Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи аналізу даних для інтелектуальних систем з використанням машинного навчання

(тема)

Виконав:

студент II курсу, групи СПзм-22-1

Шмельова В.С.
(прізвище, ініціали)

Спеціальність

123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма

Системне програмування
(повна назва освітньої програми)

Керівник: доц. Федорченко В.М.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Шмельовій Валерії Сергіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Методи аналізу даних для інтелектуальних систем з використанням машинного навчання

затверджена наказом по університету від " 01 " квітня 2024 р. № 45 Стз

2. Термін подання студентом роботи до екзаменаційної комісії 15 червня 2024 р.

3. Вхідні дані до роботи
прецедент
СКБД
Інтелектуальний аналіз даних
Машинне навчання

4. Перелік питань, що потрібно опрацювати у роботі
Аналіз предметної області
ІАД з використанням прецедентів
Програмна реалізація ІАД

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 17 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Отримання завдання та аналіз літератури	01.04.2024 – 06.04.2024	
2	Огляд існуючих рішень та методів	07.04.2024 – 12.04.2024	
3	Розробка моделі	13.04.2024 – 18.04.2024	
4	Вибір програмних засобів	19.04.2024 – 25.04.2024	
5	Програмна реалізація	26.04.2024 – 02.05.2024	
6	Аналіз отриманих результатів	03.05.2024 – 16.05.2024	
7	Оформлення записки	17.05.2024 – 14.06.2024	
8	Представлення роботи в ЕК	15.06.2024	

Дата видачі завдання 01 квітня 2024 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Федорченко В.М.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 70 с., 18 рис., 1 дод., 21 джерело.

СИСТЕМА КЕРУВАННЯ БАЗАМИ ДАНИХ, БАЗА ЗНАНЬ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ШТУЧНА НЕЙРОННА МЕРЕЖА, DATA MINING.

Метою кваліфікаційної роботи *є* дослідження засобів та методів інтелектуального аналізу даних для систем керування базами даних.

У ході виконання кваліфікаційної роботи проведено аналіз засобів та методів інтелектуального аналізу даних на основі прецедентів для систем керування базами даних. Проведено дослідження різних технологій, методів і програмних засобів ІАД, що включаються до складу сучасних СКБД. Однією з перспективних можливостей розширення засобів ІАД і аналітичних інструментів СКБД є використання прецедентного підходу. Розроблено модифікацію алгоритму вилучення прецедентів на основі k-NN для ІАД, яка полягає в зміні значення k в залежності від розміру БП. Дана модифікація дозволяє підвищити якість рішення задач ІАД, зокрема, підвищити якість класифікації даних з використанням CBR методу.

ABSTRACT

Master’s thesis: 70 pages, 18 figures, 1 appendices, 21 sources.

DATABASE MANAGEMENT SYSTEM, KNOWLEDGE BASE, INTELLECTUAL DATA ANALYSIS, ARTIFICIAL NEURAL NETWORK, DATA MINING.

The purpose is to research means and methods of intelligent data analysis for database management systems.

In order to the means and methods of intelligent data analysis based on precedents for database management systems were analyzed. A study of various technologies, methods and software tools of IAD, which are included in the composition of modern DBMSs, was conducted. One of the promising opportunities for expanding IAD tools and DBMS analytical tools is the use of a precedent approach. A modification of the algorithm for extracting precedents based on k-NN for IAD has been developed, which consists in changing the value of k depending on the size of the BP. This modification makes it possible to improve the quality of solving IAD problems, in particular, to improve the quality of data classification using the CBR method.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ 7
ВСТУП 8
1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ 10
1.1 Терміни, що відносяться до ІАД 10
1.2 Виявлення знань 12
1.2.1 Процес виявлення знань. Етапи 12
1.2.2 Підготовка вхідних даних 14
1.3 Завдання ІАД 15
1.4 Існуючі методи ІАД 19
1.5 Програмні засоби для ІАД 27
2 ІАД З ВИКОРИСТАННЯМ ПРЕЦЕДЕНТІВ 34
2.1 Навчання на основі прецедентів 34
2.2 Модифікований CBR-цикл 35
2.2.1 Метод найближчого сусіда 39
2.2.2 Метод вилучення прецедентів на основі дерев рішень 41
2.2.4 Метод вилучення прецедентів з урахуванням їх застосовності 42
2.3 Повторне використання прецедентів 46
2.4 Збереження прецедентів 48
2.5 Етапи розробки CBR-систем 49
3 РЕАЛІЗАЦІЯ ПІДСИСТЕМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ 50
3.1 Програмна реалізація 53
ВИСНОВКИ 58
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ 59
ДОДАТОК А Графічний матеріал кваліфікаційної роботи 61

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних

БЗ – база знань

БП – база прецедентів

ІАД – інтелектуальний аналіз даних

ІС – інтелектуальна система

ЛПР – людина, що приймає рішення

СКБД – система керування базами даних

ШНМ – штучна нейронна мережа

BI – Business Intelligence

CBR – Case-Based Reasoning

DM – Data Mining

KDD – Knowledge Discovery in Databases

ВСТУП

Методи інтелектуального аналізу даних (ІАД) активно застосовуються в інтелектуальних системах , зокрема, в інтелектуальних системах підтримки прийняття рішень, а також в системах керування базами даних (СКБД) і знань, бізнес-додатках, системах машинного навчання, системах електронного документообігу та ін.

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

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

Інтелектуальний аналіз даних – область знань, яка відноситься до обробки даних, що вивчає пошук і опис прихованих, нетривіальних і практично корисних закономірностей у досліджуваних даних. До задач інтелектуального аналізу даних відноситься множина напрямків, такі як пошук документів в локальних і глобальних мережах, сортування, класифікація і кластеризація документів, автоматичне анотування та реферування, системи автоматичного контролю, діалогові системи, системи, які навчаються, модифікація і поповнення баз знань, експертні системи і машинний переклад.

Метою кваліфікаційної роботи є дослідження засобів та методів інтелектуального аналізу даних для систем керування базами даних.

Об’єктом дослідження є методи інтелектуального аналізу даних на основі прецендентів.

Завдання:

- аналіз існуючих технологій, методів та засобів інтелектуального

аналізу даних для сучасних систем керування базами даних;

- аналіз методів інтелектуального аналізу даних на основі прецендентів для інтелектуальних систем та систем керування базами даних;
- розробка/модифікація методу на основі прецедентів;
- розробка програмних засобів.

