Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ННЦ ЗФН_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження методів витягу нових знань в системах підтримки прийняття рішень
(тема)

Виконав: студент 2 курсу, групи ПЗСзм-18-1
спеціальності _121- Інженерія програмного забезпечення_____
(код і повна назва спеціальності)
освітньо-професійної програми _Програмне__
_____забезпечення_систем_____
(повна назва освітньої програми)

_____Фареник І.В._____
(прізвище, ініціали)

Керівник _____проф. Шостак І.В._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2019 р.

Харківський національний університет радіоелектроніки

Факультет _Комп'ютерних наук_____

Кафедра _Програмної інженерії_____

Рівень вищої освіти _____другий (магістерський)_____

Спеціальність _121– Інженерія програмного забезпечення_____
(код і повна назва)

Освітньо-професійна програма _Програмне забезпечення систем_____
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

Студентові _Фаренику Іллі Вадимовичу_____
(прізвище, ім'я, по батькові)

1. Тема роботи _Дослідження методів витягу нових знань в системах підтримки прийняття рішень_____

затверджена наказом по університету від «___» _____ 2019 р № __Стз__
заповнюється вручну після отримання наказу

2. Термін подання студентом роботи до екзаменаційної комісії
10 грудня 2019 р.

3. Вихідні дані до роботи __*проаналізувати існуючі алгоритми, що використовуються для вимог підтримки прийнятя рішень, , мова розробки C#, сервер IIS Express*_____

4. Перелік питань, що потрібно опрацювати в роботі _*мета роботи, аналіз проблемної галузі і постановка задачі, опис запропонованих варіантів оптимізації, використовувані методи та алгоритми, опис розробленої програмної системи, опис застосованих оптимізацій, аналіз можливих застосувань*_____

5. Консультанти розділів роботи

Найменування розділу	Консультант (посаду, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Шостак І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	10 жовтня 2019 р.	
2.	Огляд існуючих методів	27 жовтня 2019 р.	
3.	Проектування та розробка ПЗ	15 листопада 2019 р.	
4.	Підготовка пояснювальної записки	25 листопада 2019 р.	
5.	Спецчастина	26 листопада 2019 р.	
6.	Підготовка презентації та доповіді	30 листопада 2019 р.	
7.	Попередній захист	10 грудня 2019 р.	
8.	Нормоконтроль, рецензування	11 грудня 2019 р.	
9.	Занесення диплома в електронний архів	12 грудня 2019 р.	
10.	Допуск до захисту в зав. кафедри	14 грудня 2019 р.	
* заповнюється вручну після виконання чергового пункту			

Дата видачі завдання 2019 р.

Студент _____
(підпис)

Керівник роботи _____ __проф. Шостак І.В._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 87 с., 37 рис., 3 додатки, 29 джерел.

ВЕБ-СЕРВІС, КЛАСИФІКАЦІЯ, КЛАСТЕРІЗАЦІЯ, НЕЙРОННА МЕРЕЖА, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, CBR-МЕТОД.

Об'єктом дослідження є алгоритми інтелектуального аналізу даних.

Метою роботи є дослідження й розробка методів і програмних засобів ІАД на основі прецедентів для СКБД і СКБЗ.

Методи розробки базуються на теорії статистичних та індуктивних процедур, штучних нейронних мереж, кластерного аналізу й ін.

У результаті роботи було проаналізовано предметну галузь, поставлено задачу та спроектовано модифікований CBR метод (CBR цикл), що використовує експертну інформацію для витягу прецедентів.

WEB SERVICE, CLASSIFICATION, CLUSTERING, NEURAL NETWORK, INTELLECTUAL DATA ANALYSIS, CBR-METHOD.

The object of the study is data mining algorithms.

The purpose of the work is to research and develop IAD methods and software on the basis of precedents for DBMS and DBMS.

Development methods are based on the theory of statistical and inductive procedures, artificial neural networks, cluster analysis, etc.

As a result, the subject area was analyzed, a task was set and a modified CBR method (CBR cycle) was designed that uses expert information to extract precedents..

ЗМІСТ

ВСТУП

Тема торкається актуальної в цей час проблеми в області штучного інтелекту (ШІ), пов'язану з дослідженням і розробкою ефективних методів інтелектуального аналізу даних (ІАД) і відповідних програмних засобів [1, 2]. Методи ІАД активно застосовуються в інтелектуальних системах (ІС) і, зокрема, в інтелектуальних системах підтримки прийняття рішень (ІСППР), а також у системах керування базами даних (СКБД) і знань (СКБЗ), бізнес-додатках, системах машинного навчання, системах електронного документообігу й ін.

В ІАД для витягу нових знань із наявних даних застосовуються різні методи [2, 4]: статистичні й індуктивні процедури (дерева рішень), генетичні алгоритми, штучні нейронні мережі, кластерний аналіз, прецедентні методи й ін.

Для виконання ІАД у роботі пропонується використовувати методи правдоподібних міркувань на основі прецедентів (CBR – Case-Based Reasoning) [3].

Метою роботи є дослідження й розробка методів і програмних засобів ІАД на основі прецедентів для СКБД і СКБЗ.

Для досягнення зазначеної мети необхідно розв'язати наступні завдання:

− дослідження різних технологій, методів і програмних засобів ІАД для сучасних СКБД і СКБЗ;

− аналіз проблем, пов'язаних з розробкою методів і програмних засобів ІАД для ІС і сучасних СКБД;

− розробка методів ІАД на основі прецедентів для ІС і СКБД, а також методів скорочення кількості прецедентів у базі прецедентів (БП);

− розробка відповідних алгоритмів для ІАД на основі прецедентів і алгоритмів скорочення кількості прецедентів у БП із використанням класифікаційних і кластерних методів;

− програмна реалізація прототипу підсистеми ІАД на основі прецедентів.

Поставлені завдання вирішуються з використанням методів дискретної математики, математичної логіки, методів ШІ, методів ІАД, методів правдоподібних міркувань на основі прецедентів, теорії програмування. Описаний процес виявлення знань на основі ІАД, який включає: вибір предметної області й релевантного знання для реалізації цілей кінцевого користувача комп'ютерної системи; вибір вихідної безлічі даних і підмножини змінних, які необхідні для витягу нового знання з бази фактів; уточнення даних і предпроцессінг; редукцію даних; вибір завдання DM і алгоритмів, що реалізують DM для пошуку закономірностей у даних; видачу результатів у формі, зручної для користувача; інтерпретацію отриманих даних; огляд і узгодження виявленого знання.

Розглянуті різні завдання ІАД (класифікація, кластеризація (сегментація), регресія, прогнозування, пошук асоціативних правил, аналіз послідовностей, аналіз відхилень, оцінювання, аналіз зв'язків, візуалізація даних і ін.) і методи ІАД (наприклад, методи класифікації й кластерного аналізу, послідовні моделі, дерева рішень, метод опорних векторів, байєсовські мережі, лінійна регресія, кореляційно-регресійний аналіз, алгоритми обмеженого перебору, методи на основі прецедентів і апарата штучних нейронних мереж і ін.).

Розглянуті особливості платформи Business Intelligence (BI) і технологій On-Line Analytical Processing (OLAP) і DM, застосовуваних у сучасних СКБД.

Виконаний огляд сучасних програмних засобів ІАД для СКБД (Microsoft SQL Server, Oracle, SAP BI, КРОК, Business Objects, Cognos, Information Builders, SAS і зазначені їхні переваги й недоліки, а також відзначено, що однією з перспективних можливостей розширення засобів ІАД і, зокрема, аналітичних інструментів СКБД є використання прецедентного підходу.

Запропонований модифікований CBR метод (CBR цикл), що використовує експертну інформацію (тестові набори даних) для витягу прецедентів, який підвищує якість рішення завдань ІАД на основі прецедентів за рахунок формування бази вдалих (відповідних) і невдалих (невідповідних) прецедентів у процесі виконання

CBR циклу.

Описані алгоритми витягу прецедентів і розроблений модифікований алгоритм витягу прецедентів на основі k-Nn для ІАД, що полягає в зміні значення k залежно від розміру БП. Дана модифікація дозволяє підвищити якість рішення завдань ІАД, зокрема, підвищити якість класифікації даних з використанням CBR методу при збільшенні розміру БП.

В роботі наведена архітектура розробленого прототипу CBR-системи для ІАД, що включає в себе наступні основні компоненти: користувацький інтерфейс, блок витягу прецедентів, БЗ, БП, БНП, набір тестових вибірок і модуль оптимізації БП для скорочення кількості прецедентів у БП. Описані особливості програмної реалізації прототипу CBR системи з використанням мови C# і середовища програмування MS Visual Studio 2010, а також технології Windows Forms, ADO.NET Entity Framework, аналітичної платформи Deductor 5.3 і SQL Server Analysis Services.

1 АНАЛІЗ СТАНУ РОЗВ'ЯЗАННЯ ПРОБЛЕМИ ТА ОБҐРУНТУВАННЯ ЦІЛЕЙ ДОСЛІДЖЕННЯ

1.1 Основні визначення й етапи ІАД

Методи ІАД і, зокрема, технології Data Mining (DM) сьогодні широко застосовуються в СКБД для рішення актуального завдання виявлення в даних раніше невідомих і практично корисних знань, необхідних для прийняття рішень у різних сферах людської діяльності [4].

ІАД – це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людських діяльності [5].

Під сирими даними розуміється формат даних, що не має чіткої специфікації й утримуючий неопрацьовані (або оброблені в мінімальному ступені) дані, що дозволяє уникнути втрат інформації [6]. В закордонній літературі термін ІАД трактується як Knowledge Discovery (KD) [7] і Data Mining (DM) [8]. Під KD (виявленням знанням) у базах даних (БД) розуміють який-небудь нетривіальний процес ідентифікації достовірних, нових, потенційно корисних і добре інтерпретованих зразків (структур) у даних. Таким чином, під процесом KD розуміють багатокрокову систему процедур, що включає підготовку даних, пошук зразків у БД, оцінку витягнутого знання, коректування й ітерацію процедур [9].

Під DM розуміють етап процесу KD, що полягає в застосуванні специфічних алгоритмів породження зразків, витягнутих із БД. Безліч зразків може бути відкритим, а їх перерахування реалізується спеціальним алгоритмом [4].

Особливістю завдань ІАД є те, що вихідні дані недостатньо формалізовані, але можна витягати з них нові знання, використовуючи спеціальні програми. Англійський термін «Data Mining» не має однозначного перекладу на російську мову (видобуток даних, розкриття даних, інформаційна проходка, витяг даних/інформації)

тому в більшості випадків використовується в оригіналі. Дане поняття, що з'явилося в 1978 році, набуло високу популярності в сучаснім трактуванні приблизно з першої половини 1990-х років. До цього часу, обробка й аналіз даних здійснювався в рамках прикладної статистики, при цьому в основному вирішувалися завдання обробки невеликих БД. DM – дослідження й виявлення «машиною» (алгоритмами, засобами штучного інтелекту) в сирих даних схованих знань, які раніше не були відомі, нетривіальні, практичні й доступні для інтерпретації людиною [4]. DM є одним з етапів процесу виявлення знань у базах даних (Knowledge Discovery in Databases, KDD). Під даними розуміється неопрацьований матеріал, надаваний постачальниками даних і використовуваний споживачами для формування інформації на основі даних.

Об'єкт описується як набір атрибутів. Об'єкт також відомий як запис, випадок, приклад, рядок таблиці і т.д. Атрибут – властивість, що характеризує об'єкт. Наприклад, колір очей людини, температура води і т.д. Атрибут також називають змінною, полем таблиці, виміром, характеристикою.

Генеральна сукупність (population) – уся сукупність досліджуваних об'єктів, що цікавить дослідника. Вибірка (sample) – частина генеральної сукупності, певним способом відібрана з метою дослідження й одержання висновків про властивості й характеристики генеральної сукупності.

Параметри – числові характеристики генеральної сукупності.

Статистика – числові характеристики вибірки.

Гіпотеза – частково обґрунтована закономірність знань, що служить або для зв'язку між різними емпіричними фактами, або для пояснення факту або групи фактів.

В ІАД до знань, що виявляються, пред'являються наступні вимоги [10]:

− знання повинні бути новими, раніше невідомими. Затрачувані зусилля на відкриття знань, які вже відомі користувачеві, не окупаються. Тому цінність представляють саме нові, раніше невідомі знання;

− знання повинні бути нетривіальні. Результати аналізу повинні відбивати неочевидні, несподівані закономірності, що становлять так звані сховані знання. Результати, які могли б бути отримані більш простими способами (наприклад, візуальним переглядом), не виправдовують залучення потужних методів DM;

− знання повинні бути практично корисними. Знайдені знання повинні бути застосовні, у тому числі й на нових даних, з досить високим ступенем вірогідності. Корисність полягає в тому, щоб ці знання могли принести певну вигоду при їхньому застосуванні;

− знання повинні бути доступні для інтерпретації людиною.

Знайдені закономірності повинні бути логічно з'ясовні, а якщо ні, то існує ймовірність, що вони є випадковими. Крім того, виявлені знання повинні бути представлені в зрозумілому для людини виді.

В DM для представлення знань служать різні моделі. Види моделей залежать від методів їх створення. Найпоширенішими є: правила, дерева рішень, кластери й математичні функції.

В ІАД для витягу нових знань із баз фактів застосовуються різні методи: статистичні й індуктивні процедури (дерева рішень), генетичні алгоритми, штучні нейронні мережі (ІНС), кластерний аналіз, прецедентні методи (CBR методи) і ін.

Процес ІАД включає чотири основні етапи [2]:

− на першому етапі аналітик формулює постановку завдання в термінах цільових змінних.

− на другому етапі здійснюється підготовка даних для аналізу.

− на третьому етапі проводиться аналіз даних за допомогою методів DM.

− на четвертому етапі здійснюється верифікація й інтерпретація отриманих результатів (витягнутих знань). При верифікації застосовується тестовий набір записів, виділених з вихідних даних, що не зазнали аналізу.

1.2 Аналіз процесу виявлення знань

Процес виявлення знань у БД (процес KDD) – це процес пошуку корисних знань у сирих даних. KDD містить у собі питання: підготовки даних, вибору інформативних ознак, очищення даних, застосування методів DM, постобробці даних і інтерпретації отриманих результатів.

Основними етапами процесу виявлення знань є наступні установки й процедури [2]:

– вибір предметної області й релевантного знання для реалізації цілей кінцевого користувача комп'ютерної системи;

– вибір вихідної безлічі даних і підмножини змінних, які необхідні для витягу нового знання з бази фактів;

– уточнення даних і предпроцесінг: вибір основних операцій над даними так, що вони можуть сприяти зменшенню «шумів», визначення стратегій для їхньої мінімізації;

– редукція даних: виявлення корисних особливостей даних, щоб представлення даних було адекватному рішенню завдань, відповідних до поставлених цілей;

– вибір завдання DM, тобто специфікація процесу KDD як класифікації, кластеризації і т.д.;

– вибір алгоритмів, що реалізують DM для пошуку зразків (patterns) у даних. Цей вибір повинен бути погоджений з моделями й параметрами представлення даних;

– DM: пошук зразків у формі, зручної для користувача (правила класифікації й кластеризації, регресія, дерева рішень і т.д.);

– інтерпретація породжених зразків з можливим повторенням етапів 1-7 для подальшої ітерації. Цей етап часто має на увазі використання методів, що

перебувають на стику технології DM і технології експертних систем (ЄС). Від того, наскільки ефективним він буде, у значній мірі залежить успіх рішення поставленого завдання;

– огляд і узгодження виявленого знання.

Після етапу 7 також може здійснюватися перевірка побудованих моделей. Дуже простий і часто використовуваний спосіб полягає в тому, що всі наявні дані, які необхідно аналізувати, розбиваються на дві групи. Одна з них більшого розміру, інша – меншого. На більшій групі, застосовуючи ті або інші методи DM, одержують моделі, а на меншій – перевіряють їх. По різниці в точності між тестовою й навчальною групами можна судити про адекватність побудованої моделі.

Остаточна оцінка цінності добутого нового знання виходить за рамки аналізу, автоматизованого або традиційного, і може бути проведена тільки після перетворення в життя рішення, прийнятого на основі добутого знання. Дослідження досягнутих практичних результатів завершує оцінку цінності добутого засобами DM нового знання.

