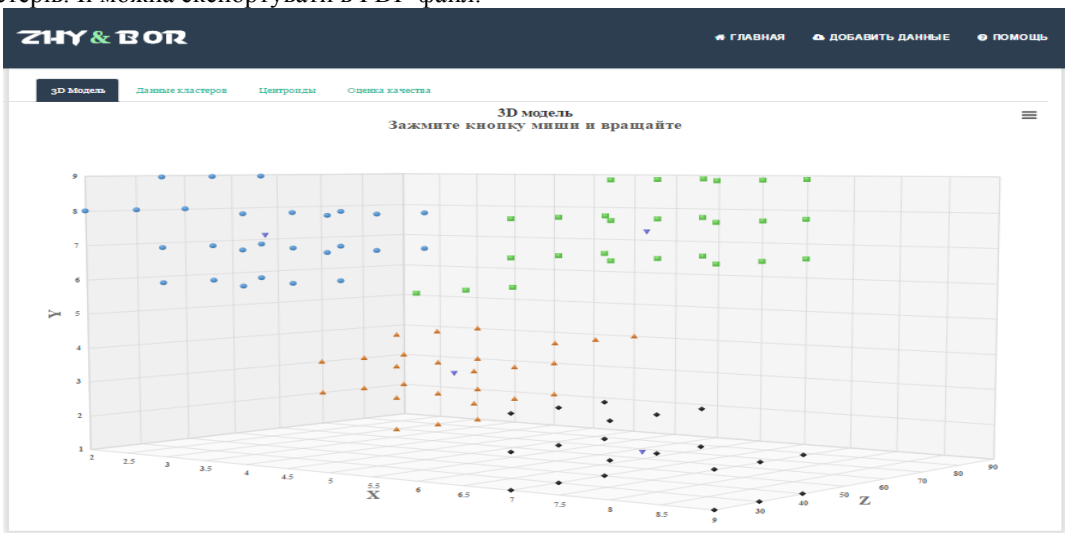

Рис. 5 – Представлення результатів кластерізації у вигляді 3D моделі

2. **Табличний** – надає можливість переглянути структуру кластерів у табличному вигляді.
3. **Центроїди** – якщо нам не потрібно дивитися результат кластеризації по всім об’єктам, а лише порівняти як

відрізняються між собою центроїди, то цей варіант дає можливість їх переглянути у вигляді гістограм.

Для перегляду оцінок якості кластеризації, необхідно перейти у вкладку «Оцінка якості». А якщо досліднику необхідно змінити будь-які налаштування, то в інтерфейсі знаходиться блок «налаштування», де є можливість вести свої корегування. Після натискання на кнопку «застосувати» результат кластеризації перерахується.

Висновки

1. Проведений аналіз та порівняння більшості методів і алгоритмів кластеризації, підтвердило, що цінність підвищується для тих, які дають можливість автоматичних (або напівавтоматичних) розрахунків кількості кластерів та оцінок якості результатів.

2. Розроблено новий програмний модуль «Zhy&Bor», який реалізує процедуру вдосконаленої кластеризації даних на основі методу k-means. У модулі реалізовано алгоритм автоматичного визначення кількості кластерів, яке пов'язано з розрахунком положень початкових центроїдів по частотним полігонам ознак об'єктів.

3. Тестування модулю на різних прикладах наборів даних продемонструвало його переваги при наявності "суперечливих" об'єктів перед результатами використання методу k-means в інструментах Deductor Studio Academic, Statistica StatSoft, SPSS Modeler IBM.

Література

[1] ДСТУ ISO 9000:2007. Системи управління якістю. Основні положення та словник термінів. (ISO 9000:2005, IDT). Київ, Держспоживстандарт України, 2008.

[2] Жигайло О.М. Використання технології Data Mining в автоматизованій системі простежуваності виробництва сирої соняшникової олії / О. М. Жигайло // Автоматизація технологічних і бізнес-процесів. – Одеса, 2014. – № 9. – с. 30-38.

[3] Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие. / Н. Б. Паклин. - 2-е изд., испр. // СПб.: Питер, 2013. - 706 с.: ил.

[4] Чубукова И.А. Data Mining. Учебное пособие. / И.А. Чубукова // М.: Интернет-Университет Информационных технологий; БИНОМ. Лаборатория знаний, 2006. – 382 с.: ил., табл. – (Серия «Основы информационных технологий»).

[5] Бериков В.Б., Лбов Г.С. Современные тенденции в кластерном анализе. / В. Б. Бериков, Г. С. Лбов // Новосибирск: ин-т математики им. С. Л. Соболева. СО РАН, 2008. — 26 с.

[6] Нейский, И.М. Классификация и сравнение методов кластеризации / И.М. Нейский // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. – М.: Изд-во ООО "Эликс +", 2008. – Выпуск 8. – С. 111-122.

[7] Миркин Б.Г. Методы кластер-анализа для поддержки принятия решений: обзор / - серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике» / Ф.Т. Алесеров, В.В. Подиновский, Б.Г. Миркин // НИУ «Высшая школа экономики», 2011. - 88 стр.

[8] Болдак А. А. Определение количества кластеров в статистических данных / А. А. Болдак, Д. Л. Сухарев // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка : збірник наукових праць. – 2011. – № 53. – С. 118–122.

References

[1] DSTU ISO 9000:2007. Systemy upravlinnya yakistyu. Osnovni polozhennya ta slovnyk terminiv. (ISO 9000:2005, IDT). Kyyiv, Derzhspozhyvstandart Ukrayiny, 2008.

[2] Zhygailo O.M. Vykorystannya tekhnolohiyi Data Mining v avtomatyzovaniy systemi prostezhuvanosti vyrobnytstva syroyi sonyashnykovoyi oliyi // Avtomatyzatsiya tekhnolohichnykh i biznes-protsesiv. – Odesa, 2014. – № 9. – s. 30-38.

[3] Paklin N.B., Oreshkov V.I. Biznes-analitika: ot dannyh k znaniyam (+SD): Uchebnoe posobie. 2-e izd., ispr. — SPb.: Piter, 2013. - 706 s.: il.

[4] Chubukova I.A. Data Mining. Uchebnoe posobie. – M.: Internet-Universitet Informacionnyh tekhnologij; BINOM. Laboratoriya znanij, 2006. – 382 s.: il., tabl. – (Seriya «Osnovy informacionnyh tekhnologij»).

[5] Berikov V.B., Lbov G.S. Sovremennye tendencii v klasternom analize. Novosibirsk: in-t matematiki im. S. L. Soboleva. SO RAN, 2008. — 26 s.

[6] Nejskij, I.M. Klassifikaciya i sravnenie metodov klasterizacii. Intellektual'nye tekhnologii i sistemy. Sbornik uchebno-metodicheskih rabot i statej aspirantov i studentov. – M.: Izd-vo OOO "EHliks +", 2008. – Vypusk 8. – S. 111-122.

[7] Mirkin B.G. Metody klaster-analiza dlya podderzhki prinyatiya reshenij: obzor / - seriya WP7 «Matematichesie metody analiza reshenij v ehkonomike, biznese i politike». - NIU «Vysshaya shkola ehkonomiki», 2011. - 88 str.

[8] Boldak A. A. Opredelenie kolichestva klasterov v statisticheskih dannyh. Visnik NTUU «KPI». Informatyka, upravlinnya ta obchyslyuval'na tekhnika : zbirnyk naukovykh prats'. – 2011. – # 53. – S. 118–122.