Особливістю завдань ІАД є те, що вихідні дані недостатньо формалізовані, але можна витягати з них нові знання, використовуючи спеціальні програми. Аналізу та реалізації подібних засобів і присвячена ця робота.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Терміни, що відносяться до ІАД

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

Під сирими даними розуміється формат даних, який не має чіткої специфікації і містить необроблені (або оброблені в мінімальному ступені) дані, що дозволяє уникнути втрат інформації [6]. У зарубіжній літературі термін ІАД трактується як Knowledge Discovery (KD) [7] і Data Mining (DM) [8]. Під KD (виявленням знанням) в базах даних (БД) розуміють будь-який нетривіальний процес ідентифікації достовірних, нових, потенційно корисних і зразків (структур), що добре інтерпретуються в даних. Таким чином, під процесом KD розуміють багатокрокову систему процедур, що включає підготовку даних, пошук зразків в БД, оцінку витягнутого знання, коригування та ітерацію процедур [9]. Під DM розуміють етап процесу KD, що складається в застосуванні специфічних алгоритмів породження зразків, витягнутих з БД. Безліч зразків може бути відкритими, 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.

Під даними розуміється необроблений матеріал, що надається постачальниками даних і використовується споживачами для формування інформації на основі даних.

Об'єкт описується як набір атрибутів. Об'єкт також відомий як запис, випадок, приклад, рядок таблиці та т.і.

Атрибут – властивість, що характеризує об'єкт. Наприклад, колір очей людини, температура води і т.і. Атрибут також називають змінної, полем таблиці, вимірюванням, характеристикою.

Генеральна сукупність (population) – вся сукупність досліджуваних об'єктів, яка цікавить дослідника.

Вибірка (sample) – частина генеральної сукупності, певним способом відібрана з метою дослідження і отримання висновків про властивості та 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 групи фактів.

В ІАД до знань, що виявляється, ставляться такі вимоги [10].

- Знання повинні бути новими, раніше невідомими. Зусилля, що витрачаються на відкриття знань, які вже відомі користувачеві, не окупаються. Тому цінність представляють саме нові, раніше невідомі знання.

- Знання повинні бути нетривіальні. Результати аналізу повинні відображати неочевидні, несподівані закономірності, що становлять так звані приховані 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, в тому числі і на нових даних, з досить високим ступенем достовірності. Корисність полягає в тому, щоб ці знання могли принести певну вигоду при їх застосуванні.

- Знання повинні бути доступні для інтерпретації людиною. Знайдені закономірності повинні бути логічно пояснити, в іншому випадку існує ймовірність, що вони є випадковими. Крім того, виявлені знання повинні бути представлені в зрозумілому для людини вигляді.

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

У ІАД для вилучення нових знань з баз фактів застосовуються різні методи: статистичні та індуктивні процедури (дерева рішень), генетичні алгоритми, штучні нейронні мережі (ШНМ), кластерний аналіз, прецедентні методи (CBR методи) та ін. Процес ІАД включає чотири основні етапи [2].

- На першому етапі аналітик формулює постановку задачі 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 та інтерпретація отриманих результатів (витягнутих знань). При верифікації застосовується тестовий набір записів, виділених з вихідних даних і не піддавалися аналізу.

1.2 Виявлення знань

Процес виявлення знань в БД (процес KDD) - це процес пошуку корисних знань в сирих даних. KDD включає в себе питання: підготовки даних, вибору інформативних ознак, очищення даних, застосування методів DM, постобки даних та інтерпретації отриманих результатів.

1.2.1 Процес виявлення знань. Етапи

Основними етапами процесу виявлення знань є наступні установки і процедури [2].

- Вибір предметної області та релевантного знання для реалізації цілей кінцевого користувача комп'ютерної системи.
- Вибір вихідної безлічі даних і підмножини змінних, які необхідні для

отримання нового знання з бази фактів.

- Уточнення даних і предпроцессінг: вибір основних операцій над даними таким чином, що вони можуть сприяти зменшенню «шумів», визначення стратегій для їх мінімізації.

- Редукція даних: виявлення корисних особливостей даних, щоб подання даних було адекватним рішенням завдань, що відповідають поставленим цілям.

- Вибір завдання DM, тобто специфікація процесу KDD як класифікації, кластеризації та ін.

- Вибір алгоритмів, що реалізують DM для пошуку зразків (patterns) в даних. Цей вибір повинен бути узгоджений з моделями і параметрами представлення даних.

- DM: пошук зразків у формі, зручній для користувача (правила класифікації і кластеризації, регресія, дерева рішень та т.і.).

- Інтерпретація породжених зразків з можливим повторенням етапів 1-7 для подальшої ітерації. Цей етап часто має на увазі використання методів, які перебувають на стику технології DM і технології експертних систем (ЕС).

Від того, наскільки ефективним він буде, в значній мірі залежить успіх вирішення 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 чи інші методи DM, отримують моделі, а на меншій - перевіряють їх. За різницею в точності між тестової та навчальної групами можна судити про адекватність побудованої моделі. Остаточна оцінка цінності видобутого нового знання виходить за рамки аналізу, автоматизованого або традиційного, і може бути проведена тільки після втілення в життя рішення, прийнятого на основі здобутого знання.

Дослідження досягнутих практичних результатів завершує оцінку цінності видобутого засобами DM нового знання.

1.2.2 Підготовка вхідних даних

Для застосування того чи 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 сенсі, що вибір системи даних параметрів виробляється людиною, хоча, звичайно, їх значення можуть обчислюватися автоматично. Після вибору параметрів досліджувані дані можуть бути представлені у вигляді прямокутної таблиці, в якій кожен рядок являє собою окремий випадок, об'єкт або стан досліджуваного об'єкта, а кожна колонка - 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 дані, що входять до неї, необхідно попередньо обробити. По-перше, таблиця може містити параметри, що мають однакові значення для всієї колонки (тобто такі ознаки ніяк не індивідуалізують досліджувані об'єкти), отже, їх треба виключити з аналізу. По-друге, таблиця може містити деякі категоріальні ознаки, значення яоих у всіх записах таблиці різне (тобто не можна використовувати це поле для аналізу даних), і його треба виключити. Нарешті, просто цих полів може бути дуже багато, і якщо все їх включити в дослідження, то це істотно збільшить час обчислень, оскільки практично для всіх методів DM характерна сильна залежність часу роботи від кількості параметрів (квадратична, а нерідко і експоненціальна). У той же час залежність часу від кількості досліджуваних об'єктів лінійна або близька до лінійної. Тому в якості попередньої обробки даних необхідно, по-перше, виділити ту безліч ознак, які найбільш важливі в контексті даного

дослідження, відкинути явно непридатні і виділити ті, які з найбільшою ймовірністю увійдуть в шукану залежність. Для цього, як правило, використовуються статистичні методи, засновані на застосуванні кореляційного аналізу, лінійних регресій та ін. Такі методи дозволяють швидко, хоча і приблизно оцінити вплив одного параметра на інший.