Для застосування того або іншого методу DM до даних їх необхідно підготувати до цього [10]. На початковому етапі необхідно виробити якийсь чіткий набір числових і нечислових параметрів, що характеризують розглянуту проблемну область. Це завдання найменш автоматизоване в тому розумінні, що вибір системи даних параметрів проводиться людиною, хоча, звичайно, їхні значення можуть обчислюватися автоматично. Після вибору параметрів, що описують, досліджувані дані можуть бути представлені у вигляді прямокутної таблиці, у якій кожний рядок являє собою окремий випадок, об'єкт або стан досліджуваного об'єкта, а кожний стовпчик – параметри, властивості або ознаки досліджуваних об'єктів. Більшість методів DM працюють тільки з подібними прямокутними таблицями.

Подібна прямокутна таблиця є занадто сирим матеріалом для застосування методів DM і вхідні в неї дані необхідно попередньо обробити. По-перше, таблиця може містити параметри, що мають однакові значення для всього стовпчика (тобто

такі ознаки ніяк не індивідуалізують досліджувані об'єкти), отже, їх треба виключити з аналізу. По-друге, таблиця може містити деяку категоріальну ознаку, значення якої у всіх записах таблиці різні (тобто не можна використовувати це поле для аналізу даних), і його треба виключити. Нарешті, просто цих полів може бути дуже багато, і якщо всі їх включити в дослідження, то це суттєво збільшить час обчислень, оскільки практично для всіх методів DM характерна сильна залежність часу роботи від кількості параметрів (квадратична, а нерідко й експонентна).

Крім «очищення» даних по стовпцях таблиці (ознакам), іноді буває необхідно провести попереднє «очищення» даних по рядках таблиці (записам). Будь-яка реальна БД звичайно містить помилки, дуже приблизні певні значення, записи, що відповідають якимсь рідким, винятковим ситуаціям, і інші дефекти, які можуть різко понизити ефективність методів DM, застосовуваних на наступних етапах аналізу. Такі записи необхідно відкинути. Навіть якщо подібні «викиди» не є помилками, а являють собою рідкі виняткові ситуації, вони однаково чи навряд можуть бути використані, оскільки по декільком крапкам статистично значимо судити про шукану залежність неможливо.

1.3 Завдання ІАД

ІАД допомагає вирішувати багато завдань, з якими зустрічається аналітик. Серед них основними на даний момент є завдання класифікації, регресії, кластеризації й пошуку асоціативних правил [10].

По призначенню дані завдання можна розділити на описові (descriptive) і завдання прогнозування (predictive) [11].

Завдання першого класу приділяють увагу поліпшенню розуміння аналізованих даних. Ключовий момент при цьому – легкість і прозорість результатів

для сприйняття людиною. Можливо, виявлені закономірності будуть специфічною рисою саме конкретних досліджуваних даних і більше ніде не зустрінуться, але це однаково може бути корисно для аналітика. До такого виду завдань ставляться кластеризація й пошук асоціативних правил.

Рішення завдань прогнозування розбивається на два етапи. На першому етапі на підставі набору даних з відомими результатами будується модель. На другому етапі вона використовується для пророкування результатів на підставі нових наборів даних. При цьому, звичайно, потрібно, щоб побудовані моделі працювали максимально точно. До даного виду завдань відносять завдання класифікації й регресії, а також пошук асоціативних правил, якщо отримані правила можуть бути використані для пророкування появи деяких подій.

За способами рішення завдання розділяють на категорії: навчання із учителем (supervised learning) і навчання без учителя (unsupervised learning). Дана назва утворилася від терміна «машинне навчання» (machine learning), часто використовуваного в англомовній літературі [10].

У випадку навчання із учителем завдання аналізу даних вирішується в кілька етапів. Спочатку за допомогою деякого алгоритму DM будується модель аналізованих даних – класифікатор. Потім побудована модель зазнає навчання. Інакше кажучи, перевіряється якість роботи класифікатора, і якщо воно незадовільне, то проводиться додаткове навчання моделі. Так триває доти, поки не буде досягнутий необхідний рівень якості або не стане ясно, що обраний алгоритм не працює коректно з даними, або ж дані не мають структури, яку можна було б виявити. До цього типу завдань відносять завдання класифікації й регресії.

Навчання без учителя поєднує завдання, що виявляють описові моделі, наприклад, закономірності в покупках, учинених клієнтами великого магазину. Очевидно, якщо ці закономірності існують, то модель повинна їх представити. Гідністю таких завдань є можливість їх рішення без яких-небудь попередніх знань про аналізовані дані. До цих завдань ставиться кластеризація й пошук асоціативних

правил.

До основних завданням ІАД можна віднести [12, 13]:

– класифікацію;
– кластеризацію (сегментацію);
– регресію;
– прогнозування;
– пошук асоціативних правил;
– аналіз послідовностей;
– аналіз відхилень;
– оцінювання;
– аналіз зв'язків;
– візуалізацію даних і ін.

Завдання класифікації полягає в тому, що для кожного варіанта визначається категорія або клас, якому він належить [11]. Як приклад можна привести оцінку кредитоспроможності потенційного позичальника: призначувані класи тут можуть бути «кредитоспроможний» і «некредитоспроможний». Необхідно відзначити, що для рішення завдання необхідно, щоб безліч класів була відома заздалегідь і було б кінцевим і рахунковим.

Завдання кластеризації полягає в розподілі безлічі об'єктів на групи (кластери) схожих по параметрах. При цьому, на відміну від класифікації, число кластерів і їх характеристики можуть бути заздалегідь невідомі й визначатися в ході побудови кластерів виходячи зі ступеня близькості поєднуваних об'єктів по сукупності параметрів. Інша назва цього завдання – сегментація. Наприклад, Інтернет-Магазин може бути зацікавлений у проведенні подібного аналізу бази своїх клієнтів, для того, щоб потім сформувати спеціальні пропозиції для виділених груп, враховуючи їх особливості. Кластеризація відноситься до завдань навчання без учителя.

Завдання регресії багато в чому схоже із завданням класифікації, але в ході його рішення проводиться пошук шаблонів для визначення числового значення.

Іншими словами, тут параметром передвіщається, як правило, число з безперервного діапазону.

Окремо виділяється завдання прогнозування нових значень на підставі наявних значень числової послідовності (або декількох послідовностей, між значеннями в яких спостерігається кореляція) [11]. При цьому можуть ураховуватися наявні тенденції (тренди), сезонність, інші фактори. Класичним прикладом є прогнозування цін акцій на біржі.

Завдання пошуку асоціативних правил, яке також називається завданням визначення взаємозв'язків, полягає у визначенні наборів об'єктів, що часто зустрічаються, серед безлічі подібних наборів. Класичним прикладом є аналіз споживчого кошика, який дозволяє визначити набори товарів, що найчастіше зустрічаються в одному замовленні (або в одному чеку). Ця інформація може потім використовуватися маркетологами при розміщенні товарів у торговельному залі або при формуванні спеціальних пропозицій для групи зв'язаних товарів. Дане завдання також відноситься до класу навчання без учителя.

Аналіз послідовностей або секвенціальний аналіз одними авторами розглядається як варіант попереднього завдання, іншими – виділяється окремо [11]. Метою, у цьому випадку, є виявлення закономірностей у послідовностях подій. Подібна інформація дозволяє, наприклад, попередити збій у роботі інформаційної системи, одержавши сигнал про настання події, що часто передує збою подібного типу. Інший приклад застосування – аналіз послідовності переходів по сторінках користувачів web-сайтів.

Аналіз відхилень дозволяє відшукати серед безлічі подій ті, які суттєво відрізняються від норми. Відхилення може сигналізувати про якусь незвичайну подію (несподіваний результат експерименту, шахрайська операція по банківській карті та ін.) або, наприклад, про помилку введення даних оператором.

Завдання оцінювання зводиться до пророкування безперервних значень ознаки.

Аналіз зв'язків – завдання знаходження залежності у наборі даних.

Для рішення завдання візуалізації використовуються графічні методи, що показують наявність закономірностей у даних [14]. У результаті візуалізації створюється графічний образ аналізованих даних (наприклад, методи візуалізації на основі представлення даних в 2-D і 3-d вимірах).

1.4 Методи ІАД

Усі методи ІАД підрозділяються на дві великі групи за принципом роботи з вихідними навчальними даними [15]. У цій класифікації верхній рівень визначається на підставі того, чи зберігаються дані після аналізу або вони дистилюються для наступного використання.

У випадку безпосереднього використання (збереження) даних вихідні дані зберігаються в явному деталізованому виді й безпосередньо використовуються на стадіях прогнозування (побудови прогнозних моделей) (а також на стадії аналізу виключень). Проблема цієї групи методів полягає в тому, що при їхньому використанні можуть виникнути складності аналізу надвеликих БД (big data). До методів даної групи відносять кластерний аналіз, методи класифікації, міркування на основі аналогій і прецедентів (CBR методи).

При використанні технології дистиляції шаблонів один зразок (шаблон) інформації витягається з вихідних даних і перетворюється в якісь формальні конструкції, вид яких залежить від конкретного використовуваного методу DM. На етапах прогнозування й аналізу виключень використовуються отримані формальні конструкції, які значно компактніше самих БД. До методів даної групи відносять різні логічні методи (нечіткі запити й аналіз, символьні правила, дерева рішень, генетичні алгоритми), методи візуалізації, методи крос-табуляції (байєсовські

мережі довіри, крос-таблична візуалізація), статистичні методи, методи, засновані на нейронних мережах [15].

Статистичні методи найбільше часто застосовуються для рішення завдань прогнозування. Існує безліч методів статистичного аналізу даних, серед них, наприклад, кореляційно-регресійний аналіз, кореляція рядів динаміки, виявлення тенденцій динамічних рядів, гармонійний аналіз [15].

Методи DM також можна класифікувати по завданням DM. Відповідно до такої класифікації можна виділити дві групи. Перша з них – це поділ методів DM на вирішальні завдання сегментації (тобто завдання класифікації й кластеризації) і завдання прогнозування.

Якщо класифікувати методи відповідно до використовуваних у них моделей, то можна виділити методи, спрямовані на одержання описових результатів, і методи, спрямовані на одержання результатів прогнозування.

Описові методи служать для знаходження шаблонів або зразків, що описують дані, які піддаються інтерпретації з погляду аналітика. До них відносяться ітеративні методи кластерного аналізу, у тому числі: алгоритм k-середн, k-медіан, ієрархічний метод кластерного аналізу карти, що самоорганізовується, Кохонена, методи кросс-табличної візуалізації та ін. [15].

Прогнозуючі методи використовують значення одних змінних для прогнозування невідомих (пропущених) або майбутніх значень інших (цільових) змінних. До даної групи методів відносять нейронні мережі, дерева рішень, лінійну регресію, метод найближчого сусіда, метод опорних векторів і ін. [15].

Далі наведений ряд основних методів, які використовуються для ІАД.

Асоціація (або відношення), найбільш відомий і простий метод ІАД [16]. Для виявлення моделей робиться просте зіставлення двох або більше елементів, часто того самого типу. Наприклад, відстеження переваг і звичок покупців при покупці товарів у магазині.

Класифікацію можна використовувати для одержання уявлення про тип

покупців, товарів або об'єктів, описуючи кілька атрибутів для ідентифікації певного класу [17]. Наприклад, автомобілі легко класифікувати по типу (седан, позашляховик, кабріолет), визначивши різні атрибути (кількість місць, форма кузова, ведучі колеса). Вивчаючи новий автомобіль, можна віднести його до певного класу, порівнюючи атрибути з відомим визначенням. Ті ж принципи можна застосувати й до покупців, наприклад, класифікуючи їх за віком і соціальній групі.

Крім того, класифікацію можна використовувати в якості вхідних даних для інших методів. Наприклад, для визначення класифікації можна застосовувати дерева прийняття рішень. Кластеризація дозволяє використовувати загальні атрибути різних класифікацій з метою виявлення кластерів.

Кластерний аналіз – це спосіб угруповання багатомірних об'єктів, заснований на представленні результатів окремих спостережень крапками відповідного геометричного простору з наступним виділенням груп як «згустків» цих крапок (кластерів, таксонів) [18]. Даний метод дослідження одержав розвиток в останні роки у зв'язку з можливістю комп'ютерної обробки великих БД. Кластерний аналіз припускає виділення компактних, вилучених друг від друга груп об'єктів, відшукує «природню» розбивку сукупності на області скупчення об'єктів. Він використовується, коли вихідні дані представлені у вигляді матриць близькості або відстаней між об'єктами або у вигляді крапок у багатомірному просторі. Найпоширеніші дані другого виду, для яких кластерний аналіз орієнтований на виділення деяких геометрично вилучених груп, усередині яких об'єкти близькі.

Досліджуючи один або більше атрибутів, або класів, можна згрупувати окремі елементи даних разом, одержуючи структурований висновок. На простому рівні при кластеризації використовується один або кілька атрибутів в якості основи для визначення кластера подібних результатів. Кластеризація корисна при визначенні різної інформації, тому що вона корелюється з іншими прикладами, так що можна побачити, де подоби й діапазони узгодяться між собою [17].

Метод кластеризації працює в обидва боки. Можна припустити, що в певній

крапці є кластер, а потім використовувати свої критерії ідентифікації, щоб перевірити це. Графік, зображений на рис. 1, демонструє наочний приклад [17].

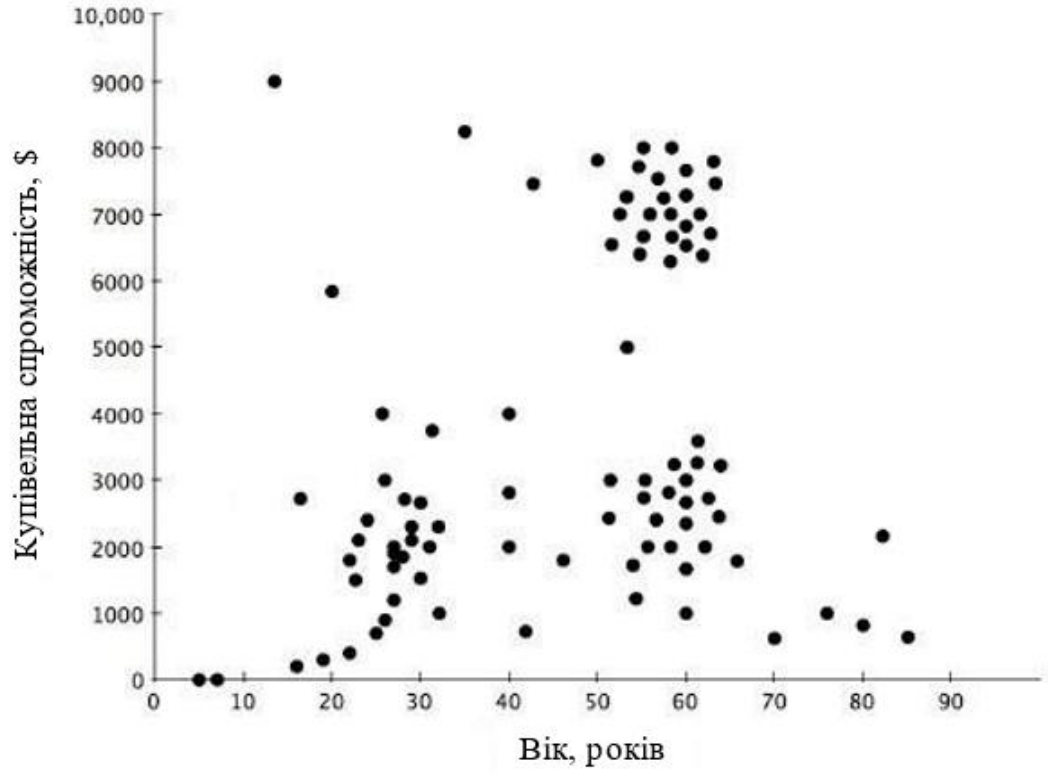

Рисунок 1 – Приклад кластеризації покупців

В цьому прикладі видно два кластери, у цьому випадку висунули гіпотезу й перевірили її на простому графіку, який можна побудувати за допомогою будь-якого відповідного програмного забезпечення для побудови графіків. Для більш складних комбінацій потрібен повний аналітичний пакет, особливо якщо потрібно автоматично засновувати рішення на інформації про найближчого сусіда. Така побудова кластерів являє собою спрощений приклад так званого образа найближчого сусіда. Окремих покупців можна розрізняти по їхній буквальній близькості один до другого на графіку. Досить імовірно, що покупці з того самого кластера розділяють і інші загальні атрибути, і це припущення можна використовувати для пошуку, класифікації й інших видів аналізу членів набору даних.

Метод кластеризації можна застосувати й у зворотну сторону: враховуючи

певні вхідні атрибути, виявляти різні артефакти. Наприклад, недавнє дослідження чотиризначних Pin-*Кодів* виявили кластери чисел у діапазонах 1-12 і 1-31 для першої й другої пар. Зобразивши ці пари на графіку, можна побачити кластери, пов'язані з датами (дні народження, ювілеї).

Прогнозування – це широка тема, яка простирається від пророкування відмов компонентів устаткування до виявлення шахрайства і навіть прогнозування прибутків компанії [17]. У комбінації з іншими методами ІАД прогнозування припускає аналіз тенденцій, класифікацію, зіставлення з моделлю й відносини. Аналізуючи минулі події або екземпляри можна пророкувати майбутнє.

Наприклад, використовуючи дані по авторизації кредитних карт, можна об'єднати аналіз дерева рішень минулих транзакцій людини із класифікацією й зіставленням з історичними моделями з метою виявлення шахрайських транзакцій [17].

Послідовні моделі, які часто використовуються для аналізу довгострокових даних, – корисний метод виявлення тенденцій, або регулярних повторень подібних подій. Наприклад, за даними про покупців можна визначити, що в різну пору року вони купують певні набори продуктів [17]. По цій інформації додаток прогнозування купівельного кошика, ґрунтуючись на частоті й історії покупок, може автоматично припустити, що в кошик будуть додані ті або інші продукти.

Дерево рішень, пов'язане з більшістю інших методів (головним чином, із класифікацією й прогнозуванням), можна використовувати або в рамках критеріїв відбору, або для підтримки вибору певних даних у рамках загальної структури [17]. Дерево рішень починає функціонувати із простого питання, яке має дві відповіді (іноді більше). Кожна відповідь приводить до наступного питання, допомагаючи класифікувати й ідентифікувати дані або робити прогнози.

На рис. 2 наведений приклад класифікації несправних станів технічного об'єкта [17].

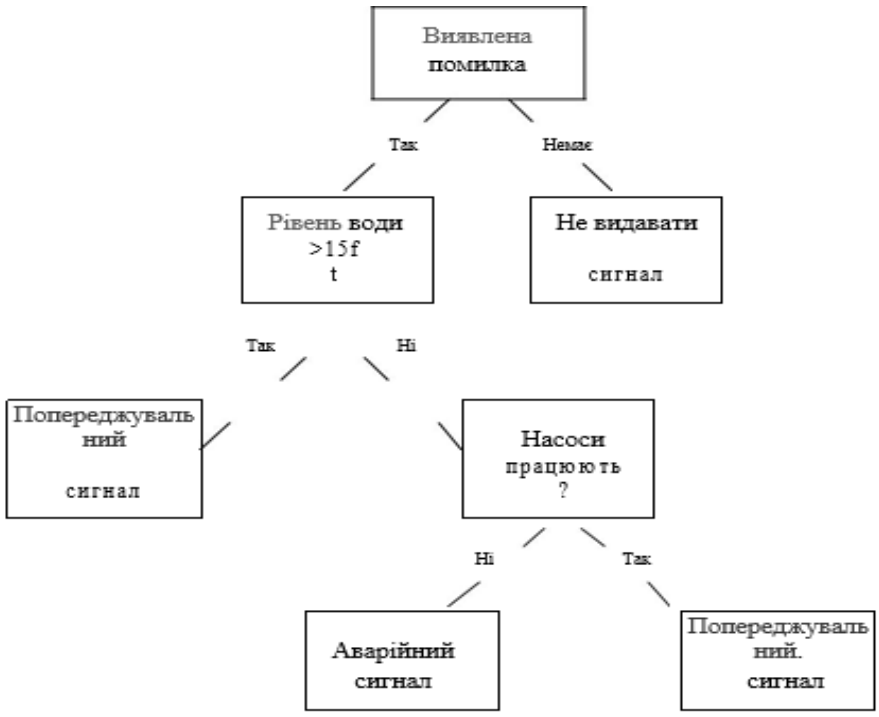

Рисунок 2 –Дерево рішень класифікації несправних станів технічного об'єкта

Дерева рішень часто використовуються із системами класифікації інформації про властивості й із системами прогнозування, де різні прогнози можуть ґрунтуватися на минулому історичному досвіді, який допомагає побудувати структуру дерева рішень і одержати результат.

Обробка із запам'ятовуванням – для всіх зазначених методів часто має сенс записувати й згодом вивчати отриману інформацію [17]. Для деяких методів це зовсім очевидно. Наприклад, при побудові послідовних моделей і навчанні з метою прогнозування аналізуються історичні дані з різних джерел і екземплярів інформації.

Метод опорних векторів – набір схожих алгоритмів навчання із учителем, що використовуються для завдань класифікації й регресійного аналізу. Належить до сімейства лінійних класифікаторів, може також розглядатися як спеціальний випадок регуляризації по Тихонову [19]. Особливою властивістю методу опорних векторів є безперервне зменшення емпіричної помилки класифікації й збільшення зазору, тому метод також відомий як метод класифікатора з максимальним зазором.

Байєсовська мережа – графічна ймовірнісна модель, що представляє собою безліч змінних і їх ймовірнісних залежностей по Байєсу [20]. Наприклад, байєсовська мережа може бути використана для обчислення ймовірності того, чи хворий пацієнт по наявності або відсутності ряду симптомів, ґрунтуючись на даних про залежність між симптомами й хворобами.

Лінійна регресія – використовувана в статистиці регресійна модель залежності однієї (що пояснюється, залежної) змінної *y* від інший або декількох інших змінних (факторів, регресії, незалежних змінних) *x* з лінійною функцією залежності [21]. Модель лінійної регресії є часто використовуваною й найбільш вивченою в економетриці. А саме вивчені властивості оцінок параметрів, одержуваних різними методами при припущеннях про ймовірнісні характеристики факторів, і випадкових помилок моделі. Граничні (асимптотичні) властивості оцінок нелінійних моделей також виводяться виходячи з апроксимації останніх лінійними моделями. Необхідно відзначити, що з економетричної точки зору більш важливе значення має лінійність по параметрах, ніж лінійність по факторах моделі.

Кореляційно-регресійний аналіз – класичний метод стохастичного моделювання господарської діяльності [22]. Він вивчає взаємозв'язки показників господарської діяльності, коли залежність між ними не є строго функціональною й перекручена впливом сторонніх, випадкових факторів. При проведенні кореляційно-регресійного аналізу будують різні кореляційні й регресійні моделі господарської діяльності. У цих моделях виділяють факторні й результативні показники (ознаки).

Кореляційний аналіз ставить завдання виміряти тісноту зв'язку між змінними, що варіюють й оцінити фактори, що виявляють найбільший вплив на результативну ознаку. Регресійний аналіз призначений для вибору форми зв'язку й типу моделі для визначення розрахункових значень залежної змінної (результативної ознаки).

Алгоритми обмеженого перебору у минулому запропоновані в середині 60-х років М.М. Бонгардом для пошуку логічних закономірностей у даних [14]. З тих пір вони продемонстрували свою ефективність при рішенні безлічі завдань із всіляких

областей. Ці алгоритми обчислюють частоти комбінацій простих логічних подій у підгрупах даних.

Основною метою використання прецедентного підходу є одержання рішення для поточної ситуації на основі прецедентів, які вже мали місце в минулому.

Прецедентний підхід може бути використаний у випадку, коли достовірні алгоритми не застосовні (неповнота знань про предметну область або наявність обмежень по часу і обчислювальних ресурсах) або взагалі відсутні.

Застосування методу для рішення завдань виправдане при виконанні наступних умов:

– подібні завдання повинні мати подібні рішення (принцип регулярності);

– види завдань, з якими зустрічається вирішувач, повинні мати тенденції до повторення.

Прецедент – це ситуація, рішення для якої вже відомо. Прецедент може бути отриманий у результаті роботи системи або як приклад рішення від експерта в проблемній області. Прецедент складається з опису проблемної ситуації, застосованого рішення й результату застосування рішення. Прецедент може містити не тільки позитивний результат. Інформацію про негативний результат застосування рішення треба так само зберегти в БП, щоб уникнути подібних результатів у майбутньому.

Часом для рішення завдань класифікації й кластеризації доцільно використовувати апарат ІНС. У даного підходу є ряд особливостей:

– ІНС мають здатність навчатися на прикладах;

– ІНС легко працюють у розподілених системах з великою паралелізацією у силу своєї природи;

– оскільки ІНС підбудовують свої вагові коефіцієнти, ґрунтуючись на вихідних даних, це допомагає зробити вибір значимих характеристик менш суб'єктивним.

На практиці дуже рідко використовується тільки один із цих методів.

Наприклад, класифікація й кластеризація – подібні методи. Використовуючи кластеризацію для визначення найближчих сусідів, можна додатково уточнити класифікацію. Дерева рішень часто використовуються для побудови і виявлення класифікацій, які можна простежувати на історичних періодах для визначення послідовностей і моделей [17].

1.5 Обґрунтування цілей дослідження

На основі наведених основних визначень (ІАД, *KDD*, *DM* і ін.) і етапів ІАД, що включають формулювання постановки завдання в термінах цільових змінних, підготовку даних для аналізу, аналіз даних за допомогою методів *DM*, верифікацію й інтерпретацію отриманих результатів (витягнутих знань).

В роботі потрібно описати процес виявлення знань на основі ІАД, який включає: вибір предметної області й релевантного знання для реалізації цілей кінцевого користувача комп'ютерної системи; вибір вихідної безлічі даних і підмножини змінних, які необхідні для витягу нового знання з бази фактів; уточнення даних і предпроцессинг; редукцію даних; вибір завдання *DM* і алгоритмів, що реалізують *DM* для пошуку закономірностей у даних; видачу результатів у формі, зручної для користувача; інтерпретацію отриманих даних; огляд і узгодження виявленого знання.

На засадах наведеного огляду різних завдань ІАД (класифікація, кластеризация (сегментація), регресія, прогнозування, пошук асоціативних правил, аналіз послідовностей, аналіз відхилень, оцінювання, аналіз зв'язків, візуалізація даних і ін.) і методів ІАД (методи класифікації й кластерного аналізу, послідовні моделі, дерева рішень, метод опорних векторів, байесовские мережі, лінійна регресія, кореляційно-регресійний аналіз, алгоритми обмеженого перебору, *CBR* методи,

методи на основі штучних нейронних мереж і ін.), на основі якого ухвалене рішення основну увагу в роботі приділити перспективним *CBR* методам для ІАД.

Потрібно визначити їхні переваги й недоліки, а також установлена перспективність можливості розширення засобів ІАД для сучасних СКБД на основі застосування прецедентного підходу.

2 ОПИС ПРОВЕДЕНИХ ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

2.1 Аналіз програмних засобів, що забезпечують ІАД

На світовому ринку корпоративних систем керування базами даних (СКБД) домінуюче положення займає традиційна трійка продуктів: IBM DB2, Microsoft SQL Server і Oracle. Більше 80% ринку СКБД протягом довгих років контролюється трьома компаніями виробниками: Microsoft SQL Server, Oracle і IBM [23].

На сьогоднішній день усі представлені на ринку сучасні СКБД містять у собі різні набори компонентів для ІАД, які входять до складу інструментів платформи Business intelligence.

Під Business intelligence (BI) розуміють програмне забезпечення, створене для допомоги менеджерові в аналізі інформації про свою компанію і її оточенні [24]. Більшість інструментів BI застосовуються кінцевими користувачами для доступу, аналізу й генерації звітів по даним, які найчастіше розташовуються в сховищу, вітринах даних або оперативних складах даних. Розроблювачі додатків використовують Bi-Платформи для створення й впровадження Bi-Додатків, які не розглядаються як Bi-Інструменти. Прикладом Bi-Додатка є інформаційна система керівника (EIS – Executive Information System).

Одним з основних компонентів програмних рішень класу BI є засоби ІАД, що базуються на технології аналітичної обробки в реальному часі (OLAP – On-Line Analytical Processing).

Під OLAP розуміють технологію обробки даних, що полягає в підготовці сумарної (агрегованої) інформації на основі великих масивів даних, структурованих по багатомірному принципу. Реалізації технології OLAP є компонентами програмних рішень класу BI [25].

Технологія OLAP орієнтована, головним чином, на обробку нерегламентованих запитів до сховищ даних. Створення сховищ даних викликане

тим, що аналізувати дані й, зокрема, дані OLTP-Систем (On-Line Transaction Processing – оперативна обробка транзакцій у реальному часі) [26] прямо неможливо або важко, тому що вони є розрізненими, зберігаються у форматах різних СКБД і в різних сегментах корпоративної мережі. У цілому можна сказати, що дані OLTP-Систем не орієнтовані на потреби аналітиків. Тому основним завданням сховища є представлення даних для аналізу в одному місці в рамках простої і зрозумілої структури. На рис. 3 показані компоненти, що входять у типове сховище даних [27]. Суцільні стрілки позначають потоки даних, пунктирні – метаданих.

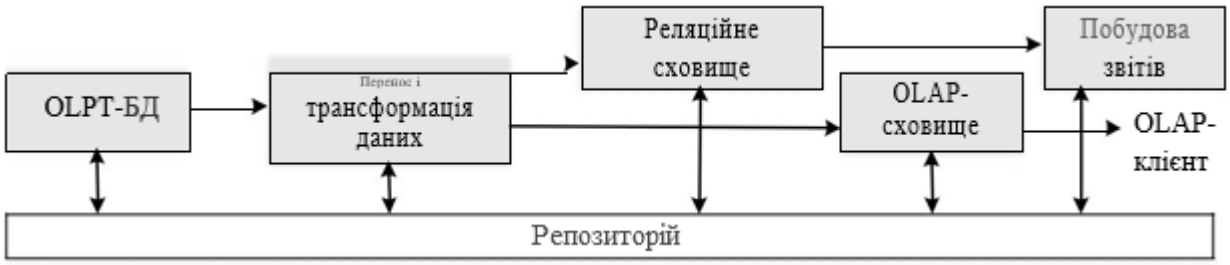

Рисунок 3 – Структура сховища даних

Основна мета аналізу даних – якісна й кількісна оцінка досягнутих результатів і (або) динаміки діяльності компанії. Принципи OLAP, використовувані для цього, у минулому сформульовані Є. Коддом [23]. Центральне місце серед них займає підтримка багатомірного представлення даних. У багатомірній моделі даних БД представляється у вигляді одного або декількох кубів даних (гіперкубів).

Осями гіперкуба служать основні атрибути аналізованого бізнесу-процесу. Незалежні виміри гіперкуба представляють багатомірний простір даних. Кожному виміру відповідає атрибут, що характеризує одне з якісних властивостей даних: час, територію, категорію продукції й т.п. На перетині осей-вимірів (dimensions), тобто в гнізді гіперкуба, утримуються дані, що кількісно характеризують аналізований процес. Ці дані називаються заходами (measures) або показниками.

В процесі аналізу виконуються операції побудови перетинів (проекцій) гіперкуба шляхом фіксації значень наборів атрибутів-координат, а також операції

стиску гіперкуба шляхом використання значень атрибутів-вимірів більш високих рівнів ієрархії й відповідного агрегування значень, асоційованих з ними показників. Ієрархічні відносини можуть бути природно введені для ряду атрибутів. Можуть застосовуватися й зворотні операції деталізації даних.

Помітимо, що куб даних розглядається як концептуальне, а не фізичне представлення. Для забезпечення зручності сприйняття даних аналітиками використовуються операції обертання куба шляхом зміни порядку вимірів. Для візуалізації даних з гіперкуба, як правило, застосовуються двовимірні представлення у вигляді таблиць, що мають складні ієрархічні заголовки рядків і стовпців. Двовимірне представлення куба можна одержати, фіксуючи значення всіх вимірів, крім двох.

Багатомірність в OLAP-Додатках втілюється в рамках двох або трирівневої архітектури. Перший рівень підтримує багатомірне представлення даних, абстраговане від їхньої фізичної структури. Він містить засоби багатомірної візуалізації й маніпулювання даними для кінцевого користувача. Другий рівень забезпечує багатомірну обробку. Він включає мову формулювання багатомірних запитів (SQL для цих цілей непридатний) і програмний процесор, здатний виконувати такі запити. На третьому рівні архітектури реалізується фізична організація зберігання багатомірних даних. У рамках його для підтримки багатомірних моделей даних використовуються або спеціальні OLAP-СКБД, або звичайні реляційні структури.