Крім «очищення» даних по стовпцях таблиці (ознаками), іноді буває необхідно провести попереднє «очищення» даних по рядках таблиці (записів). Будь-яка реальна БД зазвичай містить помилки, дуже приблизно певні значення, записи, відповідні якимось рідкісним, винятковим ситуаціям, і інші дефекти, які можуть різко знизити ефективність методів DM, що застосовуються на наступних етапах аналізу. Такі записи необхідно відкинути. Навіть якщо подібні «викиди» не є помилками, а являють собою рідкісні виняткові ситуації, вони все одно навряд чи можуть бути використані, оскільки по декількох точках статистично значимо судити про шукану залежність неможливо.

1.3 Завдання ІАД

ІАД (рисунок 1.1) допомагає вирішувати багато завдань, з якими стикається аналітик. Серед них основними на даний момент є завдання класифікації, регресії, кластеризації та пошуку асоціативних правил [10]. За призначенням ці 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 все одно може бути корисно для аналітика. До такого виду завдань відносяться кластеризація і пошук асоціативних правил.

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

За способами вирішення завдання поділяють на категорії: навчання з вчителем (supervised learning) і навчання без вчителя (unsupervised learning). Дана назва походить від терміна «машинне навчання» (machine learning), часто використовуваного в англомовній літературі [10].

У разі навчання з учителем завдання аналізу даних вирішується в кілька етапів. Спочатку за допомогою деякого алгоритму DM будується модель аналізованих даних – класифікатор. Потім побудована модель піддається навчанню. Іншими словами, перевіряється якість роботи класифікатора, і якщо вона незадовільна, то проводиться додаткове навчання моделі. Так триває до тих пір, 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 і регресії.

Навчання без вчителя об'єднує завдання, що виявляють описові моделі, наприклад, закономірності в покупках, скоєних клієнтами великого магазину. Очевидно, якщо ці закономірності існують, то модель повинна їх представити.

Рисунок 1.1 – Інтелектуальний аналіз даних

Перевагою таких завдань є можливість їх вирішення без будь-яких попередніх знань про аналізованих даних. До цих завдань належить кластеризація і пошук асоціативних правил.

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

- класифікацію;
- кластеризацию (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:

призначаються класи тут можуть бути «кредитоспроможний» і «некредітоспособен». Необхідно відзначити, що для вирішення завдання необхідно, щоб безліч класів було відомо заздалегідь і було б кінцевим і рахунковим.

Завдання кластеризації полягає в розподілі безлічі об'єктів на групи (кластери) схожих за параметрами. При цьому, на відміну від класифікації, число кластерів і їх характеристики можуть бути заздалегідь невідомі і визначатися в ході побудови кластерів виходячи зі ступеня близькості об'єктів за сукупністю параметрів. Інша назва цього завдання - сегментація. Наприклад, Інтернет-магазин може бути зацікавлений в 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 з завданням класифікації, але в ході його рішення проводиться пошук шаблонів для визначення числового значення. Іншими словами тут параметр, що передбачається, 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 враховуватися наявні тенденції (тренди), сезонність, інші чинники. Класичним прикладом є прогнозування цін акцій на біржі.

Завдання пошуку асоціативних правил. Задача визначення взаємозв'язків полягає у визначенні наборів об'єктів,що частіше зустрічаються, серед безлічі подібних наборів. Класичним прикладом є аналіз споживчого кошика, який дозволяє визначити набори товарів, які 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 пов'язаних товарів. Дане завдання також відноситься до класу навчання без учителя.

Аналіз послідовностей або секвенційний аналіз одними авторами розглядається як варіант попередньої задачі, іншими - виділяється окремо [11]. Метою, в даному випадку, є виявлення закономірностей в послідовності подій. Подібна інформація дозволяє, наприклад, попередити збій в роботі інформаційної системи, отримавши сигнал про настання події, 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 істотно відрізняються від норми. Відхилення може сигналізувати про якусь незвичному подію (несподіваний результат експерименту, шахрайська операція по банківській карті і ін.). Або, наприклад, про помилку введення даних оператором.

Завдання оцінювання зводиться до передбачення безперервних значень ознак. Аналіз зв'язків – завдання знаходження залежностей в наборі даних. Для вирішення завдання візуалізації використовуються графічні методи, що показують наявність закономірностей в даних [14]. В результаті візуалізації створюється графічний образ аналізованих даних (наприклад, методи візуалізації на основі представлення даних в 2-D і 3-D вимірах).

1.4 Існуючі методи ІАД

Всі 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 подальшого використання. У разі безпосереднього використання (збереження) даних, вихідні дані зберігаються в явному

детальному вигляді і безпосередньо використовуються на стадіях прогнозування (побудови прогнозних моделей) (а також на стадії аналізу винятків). Проблема цієї групи методів полягає в тому, що при їх 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 витягується з вихідних даних і перетворюється в якісь формальні конструкції, вид яких залежить від конкретного використовуваного методу DM.

На етапах прогнозування і аналізу винятків використовуються отримані формальні конструкції, які значно компактніше самих БД. До методів даної групи відносять різні логічні методи (нечіткі запити і 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 найбільш часто застосовуються для вирішення 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 за завданнями DM. Відповідно до такої класифікації можна виділити дві групи. Перша з них – це поділ методів DM на вирішальні завдання сегментації (тобто завдання класифікації і кластеризації) і завдання прогнозування. Якщо класифікувати методи відповідно до використовуваних в них моделям, то можна виділити методи, спрямовані на отримання описових результатів, і методи, спрямовані на отримання результатів прогнозування. Описові методи служать для знаходження шаблонів або зразків, що описують дані, які піддаються інтерпретації з точки зору аналітика.

До них відносяться ітеративні методи кластерного аналізу, в тому числі: алгоритм k-середніх, k-медіани, ієрархічні методи кластерного аналізу, карти Кохонена, методи крос-табличної візуалізації і ін. [15].