Найбільше застосування OLAP знаходить у продуктах для фінансового планування, у сховищах даних, у рішеннях класу BI.

Список найбільш відомих виробників комерційних OLAP-продуктів, згідно OLAP Report [24], включає: Microsoft (Microsoft SQL ServerAnalysis Services), Hyperion (HyperionEssbase), Cognos (CognosPowerPlay), BusinessObjects, Microsofttegy, SQL (SAP BW), Cartesis (Cartesis Magnitude), Servers Union/MIS Analysisle (Oracle Express, OLAP Option), Applix (IBM Cognos TM1).

Існує трохи open-source рішень, включаючи Mondrian і Palo [25]. Шар фізичного зберігання даних реалізується або в реляційних, або в багатомірних структурах багатомірних масивів, що представляються у представленні. Звичайно OLAP-продукти забезпечують обидва способи зберігання, а також їх комбінації:

MOLAP (Multidimensional OLAP) – і детальні дані, і агрегати даних зберігаються в багатомірній БД;

ROLAP (Relational OLAP) – детальні дані зберігаються в реляційній БД, агрегати – у спеціально створених службових таблицях;

HOLAP (Hybrid OLAP) – детальні дані зберігаються в реляційній БД, агрегати – у багатомірній БД.

В ROLAP використовуються дві схеми зберігання багатомірних даних: зірка й сніжинка. У БД входять таблиця фактів і ряд таблиць вимірів. Рядок таблиці фактів представляє набір фактів, які можуть бути як атомарними, так і агрегованими, тобто відповідними до сукупностей значень елементів вимірів. Для кожної таблиці вимірів асоційований їй рядок таблиці фактів містить значення зовнішнього ключа. У рядках таблиць вимірів зазначені значення первинних ключів, у якості яких виступають значення атрибутів для різних вимірів.

При виконанні запитів використовуються операції з'єднання таблиці фактів і таблиць вимірів. У схемі типу зірка таблиці вимірів є денормалізованими й можуть містити складені первинні ключі. Це забезпечує спрощення запитів і скорочення кількості операцій з'єднань таблиць при їхньому виконанні, а також підвищує наочність представлення даних. Розплатою за позитивний ефект є надмірність даних, що викликає ріст необхідного обсягу пам'яті. Для мінімізації надмірності використовується схема типу сніжинка. У ній таблиці вимірів нормалізовані за рахунок їх декомпозиції.

Сьогодні більшість систем OLAP загострює увагу тільки на забезпеченні доступу до багатомірних даних, а більшість засобів ІАД мають справу з одномірним представленням даних. Ці два види аналізу повинні бути тісно зв'язані, тобто

системи OLAP повинні фокусуються не тільки на доступі, але й на пошуку закономірностей. K. Parsaye уводить складений термін «OLAP DM» (багатомірний інтелектуальний аналіз) для позначення такого об'єднання (рис. 4).

Рисунок 4 – Архітектура системи багатомірного ІАД

Технологія «OLAP Mining» пропонує кілька варіантів інтеграції двох технологій:

– «Cubing then mining». Можливість виконання інтелектуального аналізу повинна забезпечуватися над будь-яким результатом запиту до багатомірного концептуального представлення, тобто над будь-яким фрагментом будь-якої проекції гіперкуба показників;

– «Mining then cubing». Подібно даним, витягнутим зі сховища, результати інтелектуального аналізу повинні представлятися в гіперкубічній формі для наступного багатомірного аналізу;

– «Cubing while mining». Цей гнучкий спосіб інтеграції дозволяє

автоматично активізувати однотипні механізми інтелектуальної обробки над результатом кожного кроку багатомірного аналізу (переходу між рівнями узагальнення, витягу нового фрагмента гіперкуба і т.д.).

Далеко не всі виробники надають сьогодні досить потужні засоби інтелектуального аналізу багатомірних даних у рамках систем OLAP. Проблема також полягає в тому, що деякі методи ІАД (байєсовські мережі, метод найближчого сусіда) незастосовні для завдань багатомірного інтелектуального аналізу, тому що засновані на визначенні подібності деталізованих прикладів і не здатні працювати з агрегованими даними.

2.2 Аналіз програмних засобів ІАД для СКБД

Розглянуто докладніше деякі додатки BI і засобу ІАД від різних виробників.

ІАД у СКБД Microsoft SQL Server – компанія Microsoft вийшла на ринок засобів BI за рахунок компаній, що бідують у відносно недорогих аналітичних рішеннях [33].

Microsoft BI являє собою набір продуктів, що дозволяють організаціям ухвалювати рішення на підставі достовірної інформації, отриманої із внутрішніх і зовнішніх джерел даних [34]. З погляду варіантів здійснення аналітичної діяльності можливі три зв'язані сценарії: персональна, колективна й корпоративна аналітика.

Компоненти Microsoft BI – Інструменти візуалізації й аналізу даних (або презентаційний рівень), у тому числі:

- інформаційні панелі й звіти Power BI;
- звіти й інформаційні панелі Sharepoint Insights;
- портал і додаток Datazen для планшетів і смартфонів.

Клієнтські інструменти, у тому числі:

– Reportbuilder для створення звітів SQL Server Reporting Services;

– Visioдля створення довільних схем і прив'язки до цих схем даних зісховища, аналітичних кубів і довільних джерел;

– Excelдля створення офісними користувачами довільних звітів велектронних таблицях і надбудови для Excel:

– Powerpivot для Excel для самостійного підключення нових джерел даних і створення нових розрахункових показників офісними користувачами;

– Power View для самостійного створення інтерактивних звітів;

– Power Map для відображення шарів даних на тривимірній карті;

– DM Add-ins для Office IAД в офісних додатках; о Power BI Desktopдля створення звітівPower BI;

– Datazen Publisherдля створення звітівDatazen.

Платформа даних:

– багатомірні й табличні аналітичні моделі(BISM)вSQL ServerAnalysis Services;

– засіб IAД–DMвSQL Server Analysis Services;

– компоненти для керування інформацією організацій(EnterpriseInformation Management, EIM) у складі:

– система ETL (extract, transform, load – витяг, перетворення й завантаження даних) – SQL Server Integration Services;

– система керування позначка-даними (або нормативно-довідковою інформацією) – SQL Server Master Data Services;

– система керування якістю даних – SQL Server Data Quality Services;

– система аналізу зв'язків між компонентами аналітичного рішення – проект «Barcelona».

– системи обробки складних подій (Complex Event Processing, CEP) – Azure Stream Analytics (ASA), SQL Server Streaminsight і інші;

– рішення для створення сховищ даних:

– реляційна БД SQL Server (традиційно – у власному центрі обробки даних компанії; бажане на базі рекомендованих архітектур Fasttrack DW для десятків терабайт інформації);

– віртуальні машини з SQL Server для зберігання даних в «хмарі» (SQL Server for Data Warehousing Azure VM);

– програмно-апаратні комплекси для завдань створення сховищ даних (SQL Server Parallel Data Warehouse);

– у якості проміжного сховища – лінійно масштабовані сховища напів-структурованих і неструктурованих даних («більші дані» – петабайти інформації) у рішенні Hdinsight (Hadoop для Windows) і засобу побудови звітності на підставі цієї інформації.

Інструменти розробника:

– SQL Server Data Tools (SSDT) – спеціальна версія Visual Studio, яка поставляється з SQL Server безкоштовно й дозволяє створювати рішення з пакетами інтеграції, аналітичними кубами й моделями ІАД, звітами Reporting Services.

– Performancepoint Dashboards Designer–інструмент для створення картпоказників і інтерактивних діаграм на порталі Sharepoint.

Microsoft BI можна інтегрувати з БД інших вендорів, такими як:

– Oracle;

– SAP R/3 і SAP Netweaver BI (BW);

– SAS і ін.

Microsoft BI можна розширити партнерськими рішеннями в частині:

– інструментів аналізу й відображення інформації;

– Panorama-Novaview;

– Iconics.

– відображення картографічної інформації;

– IDV Solutions Visual Fusion.

Аналітичні служби Microsoft SQL Server дозволяють працювати з будь-якими реляційними даними, доступними за допомогою OLE DB, містять два алгоритми DM і можуть також використовувати алгоритми DM, розроблені сторонніми виробниками. Клієнтські засоби можуть звертатися до даних, що зберігаються на цих серверах, за допомогою OLE DB. Слід зазначити, що є можливість звертатися до аналітичних служб із додатків Microsoft Office останніх трьох версій, а також створювати клієнтські додатки для аналітичних служб на його основі.

Розглянемо реалізацію засобів ІАД у СКБД Microsoft SQL Server. Завдання ІАД вирішуються за допомогою служб Analysis Services. На рис. 5 схематично представлені компоненти СКБД MS SQL Server і виділена підсистема ІАД [15].

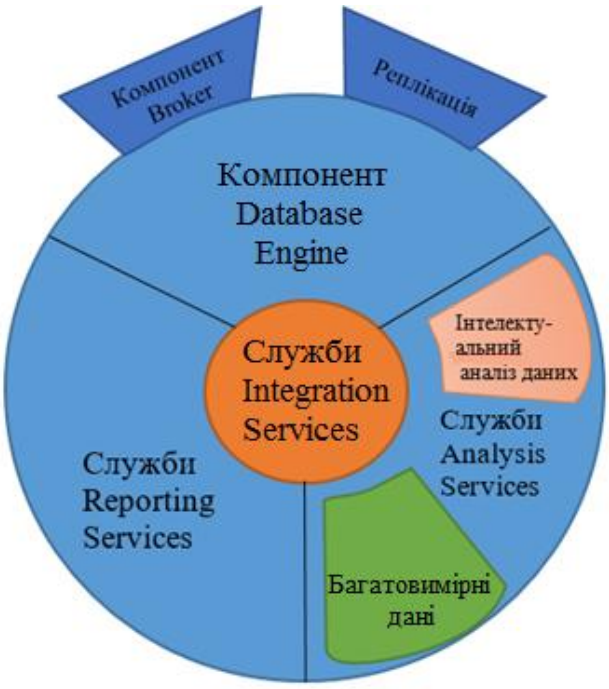

Рисунок 5 – Служби й компоненти СКБД Microsoft SQL Server

Служби Analysis Services надають наступні функції й засоби для створення рішень по ІАД:

– набір стандартних алгоритмів ІАД;

– конструктор ІАД, призначений для створення й перегляду моделей ІАД,

керування ними й побудови прогнозів;

– мова розширень ІАД (DM extensions to SQL, DMX).

Для роботи з надаваними засобами ІАД використовується середовище BI Development Studio, скорочено BI Devstudio.

Структура ІАД може бути представлена як сукупність вихідних даних і опису способів їх обробки. Структура містить моделі, які використовуються для аналізу її даних. Зокрема, одна структура може підтримувати кілька моделей. У структурі ІАД можна виділити навчальний і перевірочний набір даних, задавши відсоткове відношення або обсяг даних.

Модель ІАД являє собою комбінацію самих даних, алгоритму ІАД і колекції значень параметрів і фільтрів, керуючих використанням і обробкою даних. Модель ІАД визначається мовою розширень ІАД або за допомогою майстра ІАД у середовищі BI Devstudio.

Алгоритм ІАД являє собою механізм, що створює модель ІАД. Щоб створити модель, алгоритм спочатку аналізує набір даних, здійснюючи пошук певних закономірностей і трендів. Алгоритм використовує результати цього аналізу для визначення параметрів моделі ІАД. Потім ці параметри застосовуються до всього набору даних, щоб виявити придатні до використання закономірності й одержати докладну статистику. Нижче перераховані алгоритми ІАД, реалізовані в Microsoft SQL Server:

– спрощений алгоритм Байєса – Microsoft Naive Bayes;

– алгоритм дерева прийняття рішень – Microsoft Decision Trees;

– алгоритм тимчасових рядів – Microsoft Time Series;

– алгоритм кластеризації – Microsoft Clustering;

– алгоритм кластеризації послідовностей – Microsoft Sequence Clustering;

– алгоритм взаємозв'язків – Microsoft Association Rules;

– алгоритм нейронної мережі – Microsoft Neural Network;

– алгоритм лінійної регресії – Microsoft Linear Regression;

– алгоритм логістичної регресії – Microsoft Logistic Regression.

2.3 Засоби аналізу даних СКБД Oracle

Bi-засоби Oracle засновані на Oracle OLAP – засобах аналітичної обробки даних, вбудованих безпосередньо в реляційну СКБД Oracle. Крім власне OLAP-Сховища, Oracle надає кошти DM, інструменти створення звітів, доставки результатів аналізу за допомогою порталу, а також ряд засобів, що дозволяють створювати аналітичні Java-Додатки, зокрема, Java OLAP API і компоненти OLAP Beans, призначених для використання в засобах розробки Java-Додатків [33].

Крім OLAP-сервера, засобів доступу до OLAP-даних і засобів створення Bi-додатків, Oracle поставляє ряд готових аналітичних рішень на їхній основі.

Компанія Oracle робить великий спектр програмних продуктів. Це й готові додатки (Oracle E-Business Suite), і сервера додатків і засоби для колективної роботи й різні СКБД. До теперішнього часу розроблено кілька версій систем, кожна з яких включає цілу лінійку продуктів, наприклад, Oracle 8, Oracle 9i, Oracle 10g. Відповідні лінійки продуктів включають як власне СКБД (наприклад, Oracle Database 10g, Oracle Database 11g), так і засоби розробки й аналізу даних.

Сімейство продуктів BI містить набір продуктів для побудови оперативних аналітичних систем (OLAP), корпоративної звітності, виконання нерегламентованих запитів до БД Oracle, побудови простих інтерфейсів для керівників, що ухвалюють рішення (dashboard). За допомогою цих продуктів можна швидко створювати регламентовані й нерегламентовані запити, складні звіти, інтерфейси для аналітиків.

Алгоритми ІАД, що реалізовані в Oracle DM наведено в таблиці 1.

Таблиця 1 – Алгоритми, реалізовані в Oracle DM

Метод	Реалізовані алгоритми
Класифікаційні моделі	NAve Bayes, Adaptive Bayes Network
Класифікації й регресійні моделі	Support Vector Machine
Пошук істотних атрибутів	Minimal Descriptor Length
Кластеризація	Enhanced K-means, O-cluster
Пошук асоціацій	Apriory Algorithm
Виділення ознак	Non-Negative Matrix Factorization

Важлива особливість алгоритмів полягає в тому, що всі вони працюють безпосередньо з реляційними БД і не вимагають вивантаження й збереження даних у спеціальних форматах. Крім власне алгоритмів, в опцію Oracle Data Miner (ODM) входять засоби підготовки даних, оцінки результатів, застосування моделей до нових наборів даних. Використовувати всі ці можливості можна як на програмному рівні за допомогою Java API або PL/SQL API, так і за допомогою графічного середовища ODM, орієнтованого на роботу аналітиків, що вирішують завдання прогнозування, виявлення тенденцій, сегментації й ін.

Крім того, Oracle представляє сервіс BI Cloud Service [24], що дозволяє аналізувати дані з різних джерел, включаючи додаток Oracle, розгорнутий в «хмарі» або безпосередньо на підприємстві, щоб швидко створити функціонально насичені, інтерактивні аналітичні додатки. Клієнти можуть одержувати інформацію й аналізувати її в будь-який час, у будь-якому місці з мобільних обладнань. Простий інтерактивний користувацький інтерфейс із вбудованими підказками прискорює освоєння продукту й підвищує продуктивність. Користувачі з навичками роботи з Oracle BI або Oracle Cloud Applications можуть почати використовувати сервіс без додаткового навчання.

Існує цілий ряд програмних систем, що забезпечують ІАД. Наприклад, SAP BI, КРОК, Business Objects, Cognos, Information Builders, SAS, BusinessQ, Deductor.

Під SAP BI розуміють комплекс аналітичних додатків [37], що забезпечують багатомірний аналіз результатів діяльності для прогнозування й планування майбутніх показників бізнесу. Рішення включає інструментарій для створення й публікації звітів, що настроюються і додатків, які забезпечують якісно новий рівень поінформованості осіб, що ухвалюють рішення (ЛПР). Системою надаються інструменти й готові моделі для аналізу інформації й складання звітності, що забезпечує комплексний підхід до обґрунтування для ухвалення рішення.

Рішення SAP BI – система бізнес-аналізу, що дозволяє здійснювати стратегічний аналіз даних і підтримку процесу прийняття управлінських рішень у компанії. Призначена для надання доступу й обробки інформації, що втримується в різних системах або БД організацій, її аналізу. У якості джерел можуть виступати будь-які інформаційні системи, бухгалтерські й фінансові програми, спеціалізовані галузеві рішення, а також локальні джерела (наприклад, Excel або Access файли).

Основні переваги SAP BI:

– комплексне інтегроване рішення для інтелектуального аналізу даних, що надходять із різних джерел, звітності й обміну інформацією в єдиному середовищі;

– простий і зрозумілий інтерфейс, легке завантаження даних, багаті можливості візуалізації, потужний функціонал пошуку, аналізу й зіставлення даних з різних джерел;

– ліквідація витрат на створення, розробку й тестування звітів, підвищення ефективності побудови запитів і скорочення часу на складання звітів;

– можливість масштабування рішення в міру росту бізнесу, зниження витрат на інформаційну інфраструктуру й загальної вартості володіння.

В залежності від цілей впровадження рішення SAP BI, ним можуть користуватися різні співробітники, яким для роботи необхідна інформація, що зберігається в системах, – співробітники комерційного відділу, відділів IT, маркетингу, бухгалтерії й ін.

3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

3.1 Інтелектуальний аналіз даних на основі прецедентів

Цикл міркування (навчання) на основі прецедентів – CBR-методи включають чотири основні етапи, що утворюють так званий CBR цикл [21] або цикл навчання по прецедентах (прикладам), структура якого представлена на рис 6.

Основними етапами CBR циклу є:

− витяг найбільш відповідного (подібного) прецеденту (або прецедентів) для сформованої ситуації із БП;

− повторне використання витягнутого прецеденту для спроби рішення поточної проблеми;

− адаптація й застосування отриманого рішення для рішення поточної проблеми;

− збереження знову ухваленого рішення як частини нового прецеденту.

Інформація про нову проблемну ситуацію використовується для витягу із БП найбільш відповідного прецеденту (прецедентів). Витягнутий прецедент використовується повторно для одержання рішення нової проблеми (завдання). Потім запропоноване рішення, якщо буде потреба, може бути адаптоване до особливостей нової ситуації й застосоване на практиці. У випадку успішного застосування, перевірене рішення разом з описом проблемної ситуації утворює новий прецедент, який зберігається в БП. Таким чином, системою накопичується досвід (прецеденти) і реалізується машинне навчання. В CBR циклі може використовуватися не тільки БП, але й узагальнені знання предметної області для підтримки процесу міркування на основі прецедентів. Ця підтримка може бути слабкою або сильною, а може й взагалі бути відсутньою.

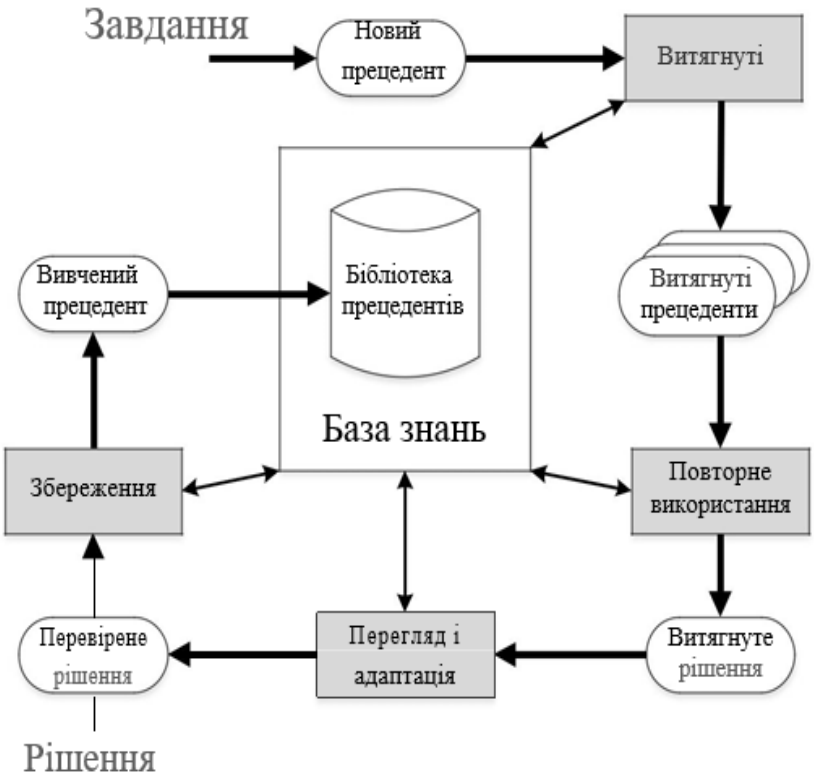

Рисунок 6 – Структура CBR циклу

Як правило, в CBR циклі на різних етапах потрібне залучення людини (ЛПР). Якщо процес витягу прецедентів може виконуватися автоматично, то для процесу адаптації й повторного використання прецедентів може знадобитися участь ЛПР.

В стандартному CBR циклі інформація про нову проблемну ситуацію використовується для витягу із БП найбільш відповідного прецеденту (прецедентів). Витягнутий прецедент використовується повторно для одержання рішення нової проблеми (завдання). Потім запропоноване рішення може бути адаптоване до особливостей нової ситуації й у випадку успішного застосування, новий прецедент зберігається в БП.

Для роботи модифікованого CBR циклу особливо в завданнях класифікації можуть застосовуватися тестові вибірки із прикладами для перевірки коректності знайденого рішення (рис. 7).

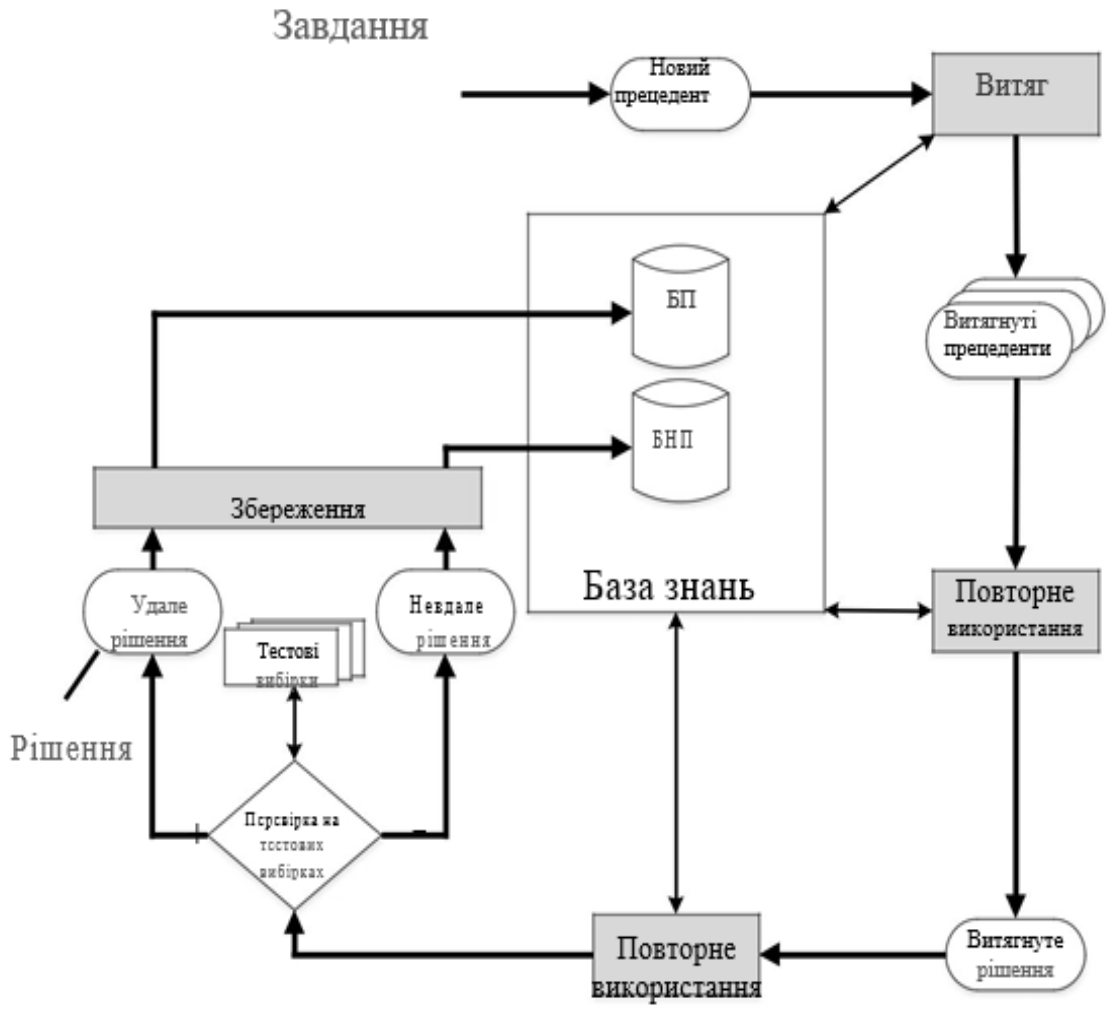

Рисунок 7 – Модифікований CBR цикл

Вдалим є прецедент, який не погіршує якість роботи (класифікації) CBR системи після його додавання в БП, а невдалим прецедентом будемо називати прецедент, який погіршує якість роботи (класифікації) CBR системи після його додавання в БП.

Представлення прецедентів у вигляді експертних правил продукційного типу – такий спосіб є найбільш зрозумілим і популярним методом представлення прецедентів. Правила забезпечують формальний спосіб представлення рекомендацій, знань або стратегій. Вони частіше підходять у тих випадках, коли предметна область виникає з емпіричних асоціацій, накопичених за роки роботи з рішення завдань у даній області. У системах, заснованих на правилах, предметні

знання представляються набором правил, які перевіряються на групі фактів і знань про поточну ситуацію (вхідної інформації). Коли частина правила ЯКЩО задовольняє фактам, то дії, зазначені в частині ТО, виконується. Коли це відбувається, то говорять, що правило спрацьовує.

Інтерпретатор правил зіставляє частини правил ЯКЩО з фактами й виконує, те правило, частина ЯКЩО якого відповідає фактам, тобто інтерпретатор правил працює в циклі «зіставити – виконати», формуючи послідовність дій.

Представлення прецедентів у структурованій формі – до такого представлення можна віднести дерева, графи, семантичні мережі. Один з методів – це концептуальний граф (conceptual graph) – це кінцевий, зв'язаний, двочастковий, орієнтований мультиграф. Вузли графа представляють поняття, або концептуальні відносини. У концептуальних графах мітки дуг не використовуються. Відносини між поняттями представляються вузлами концептуальних відносин. На рис. 8 вузли b1h, a2a (мітки правил) і 19, 56, 47, 9 (номера діагнозів) представляють поняття, а After, Diagnosisthislable – концептуальні відносини.

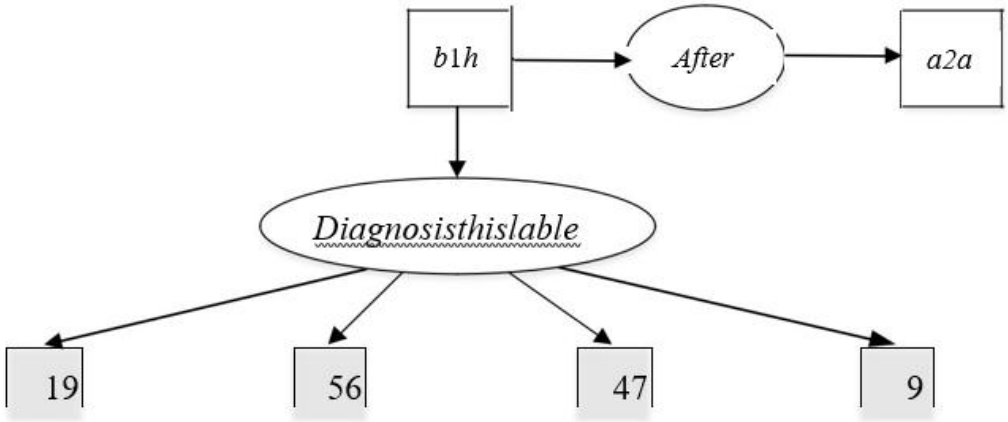

Рисунок 8 – Приклад концептуального графа

В концептуальних графах вузли понять представляють або конкретні, або абстрактні об'єкти предметної області. Вузли ж концептуальних відносин описують відносини, що включають одне або кілька понять. Однією з переваг концептуальних

графів без використання позначених дуг є простота представлення відносин будь-якої арності. N-Арне відношення представляється вузлом концептуального відношення, що має N дуг [59].

Кожний концептуальний граф представляє одне висловлення. Типова БЗ буде складатися з ряду таких графів. Графи можуть бути довільної складності, але вони повинні бути кінцевими.

В якості зберігання такої інформації може бути використаний текстовий файл, кожний рядок якого описує конкретний концептуальний граф, де першим словом є концептуальне відношення, другим – концептуальне поняття, що має вихідну дугу в це відношення, інші – концептуальні дуги, що мають вхідні дуги із цього відношення.

3.2 Витяг прецедентів

Якість роботи CBR системи при рішенні складних завдань прямо залежить від кількості прецедентів, що втримуються в БП. На відміну від пошуку в БД, де визначається конкретне значення в записах, пошук прецедентів повинен здійснюватися в умовах часткового збігу значень, тому що може не існувати прецеденту, що повністю збігається з поточним. У цьому випадку застосовуються спеціальні методи, що використовують метричні алгоритми й різні евристики.

На першому етапі CBR циклу (витяг прецедентів) виконується визначення ступеня подібності поточної ситуації із прецедентами із БП системи і наступний їхній витяг з метою дозволити нову проблемну ситуацію, що склалася на об'єкті.

Витяг прецедентів прямо зв'язаний зі способом представлення прецедентів і відповідно зі способом організації БП.

БП є важливою складовою БЗ ІС, але може виступати як окремий компонент

системи. Таким чином, структура БП впливає на різні показники роботи системи й, зокрема, на час пошуку й витягу прецедентів.

Для успішної реалізації CBR систем необхідно забезпечити коректний витяг прецедентів із БП системи.

Існують різні методи витягу прецедентів із БП системи [2]:

− метод найближчого сусіда і його модифікації;

− метод пошуку на деревах рішень;

− метод витягу на основі знань;

− метод витягу з урахуванням застосовності прецедентів.

Метод найближчого сусіда – це найпоширеніший метод порівняння й витягу прецедентів. Він дозволяє досить легко обчислити ступінь подібності поточної проблемної ситуації й прецедентів із БП системи по кожному параметру, використовуваному для опису прецедентів і поточної ситуації. З метою визначення ступеню подібності вводиться метрика (наприклад, засіб подібності Хэммінга) на просторі всіх параметрів. У цьому просторі визначається крапка, що відповідає поточній проблемній ситуації, і відповідно до обраної метрики визначається найближча до неї крапка (найближчий сусід – прецедент, який має максимальний ступінь подібності з поточною ситуацією) із крапок, що представляють прецеденти із БП. Для обраної метрики (засіб подібності Хэммінга) по методу найближчого сусіда ступінь подібності прецеденту й поточної проблемної ситуації обчислюється виходячи з того, що при збігу всіх параметрів в описі прецеденту й поточної ситуації ступінь подібності буде рівним 1, а кожний параметр, що збігся, дає внесок рівний 1/n, де n – число параметрів в описі прецеденту й поточної ситуації.

Метод визначення найближчого сусіда (найближчих сусідів) застосовується для рішення завдань класифікації, кластеризації, регресії й розпізнавання образів. На рис. 9 наведений алгоритм витягу прецедентів (алгоритм 1) [1]. Вхідні дані: T – поточна ситуація; CB – непуста множина прецедентів (БП); m – кількість розглянутих прецедентів із БП; S(C, T) – задана метрика (захід подібності); H –

граничне значення ступеня подібності.

Вихідні дані: Множина витягнутих прецедентів SC. Проміжні дані: j.