Прогнозуючі методи використовують значення одних змінних для прогнозування невідомих (пропущених) або майбутніх значень інших (цільових) змінних. До цієї групи методів відносять нейронні мережі, дерева рішень, лінійну регресію, метод найближчого сусіда, метод опорних векторів і ін. [15]. Нижче наведено ряд основних методів, які використовуються для ІАД (рисунок 1.2).

Рисунок 1.2 – Методи ІАД

Асоціація. Асоціація (або відношення), найбільш відомий і простий метод ІАД. Для виявлення моделей робиться просте зіставлення двох або більше елементів, часто одного і того ж типу. Наприклад, відстеження переваг і звичок покупців при купівлі товарів в магазині.

Класифікація. Класифікацію можна використовувати для отримання

уявлення про тип покупців, товарів або об'єктів, описуючи кілька атрибутів для ідентифікації певного класу. Наприклад, автомобілі легко класифікувати за типом (седан, позашляховик, кабріолет), визначивши різні атрибути (кількість місць, форма кузова, провідні колеса). Вивчаючи новий автомобіль, можна віднести його до певного класу, порівнюючи атрибути з відомим визначенням. Ті ж принципи можна застосувати і до покупців, наприклад, класифікуючи їх за віком і соціальної групи. 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. Кластерний аналіз – це спосіб угруповання багатовимірних об'єктів, заснований на представленні результатів окремих спостережень точками відповідного геометричного простору з подальшим виділенням груп як «згустків» цих точок (кластерів, таксонів). Даний метод дослідження отримав розвиток в останні роки в зв'язку з можливістю комп'ютерної обробки великих БД. Кластерний аналіз передбачає виділення компактних, віддалених одна від одної груп об'єктів, відшукує «природне» розбиття сукупності на області скупчення об'єктів. Він використовується, коли вихідні дані представлені в вигляді матриць близькості або відстаней між об'єктами або у вигляді точок у багатовимірному просторі. Найбільш поширені дані другого виду, для яких кластерний аналіз орієнтований на виділення деяких геометрично віддалених груп, всередині яких об'єкти близькі. Досліджуючи один або більше атрибутів, або класів, можна згрупувати окремі елементи даних разом, отримуючи структуроване 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

подібності та діапазони узгоджуються між собою.

Метод кластеризації працює в обидві сторони. Можна припустити, що в певній точці є кластер, а потім використовувати свої критерії ідентифікації, щоб перевірити це. Графік, зображений на рисунку 1.3, демонструє наочний приклад. Тут вік покупця порівнюється з вартістю покупки. Розумно очікувати, що люди у віці від двадцяти до тридцяти років (до вступу в шлюб і появи у них дітей), а також в 50-60 років (коли діти покинули рідну домівку) мають більш високий дохід на одного члена сім'ї.

У цьому прикладі видно два кластери, один в районі $ 2000 / 20-30 років і інший в районі $ 7000-8000 / 50-65 років. В даному випадку висунули гіпотезу і 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 найближчому сусідові.

Така побудова кластерів являє собою спрощений приклад так званого образу найближчого сусіда. Окремих покупців можна розрізняти по їх буквальній близькості один до одного на графіку. Досить імовірно, що покупці з одного і того ж кластера поділяють і інші загальні атрибути, і це припущення можна використовувати для пошуку, класифікації та інших видів аналізу членів набору даних.

Метод кластеризації можна застосувати і в зворотний бік: з огляду на певні вхідні атрибути, виявляти різні артефакти. Наприклад, недавнє дослідження чотиризначних PIN-кодів виявили кластери чисел в діапазонах 1-12 і 1-31 для першої і другої пар.

Зобразивши ці пари на графіку, можна побачити кластери, пов'язані з датами (дні народження, ювілеї).

Прогнозування. Прогнозування – це широка тема, яка простягається від передбачення відмов компонентів обладнання до виявлення шахрайства і навіть прогнозування прибутку компанії. У поєднанні з іншими методами

ІАД прогнозування передбачає аналіз тенденцій, класифікацію, зіставлення з моделлю і відносини. Аналізуючи минулі події або екземпляри можна передбачати майбутнє. Наприклад, використовуючи дані по авторизації кредитних карт, можна об'єднати аналіз дерева рішень минулих транзакцій людини з класифікацією і зіставленням з історичними моделями з метою виявлення шахрайських транзакцій [17].

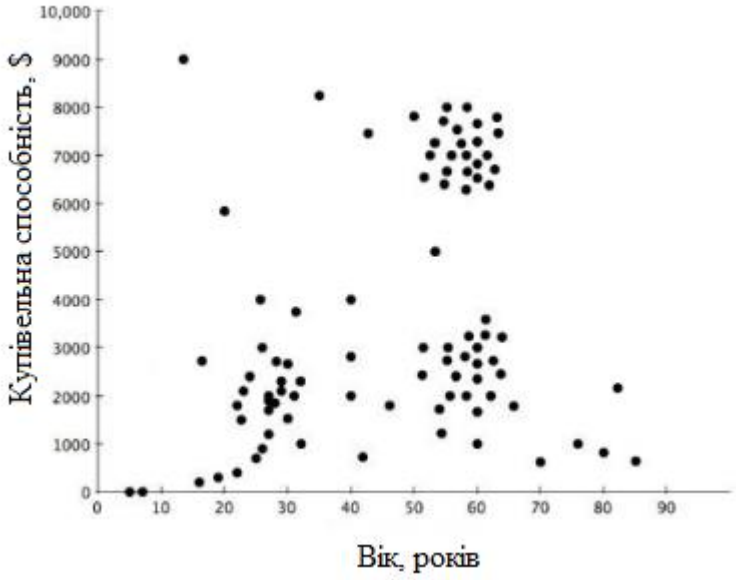

Рисунок 1.3 – Приклад кластеризації покупців

Послідовні моделі. 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 різні пори року вони купують певні набори продуктів [17].

За цією інформацією додаток прогнозування купівельної корзини, грунтуючись на частоті і історії покупок, може автоматично припустити, що в корзину будуть додані ті чи інші продукти.

Дерево рішень, пов'язане з більшістю інших методів (головним чином, з класифікацією і прогнозуванням), можна використовувати або в рамках

критеріїв відбору, або для підтримки вибору певних даних в рамках загальної структури [17]. Дерево рішень починає функціонувати з простого питання, яке має дві відповіді (іноді більше). Кожна відповідь призводить до наступного питання, допомагаючи класифікувати та ідентифікувати дані або робити прогнози.