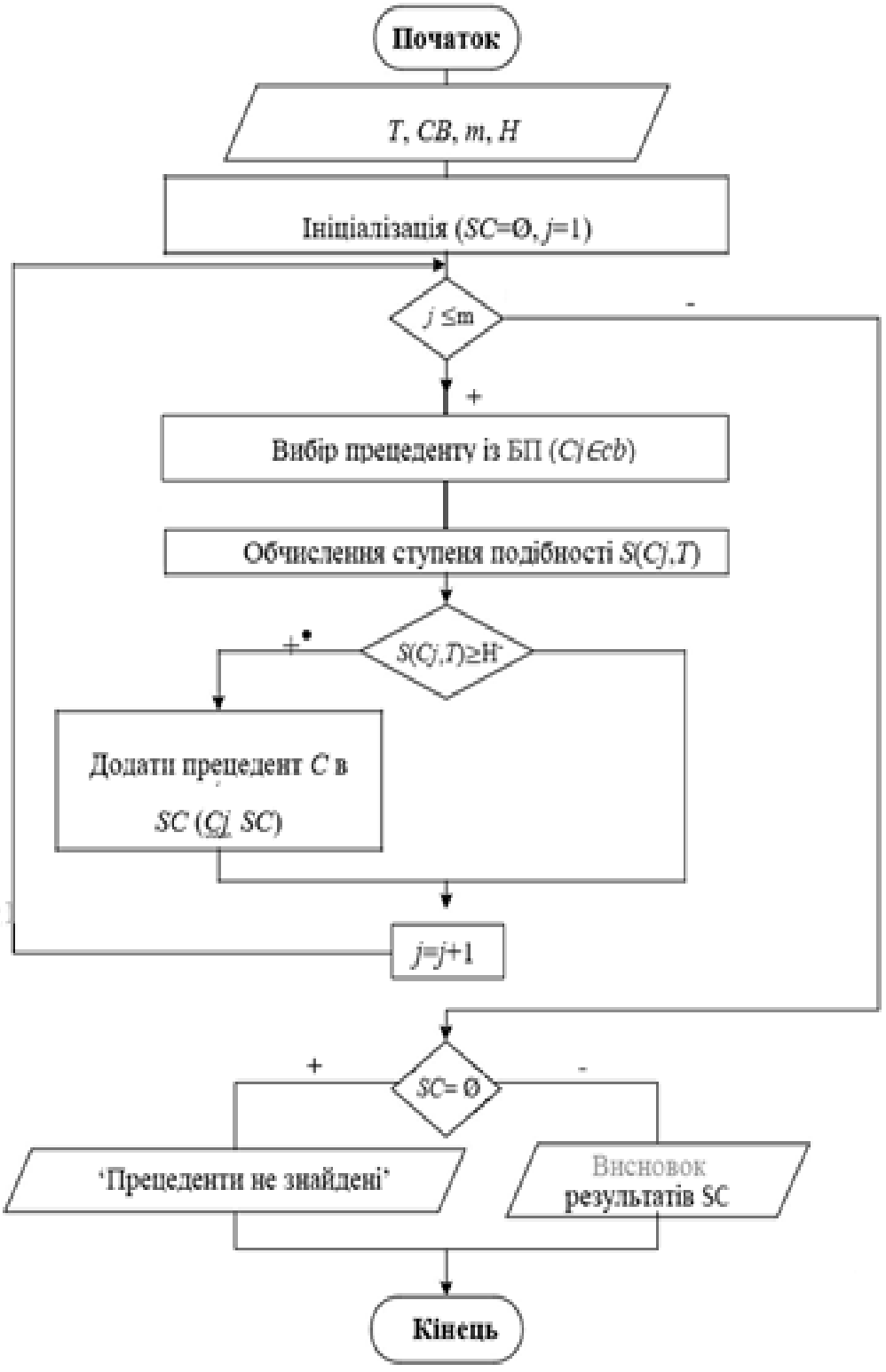

Рисунок 9 – Схема алгоритму для витягу прецедентів

3.3 Модифікація алгоритму витягу прецедентів

У роботі запропонована модифікація, яка полягає в тому, що k будуть змінюватися залежно від зміни розміру бази прецедентів (БП). Ніж більше прецедентів у БП, тем більше значення можна вибрати для k (від 1 до kmax). kmax відповідає кількості елементів, що належать класу з максимальним числом прецедентів.

У роботі пропонується вибирати k як найближче ціле число до середнього арифметичного значення між 1 і kmin (kavg=(1+kmin)/2), де kmin – кількість елементів, що належать класу з мінімальним числом прецедентів.

На рис. 10 наведена блок-схема запропонованого алгоритму витягу прецедентів на основі k-Nn (алгоритм 2). Вхідні дані:

– T – поточна ситуація;

– CB – непуста безліч прецедентів (БП);

– m – кількість розглянутих прецедентів із БП; S(C, T) – задана метрика (захід подібності);

– H – граничне значення ступеня подібності.

Вихідні дані: Безліч витягнутих прецедентів SC.

Проміжні дані:

– i, j, l – параметри циклів;

– kavg – кількість найближчих сусідів (прецедентів) для визначення рішення (приналежності класу);

– kmin – мінімальна кількість прецедентів, що мають однакове рішення (тобто приналежних одному класу).

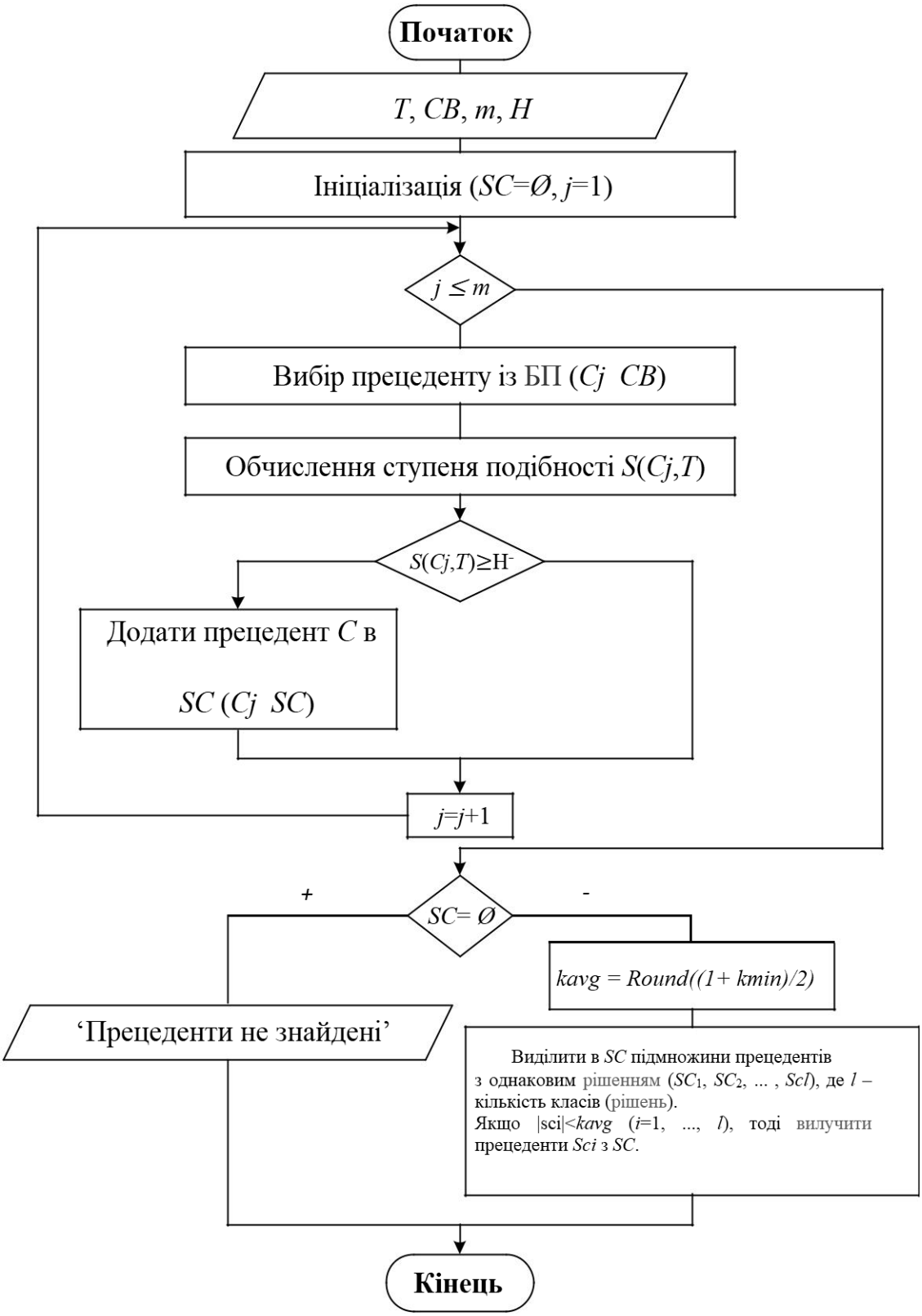

Рис унок 10 – Схема модифікованого алгоритму витягу прецедентів на основі k-Nn

При повторному використанні знайденого прецеденту в контексті нової

проблемної ситуації істотними є наступні моменти: відмінність між витягнутим і новим прецедентом, а також те, яку частину витягнутого прецеденту можна перенести на поточну ситуацію.

У простих завданнях класифікації відмінності просто ігноруються і клас рішення витягнутого прецеденту переноситися на клас рішення нового прецеденту. Це найпростіший спосіб повторного використання прецедентів.

Багато систем, тим не менш, ураховують відмінності між знайденим і наявним прецедентом, і тому рішення витягнутого прецеденту не може бути безпосередньо перенесене на нову ситуацію. Це рішення вимагає адаптації до поточної проблеми з урахуванням відмінностей між прецедентами.

Існує два методи повторного використання прецедентів: використання рішення знайденого прецеденту (трансформаційне використання) і використання методу, за допомогою якого було отримано це рішення (деривaційне використання).

4 ОПИС РОЗРОБЛЕНОЇ ПРОГРАМНОЇ СИСТЕМИ

4.1 Архітектура прототипу CBR системи

Архітектура прототипу CBR системи складається з наступних основних компонентів (рис. 11):

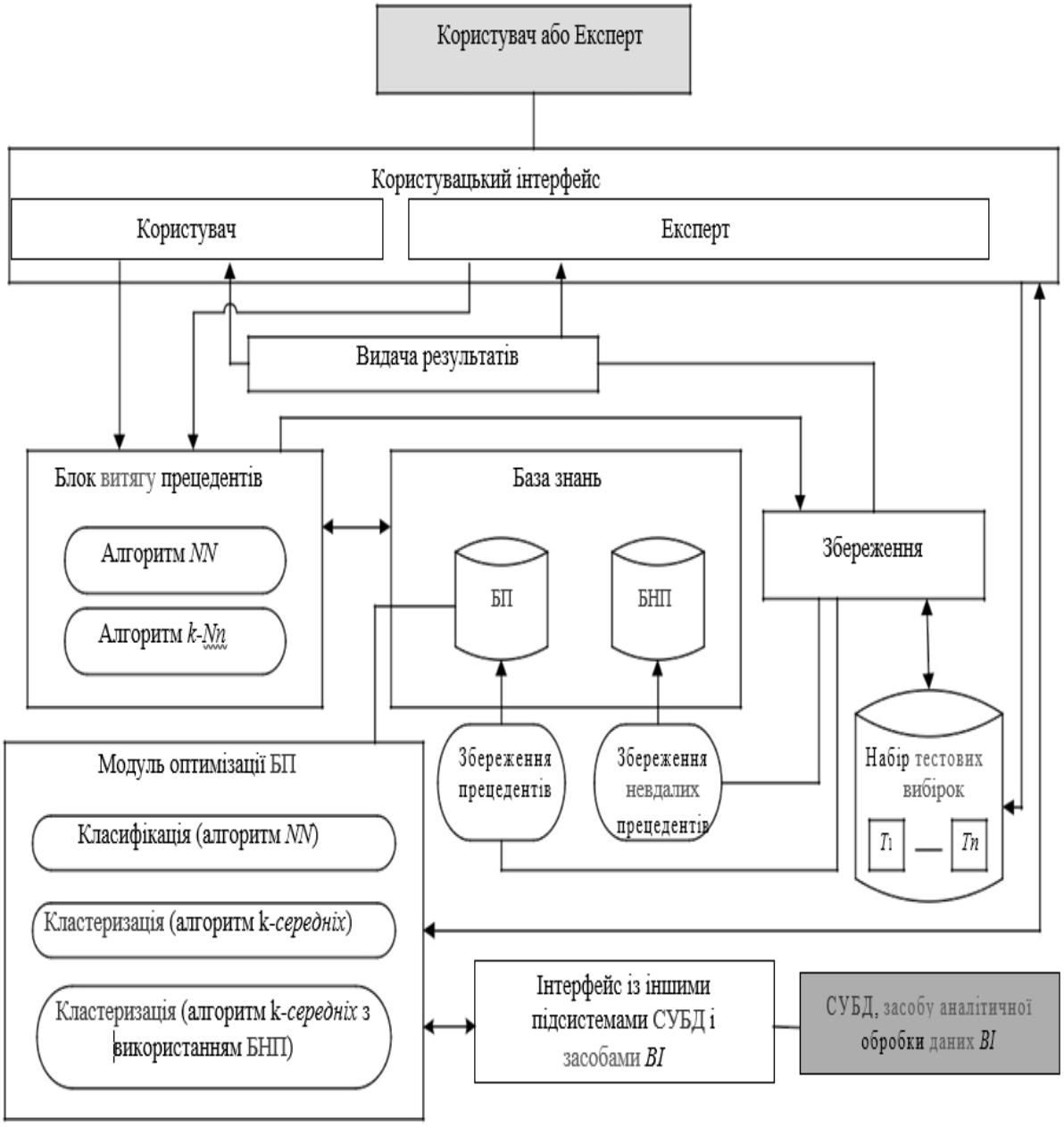

Рисунок 11 – Архітектура прототипу CBR системи

Для функціонування БП CBR системи використовується БД, структура якої

наведена на рис. 12.

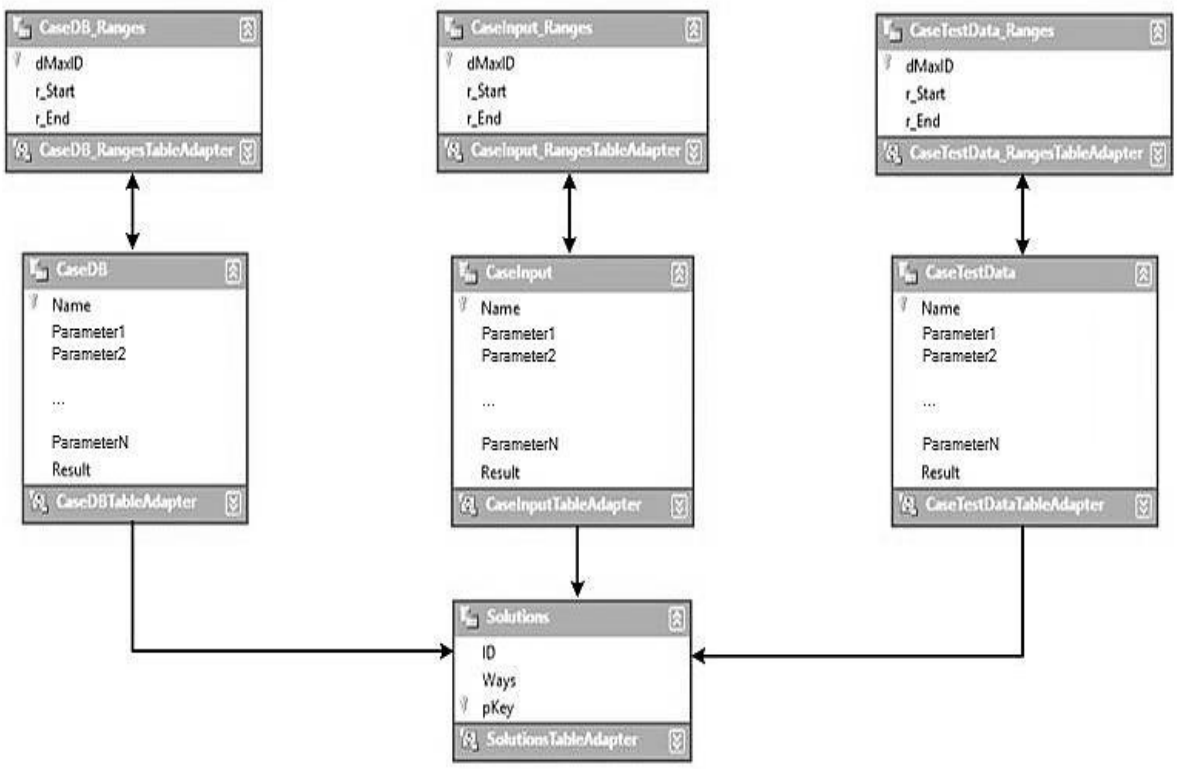

Рисунок 12 – Схема БД для CBR системи

Програмна реалізація прототипу CBR системи виконана мовою C#. У середовищу програмування MS Visual Studio для СКБД Microsoft SQL Server з використанням SQL Server Analysis Services і аналітичної платформи Deductor.

Прототип CBR системи реалізований у середовищі Microsoft Visual Studio 2010 з використанням наступних технологій і аналітичних платформ:

− Windows Forms;

− ADO.NET Entity Framework;

− аналітична платформа Deductor;

− SQL Server Analysis Services.

Deductor є аналітичною платформою Basegroup Labs, що концентрує багаторічний досвід компанії, що увібрав у себе найвдаліші архітектурні ідеї й

сучасний математичний апарат. Deductor є платформою, на базі якої створюються закінчені аналітичні рішення [11]. Платформа орієнтована на застосування експертами в різних предметних областях, дозволяє обробляти будь-яку структуровану табличну інформацію. Це доступна за ціною й проста у використанні система із прекрасними аналітичними можливостями.

Deductor надає аналітикам інструментальні засоби, необхідні для рішення найрізноманітніших аналітичних завдань: корпоративна звітність, прогнозування, сегментація, пошук закономірностей – ці й інші завдання, де застосовуються такі методики аналізу, як OLAP, KDD і DM. Deductor є ідеальною платформою для створення систем підтримки прийняття рішень.

Реалізовані в Deductor технології можуть використовуватися як у комплексі, так і окремо для рішення широкого спектра бізнес-проблем:

У роботі аналітична платформа Deductor 5.3 була використана для класифікації й кластеризації вже накопичених прецедентів у БП, а також для навчання ІНС і побудови дерева рішень на основі прецедентів.

Microsoft SQL Server Analysis Services (SSAS) забезпечують інтерактивну аналітичну обробку (OLAP) і функції ІАД для додатків бізнес-аналітики. Служби Analysis Services підтримують OLAP, дозволяючи розробляти й створювати багатомірні структури, які містять дані, зібрані з інших джерел, таких як реляційні БД, а також управляти цими структурами. Для додатків ІАД служби Analysis Services дозволяють розробляти, створювати й наочно представляти моделі ІАД, побудовані на основі інших джерел даних, використовуючи із цією метою широкий спектр стандартних алгоритмів ІАД.

У службах SSAS передбачені функції інтерактивної аналітичної обробки й ІАД для рішень в області бізнес-аналітики. Перш ніж приступитися до розробки рішення бізнес-аналітики за допомогою служб Analysis Services, слід ознайомитися з основними поняттями OLAP і ІАД, які необхідно знати для розробки ефективного рішення.

Служби Analysis Services поєднують у собі кращі аспекти традиційного аналізу на основі OLAP і реляційної звітності, дозволяючи розроблювачам визначати одну модель даних, іменовану уніфікованою багатомірною моделлю (UDM) [11], для одного або декількох фізичних джерел даних. Усі запити кінцевого користувача з OLAP, зі звітів і з користувацьких додатків бізнес-аналітики одержують доступ до даних у базових джерелах даних за допомогою уніфікованої багатомірної моделі, що забезпечує єдине бізнес-представлення таких реляційних даних.

Служби Analysis Services надають великий набір алгоритмів ІАД, які дозволяють бізнес-користувачам виконувати ІАД з метою виявлення певних закономірностей і трендів. Такі алгоритми ІАД можуть використовуватися для аналізу даних за допомогою уніфікованої багатомірної моделі або безпосередньо з фізичного сховища даних.

SSAS використовують як серверні, так і клієнтські компоненти для надання додаткам бізнес-аналітики функцій оперативної аналітичної обробки й ІАД [14]:

Серверний компонент служб Analysis Services реалізований у вигляді служби Microsoft Windows. Служби SQL Server Analysis Services підтримують роботу декількох екземплярів на одному комп'ютері, при цьому кожний екземпляр служб Analysis Services реалізований як окремий екземпляр служби Windows.

Клієнти зв'язуються зі службами Analysis Services, які розглядаються як веб-служби, за допомогою загальнодоступного стандарту XML для аналітики (XMLA) [13] – протоколу, заснованого на SOAP [14], для виконання команд і прийняття відповідей. Моделі клієнтських об'єктів також надаються через XMLA, і одержати доступ до них можна за допомогою керованого постачальника, наприклад, ADOMD.NET, або за допомогою власних постачальників даних OLE DB.

Команди запиту можуть виконуватися за допомогою наступних мов: SQL; багатомірних виразів – мови запитів галузевого стандарту, орієнтованого на аналіз; розширень ІАД – мови запитів галузевого стандарту, орієнтованого на ІАД. Також мова сценаріїв служб Analysis Services (ASSL) можна використовувати для

керування об'єктами БД служб Analysis Services.

В екземплярі служб SSAS утримуються об'єкти БД і складання для використання з інтерактивною аналітичною обробкою й ІАД:

У БД утримуються об'єкти OLAP і ІАД, такі як джерела даних, представлення джерел даних, куби, заходи, групи заходів, атрибути, ієрархії, структури й моделі ІАД, а також ролі. У складаннях містяться користувацькі функції, що розширюють функціональність внутрішніх функцій, забезпечуваних мовами багатомірних виражень і розширеннями ІАД.

У роботі SSAS були використані для класифікації й кластеризації вже накопичених БП, а також для навчання ІНС і побудови дерева рішень на основі прецедентів із БП CBR системи.

4.2 Приклад використання прототипу CBR системи

Робота прототипу була розглянута на прикладі наборів даних з репозиторія UCI Machine Learning Repository Каліфорнійського університету [18]. БД із інформацією про рівень знань, студентів по дисципліні «Електричні машини постійного струму».

БД із репозиторієм включає 258 прикладів, що характеризуються 5 атрибутами (параметрами) і приналежних одному з 4 рішень (класів): 1 – дуже низький (very low), 2 – низький (low), 3 – середній (middle) і 4 – високий (high) (рис. 13).

Name	STG	SCG	STR	LPR	PEG	Result
P001	0	0	0	0	0	1
P002	0.08	0.08	0.1	0.24	0.9	4
P003	0.06	0.06	0.05	0.25	0.33	2
P004	0.1	0.1	0.15	0.65	0.3	3
P005	0.08	0.08	0.08	0.98	0.24	2
P006	0.09	0.15	0.4	0.1	0.66	3

Рисунок 13 – Приклад даних із БД UCI Machine Learning Repository

STG (The degree of study time for goal object materails) – частка навчального часу, витраченого для вивчення матеріалів по дисципліні (область значень даного параметра [0, 1]);

SCG (The degree of repetition number of user for goal object materails) – частка від кількості повторів при вивченні матеріалів по дисципліні (областьзначень даного параметра [0, 1]);

STR (The degree of study time of user for related objects with goal object) – частка навчального часу, витраченого для вивчення матеріалів по суміжних дисциплінах (область значень даного параметра [0, 1]);

LPR (The exam performance of user for related objects with goal object) – результати екзаменів із суміжних дисциплін (область значень даного параметра [0, 1]);

PEG (The exam performance of user for goal objects) – результати екзаменів з дисципліни (область значень даного параметра [0, 1]);

Result (The knowledge level of user) – рівень знань учня:

- дуже низький (very low);
- низький (low);
- середній (middle);
- високий (high).

На першому кроці необхідно завантажити початкову БП для даного прикладу (рис. 14). Початкова БП була сформована на основі перших 20 записів із БД UCI Machine Learning Repository, а також можна вибрати тестову й навчальну вибірки.

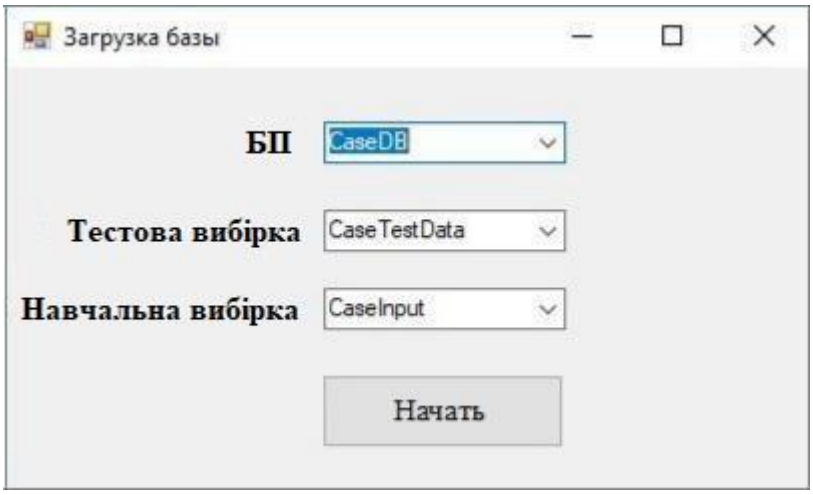

Рисунок 14 – Діалогове вікно для завантаження початкової БП

Далі можна задати нову ситуацію для пошуку рішення на основі прецедентів із БП (рис. 15).

Для прикладу була взята наступна ситуація:

– STG = 0.24;

– SCG = 0.1;

– STR = 0.25;

– LPR = 0.06;

– PEG = 0.9.

Далі натискається кнопку «Пошук» для одержання результатів (рис. 15).

На рис. 16 видно, що для поточної ситуації по алгоритму NN знайдений найближчий прецедент P011 зі ступенем подібності 84.44%, відповідно до яким для поточної ситуації знайдене рішення «4» – високий (high) рівень знань учня. За допомогою натискання кнопки «k-Nn» можна одержати результати по алгоритму k-Nn, які для даного прикладу також дають рішення «4» – високий (high) рівень знань учня, що навчається з усередненою оцінкою по найближчих прецедентах рівної 77.34%.

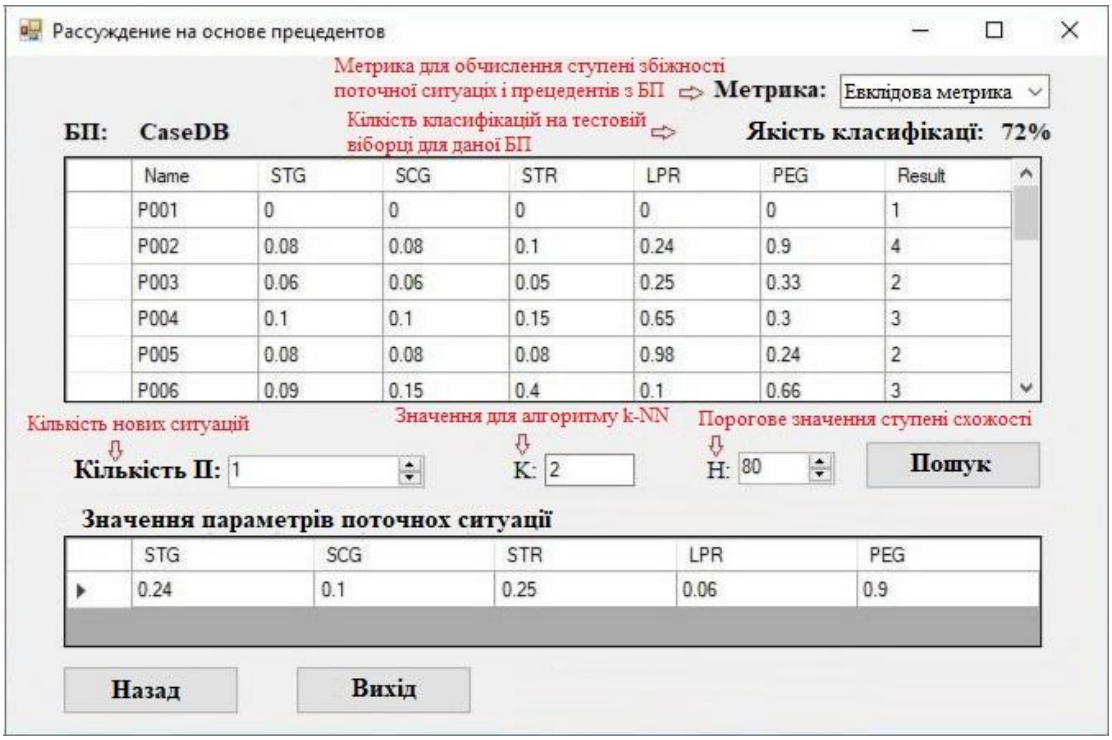

Рисунок 15 – Діалогове вікно для введення значень параметрів поточної ситуації

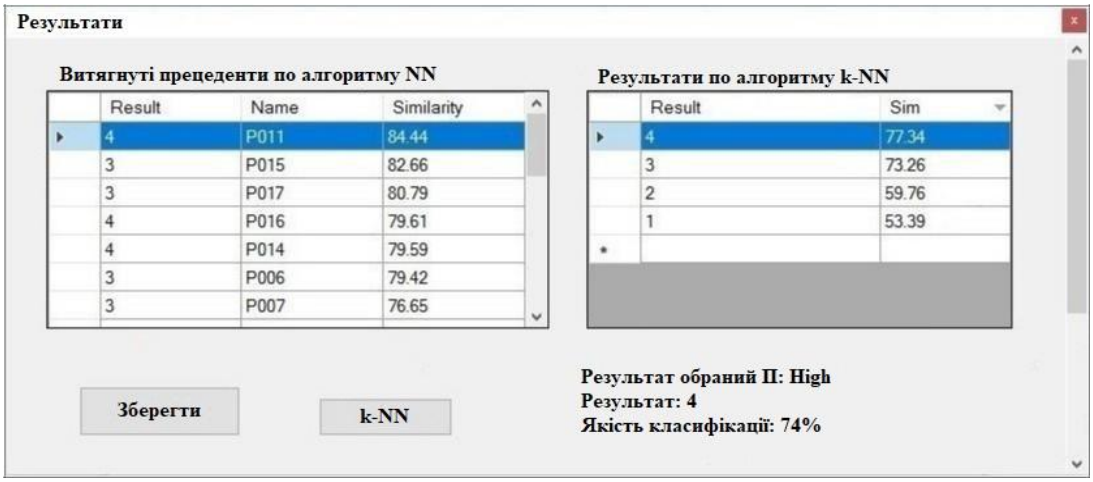

Рисунок 16 – Діалогове вікно з результатами

Для додавання прецеденту в БП необхідно нажати кнопку «Зберегти» і тоді програма занесе нову ситуацію як новий прецедент у БП. Якщо в систему завантажена тестова (експертна) вибірка, тоді перед збереженням буде виконана перевірка й у випадку, якщо новий прецедент не погіршує якість роботи CBR системи він буде доданий у БП, інакше новий прецедент потрапить у БНП.

Таким чином, для даного прикладу CBR система додасть новий прецедент у БП, класифікувавши дану ситуацію (рівень знань, що навчається) як «4» – високий (high). Зверніть увагу, що на мал. 19 і 20 видно, що при додаванні нового прецеденту зростає якість класифікації з 72% до 74%.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Робота алгоритму k-Nn була розглянута на попередньому прикладі із БД, що містить інформацію про рівень знань, що навчаються по дисципліні «Електричні машини постійного струму» і яка містить 258 записів. Початкова БП була сформована на основі перших 20 записів із БД. Для навчальної вибірки були використані наступні 238 записів, а для тестової вибірки були взяті всі приклади з тестового набору (http://archive.ics.uci.edu/ml/machine-learning-databases/00257/).

Для порівняння були обрані основні метрики: Евклідова, Мангеттенська, Чебишевська, Хемінгська й Журавльова.

На рис. 17 наведені результати по оцінці якості класифікації зі збільшенням прецедентів у БП і використанням різних метрик, які дозволяють зробити висновок, що для даного прикладу кращі результати по якості класифікації були досягнуті при використанні метрик Евкліда, Мангеттенської і Чебишева. Із цієї причини для виконання наступних обчислювальних експериментів з використанням зазначених вище наборів даних була обрана Евклідова метрика.