Дерева рішень часто використовуються з системами класифікації інформації про властивості і з системами прогнозування, де різні прогнози можуть грунтуватися на минулому історичному досвіді, який допомагає побудувати структуру дерева рішень і отримати результат.

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

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

Байєсовські мережі. Байєсова мережу - графічна модель, що представляє собою безліч змінних і їх імовірнісних залежностей по Байеса [20]. Наприклад, баєсова мережа може бути використана для обчислення ймовірності того, що хворий пацієнт за наявністю або відсутністю ряду симптомів, грунтуючись на даних про залежність між симптомами і хворобами.

Лінійна регресія. Лінійна регресія – використовувана в статистиці регресійна модель залежності однієї (що пояснюється, залежною) змінної у

від іншої або кількох інших змінних (факторів, регресорів, незалежних змінних) x з лінійною функцією залежності [21]. Модель лінійної регресії є часто респонденти користуються послугами і найбільш вивченої в економетрики. А саме вивчені властивості оцінок параметрів, одержуваних різними методами при припущеннях про імовірнісних характеристиках 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 за параметрами, ніж лінійність за факторами моделі.

1.3.1 Прецедентний підхід

Основною метою використання прецедентного підходу є отримання рішення для поточної ситуації на основі прецедентів, які вже мали місце в минулому. Прецедентний підхід може бути використано в разі, коли достовірні алгоритми неспроможні (неповнота знань про предметну область або наявність обмежень по тимчасовим і обчислювальних ресурсів) або взагалі відсутні. Застосування методу для вирішення завдань виправдано при виконанні наступних умов: 1. Подібні завдання повинні мати подібні рішення (принцип регулярності). 2. Види завдань, з якими стикається вирішувач, повинні мати тенденції до повторення. Прецедент – це ситуація, рішення для якої вже відомо. Прецедент може бути отриманий в результаті роботи системи або в якості прикладу рішення від експерта в проблемній області. Прецедент складається з опису проблемної ситуації, застосованого рішення і 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 випадків у майбутньому.

Часом для вирішення завдань класифікації та кластеризації доцільно

використовувати апарат ШНМ. У даного підходу є ряд особливостей:

- ШНМ мають здатність навчатися на прикладах;
- ШНМ легко працюють в розподілених системах з великою паралельністю в силу своєї природи;
- оскільки ШНМ підлаштовують свої вагові коефіцієнти, грунтуючись на вихідних даних, це допомагає зробити вибір значущих характеристик менш суб'єктивним.

На практиці дуже рідко використовується тільки один з цих методів. Наприклад, класифікація та кластеризація - подібні методи. Використовуючи кластеризацию 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 Програмні засоби для 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. На сьогоднішній день всі представлені на ринку сучасні СКБД включають в себе різні набори компонентів для ІАД, які входять до складу інструментів платформи Business intelligence.

Під Business intelligence (BI) розуміють програмне забезпечення, створене для допомоги менеджеру в аналізі інформації про свою компанію і її оточенні. Більшість інструментів BI застосовуються кінцевими користувачами для доступу, аналізу і генерації звітів за даними, які найчастіше розташовуються в сховищі, вітринах даних або оперативних складах даних. Розробники додатків використовують BI-платформи для створення і впровадження BI-додатків, які не розглядаються як BI-

інструменти. Прикладом BI-додатків 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, яка полягає в підготовці сумарної інформації на основі великих масивів даних, структурованих по багатовимірному принципу. Реалізації технології OLAP є компонентами програмних рішень класу BI.

Технологія OLAP орієнтована, головним чином, на обробку нерегламентованих запитів до сховищ даних. Створення сховищ даних викликано тим, що аналізувати дані і, зокрема, дані OLTP- систем (On-Line Transaction Processing - оперативна обробка транзакцій в реальному часі) безпосередньо неможливо або важко, так як вони є розрізненими, зберігаються в форматах різних СКБД і в різних сегментах корпоративної мережі. В 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 простою і зрозумілою структури. На рисунку 1.4 показані компоненти, що входять до типове сховище даних. Суцільні стрілки позначають потоки даних, пунктирні - метаданих.

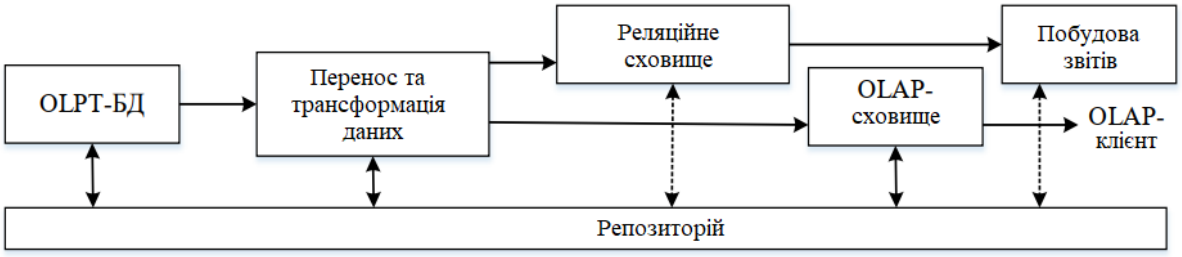

Рисунок 1.4 – Структура сховища даних

Основна мета аналізу даних – якісна і кількісна оцінка досягнутих результатів і (або) динаміки діяльності компанії. Принципи OLAP,

використовувані для цього, були сформульовані Е. Коддом. Центральне місце серед них займає підтримка багатовимірного представлення даних. У багатовимірній моделі даних БД представляється у вигляді одного або декількох кубів даних (гіперкубів). Осями гіперкуба є основні атрибути аналізованого бізнес-процесу. Незалежні вимірювання гиперкуба представляють багатовимірний простір даних. Кожному вимірюванню відповідає атрибут, який характеризує одну з якісних властивостей даних: час, територію, категорію продукції і т.п. На перетині осей-вимірювань (dimensions), тобто в осередку гіперкуба, містяться дані, що кількісно характеризують аналізований процес. Ці дані називаються заходами (measures) або показниками.