Робота модифікованого алгоритму k-Nn була розглянута на попередньому прикладі із БД, що містить інформацію про рівень знань, що навчаються по дисципліні «Електричні машини постійного струму». Для витягу прецедентів була обрана Евклідова метрика. У роботі запропонована модифікація, яка полягає в тому, що k будуть змінюватися залежно від розміру БП. Ніж більше прецедентів у БП, тем більше значення можна вибрати для k (від 1 до k_{max}). k_{max} відповідає кількості елементів, що належать класу з максимальним числом прецедентів.

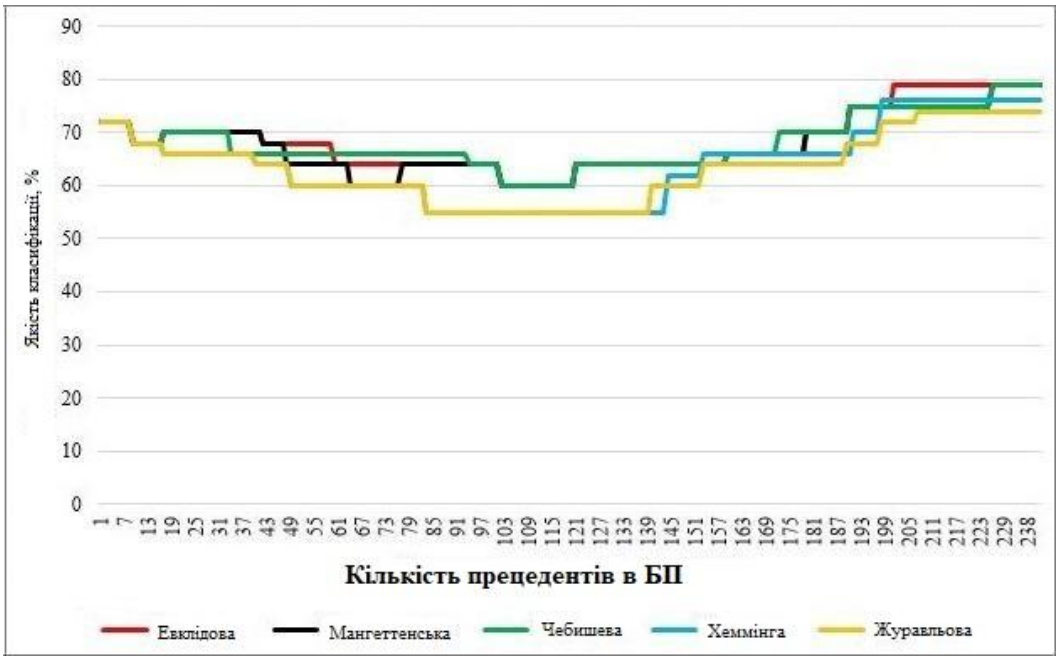

Рисунок 17 – Графік залежності якості класифікації від кількості прецедентів у БП із використанням різних метрик

У роботі пропонується вибирати k як найближче ціле число до середнього арифметичному значенню між 1 і kmin (kavg=(1+kmin)/2), де – кількість елементів, що належать класу з мінімальним числом прецедентів.

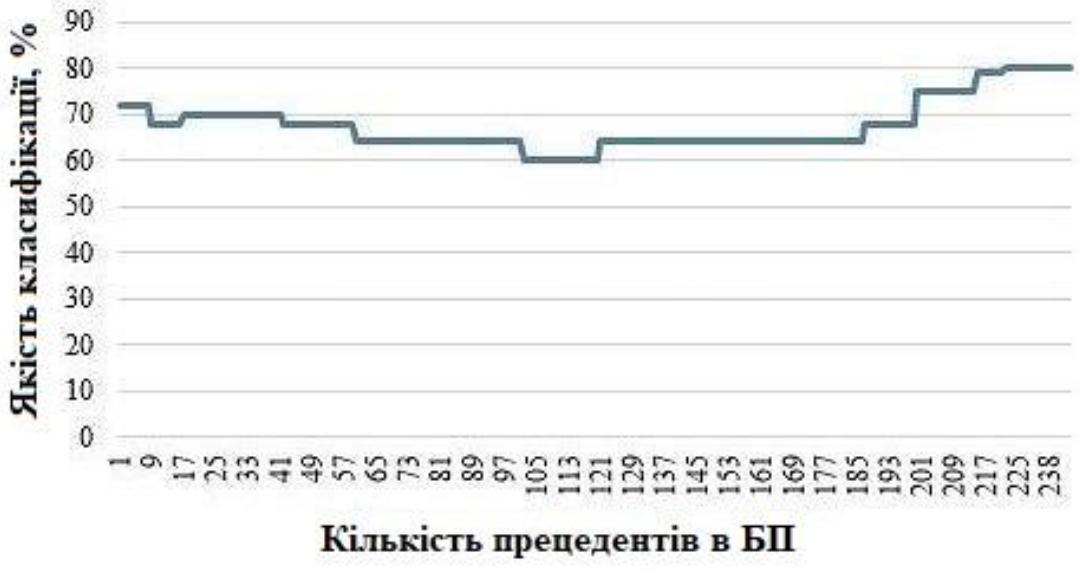

Рисунок 18 – Графік залежності якості класифікації від кількості прецедентів для k = 1

З інформації, представленої на рис. 19 видно, що з ростом прецедентів у БП показники якості класифікації вище при виборі середнього значення для k (kavg).

Слід зазначити, що можлива ситуація, коли сформувати узагальнену БП на основі тестових вибірок не вдається через суперечливість тестових (експертних) наборів.

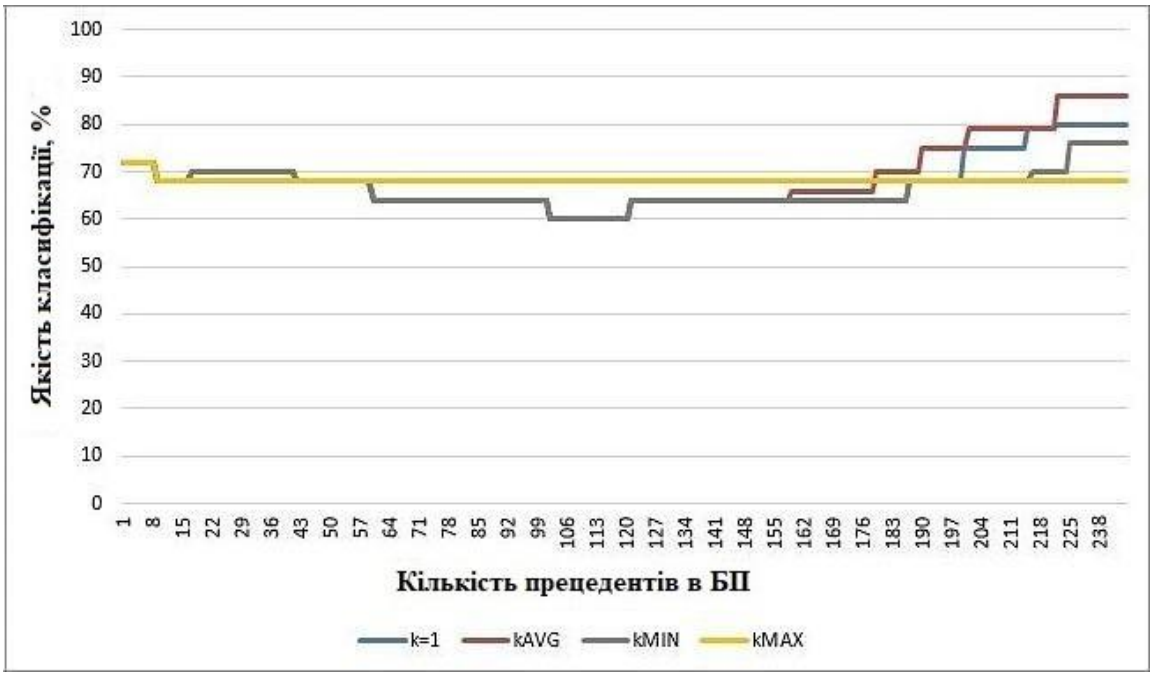

Рисунок 19 – Графік залежності якості класифікації від кількості прецедентів у БП при різних значеннях k

Таким чином, на основі отриманих результатів обчислювальних експериментів можна констатувати, що для скорочення кількості прецедентів у БП доцільно використовувати запропонований кластерний алгоритм 4 на основі k-середніх, тому що при істотнім скороченні обсягу БП якість класифікації погіршується не більше ніж на 8%, а при наявності БНП модифікований алгоритм k-середніх (алгоритм 5), який при істотнім скороченні обсягу БП знижує якість класифікації не більше ніж на 3-4%.

На основі отриманих результатів була підтверджена перспективність

можливості використання CBR методу для ІАД на початковому етапі при наявності тільки одиничних прикладів (прецедентів) для наступного нагромадження прецедентів і застосування інших методів ІАД (наприклад, методів узагальнення накопиченого досвіду (прецедентів) з використанням апарату дерев рішень і навчання ІНС на основі прецедентів із БП CBR системи).

ВИСНОВКИ

В атестаційній роботі магістра проведене дослідження різних технологій, методів і програмних засобів ІАД, що включаються до складу сучасних СКБД, і встановлене, щоодніїї з перспективних можливостей розширення засобів ІАД і аналітичних інструментів СКБД є використання прецедентного підходу.

Розроблена модифікація алгоритму витягу прецедентів на основі k-Nn для ІАД, що полягає в зміні значення k залежно від розміру БП. Дана модифікація дозволяє підвищити якість рішення завдань ІАД, зокрема, підвищити якість класифікації даних з використанням CBR методу.

Запропонований модифікований CBR цикл, що використовує експертну інформацію (тестові набори даних) для витягу прецедентів. Даний метод підвищує якість рішення завдань ІАД на основі прецедентів за рахунок формування бази вдалих (підходящих) і невдалих (невідповідних) прецедентів у процесі виконання CBR циклу.

У середовищі MS Visual Studio мовою C# з використанням MS SQL Server виконана програмна реалізація прототипу підсистеми ІАД на основі прецедентів і виконані обчислювальні експерименти для порівняння розроблених алгоритмів на наборах даних з UCI Repository, що підтвердили ефективність запропонованих у роботі методів і алгоритмів. Запропонована архітектура прототипу CBR системи для ІАД, що включає в себе наступні основні компоненти: користувацький інтерфейс, блок витягу прецедентів, БЗ із БП і БНП, набір тестових вибірок і модуль оптимізації БП для скорочення кількості прецедентів у БП CBR системи.

Виконана програмна реалізація прототипу CBR системи для розширення можливостей засобів ІАД у СКБД на прикладі Microsoft SQL Server з використанням мови C# і середовища програмування MS Visual Studio 2010, а також технології

Windows Forms, ADO.NET Entity Framework, аналітичної плаформы Deductor і SQL Server Analysis Services.

Розглянутий приклад використання розробленого прототипу системи для рішення завдання класифікації даних з репозиторію UCI Machine Learning Repository.

Виконані обчислювальні експерименти на різних наборах даних з репозиторію UCI Machine Learning Repository і Kaggledatasets (Pima Indians Diabetes Database і Human Resources Analytics) для порівняння алгоритмів витягу прецедентів і оцінки якості рішення завдання класифікації даних і швидкодії CBR системи при порівнянні алгоритмів скорочення кількості прецедентів у БП, що підтвердили ефективність запропонованого модифікованого алгоритму для витягу прецедентів і модифікованого CBR циклу, а також кластерних алгоритмів при скороченні кількості прецедентів у БП для підвищення швидкодії CBR системи.

Для підтвердження перспективності можливості використання CBR методу для ІАД на початковому етапі з наступним застосуванням інших методів ІАД (наприклад, методів узагальнення накопиченого досвіду (прецедентів) з використанням апарату дерев рішень і навчання ІНС на основі прецедентів із БП) наведені результати по оцінці якості класифікації даних навченої на прецедентах ІНС і дерева рішень, отриманого в результаті узагальнення прецедентів із БП.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЯ

1. Варшавский П.Р., Еремеев А.П. Моделирование рассуждений на основе прецедентов в интеллектуальных системах поддержки принятия решений // Искусственный интеллект и принятие решений. 2009. № 2. С. 45-47.
2. Варшавский П.Р., Еремеев А.П. Методы правдоподобных рассуждений на основе аналогий и прецедентов для интеллектуальных систем поддержки принятия решений// Новости искусственного интеллекта. – 2016. – №3. С. 39-62.
3. Ар Кар Мьо Исследование методов интеллектуального анализа данных на основе прецедентов // Радиоэлектроника, электротехника и энергетика: Двадцатая междунар. науч.-техн. конф. студентов и аспирантов: Тез. докл. В 4 т. Т. 2. М.: Издательский дом МЭИ, 2014. С. 340.
4. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта, №3, 2014, С. 3-19.
5. W. Frawley, G. Piatetsky-Shapiro, C. Matheus Knowledge Discovery in Databases: An Overview. – AI Magazine. – 1992. pp. 213-228.
6. Kitchin Rob. The Data Revolution. United States: Sage. 2014, p. 6.
7. Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Databases. USA: MIT Press, 1991.
8. Agrawal R., Mannila H., Srikant R., Toivonen H. and Verkamo I. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, Calif.: AAAI Press, 1996, pp. 307-328.
9. Fayyad U., Piatetsky-Shapiro G., Smyth P., Advances in Knowledge Discovery and Data Mining, (Chapter 1), AAAI/MIT Press, 1996.

10. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. // 2-еизд., – СПб: БХВ-Петербург, 2007.

11. А.А. Барсегян, И.И. Холод, М.Д. Тесс, М.С. Куприянов, С.И. Елизаров. Анализ данных и процессов. 3-е изд. — СПб.: БХВ-Петербург, 2009.

12. Интеллектуальный анализ данных средствами MS SQL Server 2008 – URL: http://www.intuit.ru/studies/courses/2312/612/lecture/13260: (дата звернення: 02.12.2019).

13. Data Mining – технология добычи данных – [Электронный ресурс]. URL:http://bourabai.ru/tpoi/datamining.htm: (дата звернення: 02.12.2019).

14. Дюк В.А., Самойленко А.П. Data Mining: учебный курс СПб.: Питер, 2001.

15. Чубукова И.А. Data Mining, БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий - ИНТУИТ.ру, 2006.

16. Филипов В.А. Интеллектуальный анализ данных: методы и средства. М.:Эдиториал УРСС, 2001.

17. Методы интеллектуального анализа данных – технология добычи данных – URL: http://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/: (дата звернення: 02.12.2019).

18. Judea Pearl, Stuart Russell. Bayesian Networks. UCLA Cognitive Systems Laboratory, Technical Report (R-277), November 2000.

19. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа - Methoden der Korrelation - und Regressiolynsanalyse. — М.: Финансы и статистика, 1981.

20. СУБДDB2–.URL:http://bourabai.ru/dbt/servers/Oracle.htm: (дата звернення: 02.12.2019).

21. Что такое Business Intelligence? Обзор BI систем –URL: http://www.clouderp.ru/tags/BUSINESS_INTELLIGENCE/: (дата звернення: 02.12.2019).

22. BusinessIntelligence. IT Term Definitions (англ.). Gartner (2011). – URL: http://www.webcitation.org/65AkPdIv3 /: (дата звернення: 02.12.2019).

23. Krzysztof J. Cios, Data Mining: A Knowledge Discovery Approach, Springer, 2017, ISBN 978-0-387-33333-5 – pp. 116-123.

24. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация. Т. 1: Пер. с англ. – М.: Издательский дом «Вильямс», 2011.

25. Parsaye K. OLAP and Data Mining: Bridging the Gap // Database Programming and Design. - 1997. - № 2. pp. 30-37.

26. Han J. OLAP Mining: An Integration of OLAP with Data Mining. - IFIP, 1997.

27. SAP Business Intelligence (SAP BI) – URL: http://www.tadviser.ru/index.php/Продукт:SAP_Business_Intelligence_(SAP_BI)(дата звернення: 02.12.2019).

28. Рынок платформ бизнес-аналитики – URL: https://www.bytemag.ru/articles/detail.php?ID=11919: (дата звернення: 02.12.2019).

29. Forecasting of Airfare Prices Using Time Series / I.Shubin, A.Gordiievich// Proceeding of 2015 Information Technologies in Information Business Conference (ITIB) 7 – 9 October, 2015, IEEE Catalog Number CFP15D13-PRT pp. 68-71