У процесі аналізу виконуються операції побудови перетинів (проекцій) гиперкуба шляхом фіксації значень наборів атрибутів-координат, а також операції стиснення гиперкуба шляхом використання значень атрибутів-вимірювань вищих рівнів ієрархії і відповідного агрегування значень, що асоціюються з ними показників.

Ієрархічні відносини можуть бути природним чином введені для ряду атрибутів. Можуть застосовуватися і зворотні операції деталізації даних. Треба зауважити, що куб даних розглядається як концептуальне, а не фізичне уявлення. Для забезпечення зручності сприйняття даних аналітиками використовуються операції обертання куба шляхом зміни порядку вимірювань. Для візуалізації даних з гіперкуба, як правило, застосовуються двовимірні уявлення у вигляді таблиць, що мають складні ієрархічні заголовки рядків і стовпців. Двовимірне подання куба можна отримати, фіксуючи значення всіх вимірювань, крім двох.

Багатовимірність в OLAP-додатках втілюється в рамках двох або трирівневої архітектури. Перший рівень підтримує багатовимірне представлення даних, абстраговані від їх фізичної структури. Він містить виправлення для багатовимірної візуалізації і маніпулювання даними для кінцевого користувача. Другий рівень забезпечує багатовимірну обробку. Він

включає мову формулювання багатовимірних запитів (SQL для цих цілей непридатний) і програмний процесор, здатний виконувати такі запити. На третьому рівні архітектури реалізується фізична організація зберігання багатовимірних даних. В рамках нього для підтримки багатовимірних моделей даних використовуються або спеціальні OLAP-СУБД, або звичайні реляційні структури. Найбільше застосування OLAP знаходить в продуктах для фінансового планування, в сховищах даних, в рішеннях класу BI (рисунок 1.5).

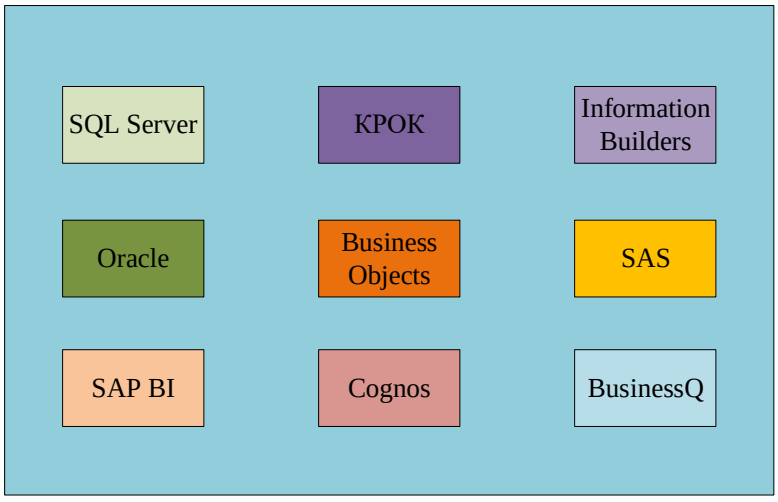

Рисунок 1.5 – Програмні засоби ІАД для СКБД

Список найбільш відомих виробників комерційних OLAP- продуктів, згідно з OLAP Report, включає: Microsoft (Microsoft SQL Server Analysis Services), Hyperion (Hyperion Essbase), Cognos (Cognos PowerPlay), Business Objects, MicroStrategy, SAP (SAP BW), Cartesis (Cartesis Magnitude), Systems Union / MIS AG, Oracle (Oracle Express, OLAP Option), Applix (IBM Cognos TM1).

Існує кілька open-source рішень, включаючи Mondrian і Palo. Шар фізичного зберігання даних реалізується або в реляційних, або в багатовимірних структурах, які подаються у вигляді багатовимірних масивів. Зазвичай OLAP-продукти забезпечують обидва ці способи зберігання.

1.6 Програмні засоби ІАД для СКБД

BI-засоби Oracle засновані на Oracle OLAP - засобах аналітичної обробки даних, вбудованих безпосередньо в реляційну СКБД Oracle. Крім власне OLAP-сховища, Oracle надає засоби DM, інструменти створення звітів, доставки результатів аналізу за допомогою порталу, а також ряд засобів, що дозволяють створювати аналітичні Java-додатка, зокрема, Java OLAP API і компоненти OLAP Beans, призначених для використання в засобах розробки Java-додатків. Відзначимо, що, крім власне OLAP-сервера, засобів доступу до OLAP- даними і засобів створення BI-додатків, Oracle поставляє ряд готових аналітичних рішень на їх основі. Компанія Oracle виробляє великий спектр програмних продуктів. Це і готові програми (Oracle E-Business Suite), і сервера додатків і засоби для колективної роботи і різні СУБД. 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 і 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, які приймають рішення (dashboard). За допомогою цих продуктів можна швидко створювати регламентовані і нерегламентовані запити, складні звіти, інтерфейси для аналітиків.

Служби Analysis Services надають такі функції і засоби для створення рішень по ІАД:

- набір стандартних алгоритмів ІАД;
- конструктор ИАД, призначений для створення і перегляду моделей ИАД, управління ними та побудови прогнозів;
- мову розширень ІАД (DM eXtensions to SQL, DMX). Для роботи з

наданими коштами ИАД використовується середу BI Development Studio, скорочено BI DevStudio.

Нижче перераховані алгоритми ІАД, реалізовані в Microsoft SQL Server (рисунок 1.6):

- спрощений алгоритм Байеса - Microsoft Naive Bayes;
- алгоритм дерева прийняття рішень - Microsoft Decision Trees;
- алгоритм часових 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.

Рисунок 1.6 – Служби та компоненти СКБД Microsoft SQL сервер

Були розглянуті існуючі програмні засоби для ІАД. На слайді представлені деякі з них. Недоліками яких є: складність продуктів, недостатня функціональність (Microsoft SQL Server); збої в роботі, недостатній рівень підтримки (Oracle); нестійкість та збої в роботі (SAP) та інші.

2 ІАД З ВИКОРИСТАННЯМ ПРЕЦЕДЕНТІВ

2.1 Навчання на основі прецедентів

Як правило, CBR методи включають чотири основні етапи, що утворюють так званий CBR цикл або цикл навчання по прецедентах (прикладів), структура якого представлена на рисунку 2.1. Основними етапами CBR циклу є:

- витяг найбільш відповідного (подібного) прецеденту (або прецедентів) для ситуації, що склалася з БП;

- повторне використання вилученого прецеденту для спроби вирішення поточної проблеми;

- адаптація і застосування отриманого рішення для вирішення поточної проблеми;

- збереження нового прийняття рішення як частини нового прецеденту.

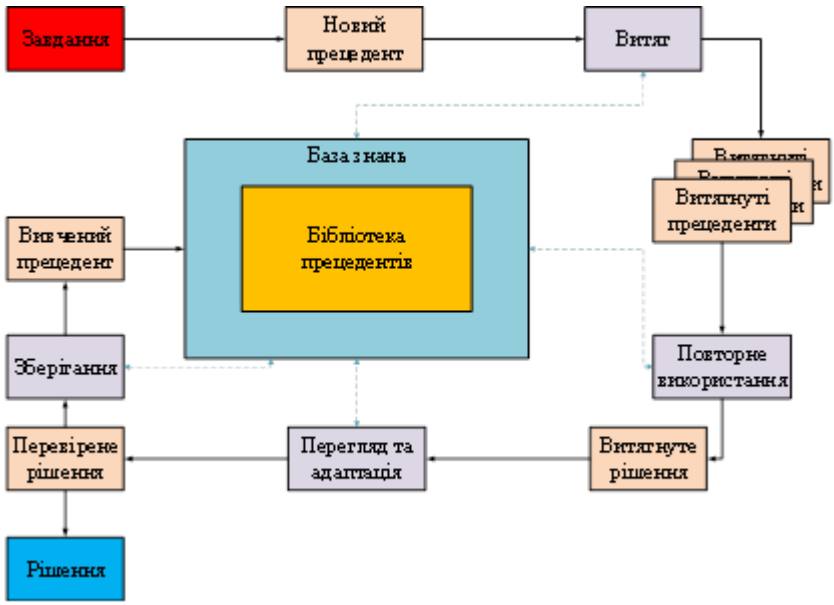

Рисунок 2.1 – Структрура CBR

Інформація про нову проблемної ситуації використовується для вилучення з БП найбільш підходящого прецеденту (прецедентів). Витягнутий прецедент використовується повторно для отримання рішення нової проблеми (завдання). Потім запропоноване рішення в разі необхідності може бути адаптоване до особливостей нової ситуації і застосовано на практиці. У разі успішного застосування, перевірене рішення спільно з описом проблемної ситуації утворює новий прецедент, який зберігається в БП. Таким чином, системою накопичується досвід (прецеденти) і реалізується машинне навчання. У CBR циклі може використовуватися не тільки БП, але й узагальнені знання про предметну область для підтримки процесу міркування на основі прецедентів. Ця підтримка може бути слабкою або сильною, а може й не бути взагалі.

2.2 Модифікований CBR-цикл

У стандартному CBR циклі інформація про нову проблемної ситуації використовується для вилучення з БП найбільш підходящого прецеденту (прецедентів). Витягнутий прецедент використовується повторно для отримання рішення нової проблеми (завдання). Потім запропоноване рішення може бути адаптоване до особливостей нової ситуації і в разі успішного застосування, новий прецедент зберігається в БП.

Для роботи модифікованого CBR циклу особливо в задачах класифікації можуть застосовуватися тестові вибірки з прикладами для перевірки коректності знайденого рішення (рисунок 2.2). При наявності експертних знань (тестових вибірок) перед збереженням повинна виконуватися перевірка правильності рішення на тестових наборах. Якщо рішення проходить перевірку і приймається користувачем, тоді воно зберігає в БП як новий прецедент. Якщо перевірка коректності рішення на тестових наборах завершується невдало, тоді прецедент зберігається в базі невдалих прецедентів.

Таким чином, пропонується використовувати тестову (експертну) вибірку на останньому етапі CBR циклу для формування бази вдалих (БП) і бази невдалих прецедентів (БНП).

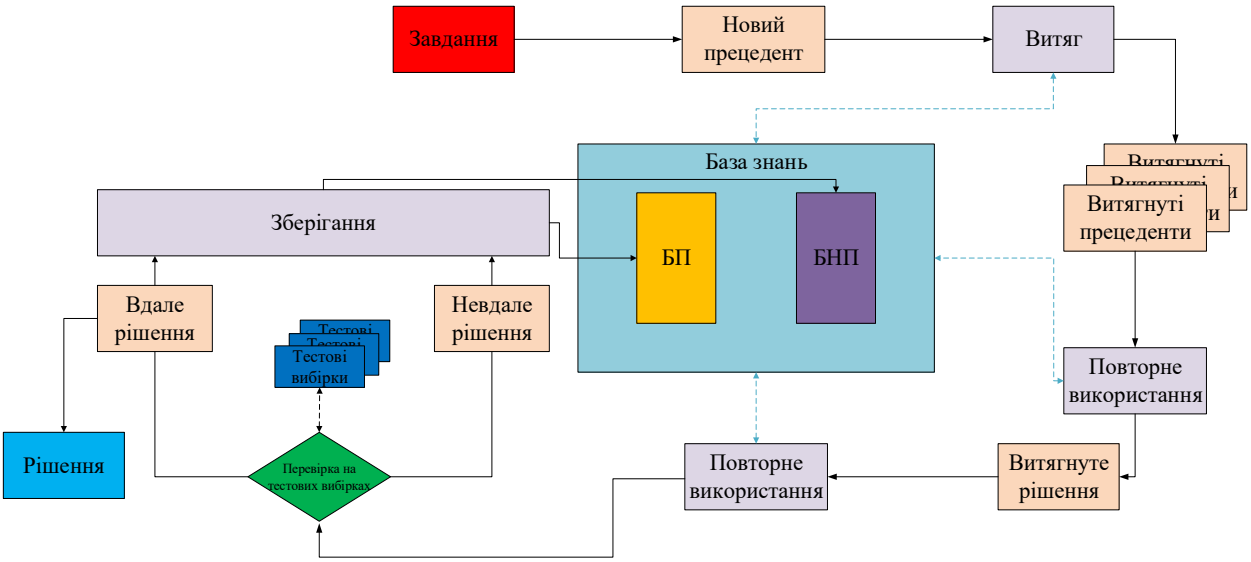

Рисунок 2.2 – Модифікований CBR цикл

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.

Існують різні способи подання та зберігання прецедентів: від простих (лінійних) до складних ієрархічних. Слід зазначити, що прості способи зберігання та подання прецедентів, що базуються на технології реляційних БД, вимагають значно менше витрат на реалізацію, а також підтримку і супровід БП системи на відміну від більш складних, але час для здійснення пошуку рішення при простому поданні прецедентів може знадобитися істотно більше в порівнянні з іншими способами представлення і збереження прецедентів.

Відповідно до прецедентів може включати наступні компоненти:

- опис завдання (проблеми або проблемної ситуації);

- рішення задачі (діагноз по проблемної ситуації і рекомендації ОПР);
- результат застосування рішення.

Опис результату може включати список виконаних дій, додаткові коментарі та посилання на інші прецеденти. Прецедент може мати як позитивний, так і негативний результат застосування рішення, а також в деяких випадках може приводитися обгрунтування вибору даного рішення і можливі альтернативи. Прецеденти можуть бути представлені у вигляді записів в БД, концептуальних графів, семантичної мережі, деревовидних структур, предикатів, фреймів, малюнків і мультимедійної інформації. Основні способи подання прецедентів можна розділити на наступні групи:

- параметричні;
- об'єктно-орієнтовані;
- спеціальні (у вигляді дерев, графів, логічних формул і т.і.).

Подання прецедентів у вигляді експертних правил продукційного типу.

Такий спосіб є найбільш зрозумілим і популярним методом уявлення прецедентів. Правила забезпечують формальний спосіб представлення рекомендацій, знань або стратегій. Вони частіше підходять в тих випадках, коли предметна область виникає з емпіричних асоціацій, накопичених за роки роботи з вирішення завдань в 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 кажуть, що правило спрацьовує. Інтерпретатор правил зіставляє частини правил ЯКЩО з фактами і виконує, то правило, частина ЯКЩО якого відповідає фактам, тобто інтерпретатор правил працює в циклі «зіставити - виконати», формуючи послідовність дій.

Подання прецедентів в структурованій формі.

До такого поданням можна віднести дерева, графи, семантичні мережі. Більш детально дане подання розглянемо на прикладі концептуальних графів.

Концептуальний граф (conceptual graph) - це кінцевий, пов'язаний, двочастковий, орієнтований мультіграф. Вузли графа представляють поняття, або концептуальні відносини. У концептуальних графах мітки дуг не використовуються. Відносини між поняттями представляються вузлами концептуальних відносин. На рисунку 2.3 8 вузли b \ h, a2a (мітки правил) і 19, 56, 47, 9 (номери діагнозів) представляють поняття, а After, DiagnosisThisLable - концептуальні відносини.

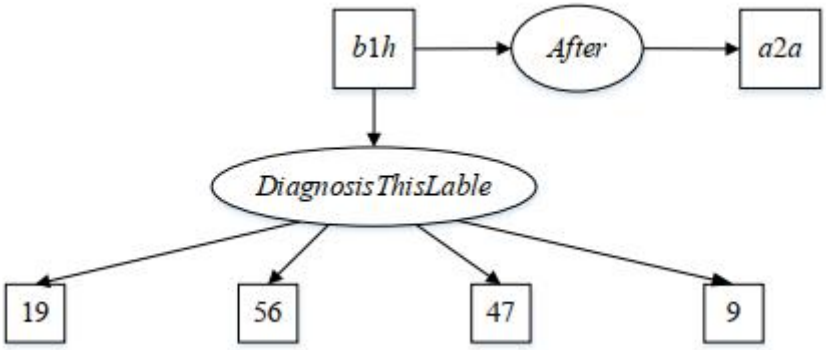

Рисунок 2.3 – Приклад концептуального графу

У концептуальних графах вузли понять представляють 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 помічених дуг є простота уявлення відносин будь-арності. Парне ставлення представляється вузлом концептуальних відносин, що має N дуг. Кожен концептуальний граф являє один вислів. Типова БЗ буде складатися з ряду таких графів. Графи можуть бути довільної складності, але вони повинні бути кінцевими.

2.2 Вилучення прецедентів

Якість роботи CBR системи при вирішенні складних завдань безпосередньо залежить від кількості прецедентів, що містяться в БП. На відміну від пошуку в БД, де визначається конкретне значення в записах, пошук прецедентів повинен здійснюватися в умовах часткового збіги значень, так як може не існувати прецеденту, повністю збігається з поточним. У цьому випадку застосовуються спеціальні методи, які використовують метричні алгоритми і різні евристики. На першому етапі CBR циклу (вилучення прецедентів) виконується визначення ступеня подібності поточної ситуації з прецедентами з БП системи і подальше їх вилучення з метою дозволити нову проблемну ситуацію, що склалася на об'єкті. Вилучення прецедентів безпосередньо пов'язано зі способом подання прецедентів і відповідно зі способом організації БП. БП є важливою складовою БЗ ІС, але може виступати як окремий компонент системи. Таким чином, структура БП надає істотний вплив на різні показники роботи системи і, зокрема, на час пошуку та вилучення прецедентів. Для успішної реалізації CBR систем необхідно забезпечити коректне вилучення прецедентів з БП системи.

Існують різні способи отримання прецедентів з БП системи [2]:

- метод найближчого сусіда і його модифікації;
- метод пошуку на деревах рішень;
- метод вилучення на основі знань;
- метод вилучення з урахуванням застосування прецедентів.

2.2.1 Метод найближчого сусіда

Це найпоширеніший метод порівняння і вилучення прецедентів. Він дозволяє досить легко обчислити ступінь подібності поточної проблемної ситуації і прецедентів з БП системи по кожному параметру, використовуваному для опису прецедентів і поточної ситуації. З метою

визначення ступеня подібності вводиться метрика (наприклад, міра подібності Хеммінга) на просторі всіх параметрів. У цьому просторі визначається точка, відповідна поточної проблемної ситуації, і відповідно до обраної метрикою визначається найближча до неї точка (найближчий сусід - прецедент, який має максимальну ступінь подібності з поточною ситуацією) з точок, що представляють прецеденти з БП. Для обраної метрики (міра подібності Хеммінга) за методом найближчого сусіда ступінь подібності прецеденту і поточної проблемної ситуації обчислюється виходячи з того, що при збігу всіх параметрів в описі прецеденту і поточної ситуації ступінь подібності буде дорівнює 1, а кожен співпав параметр дає внесок рівний 1 / n, де n - число параметрів в описі прецеденту і поточної ситуації. Метод визначення найближчого сусіда (найближчих сусідів) застосовується для вирішення завдань класифікації, кластеризації, регресії і розпізнавання образів.

До переваг даного методу вилучення прецедентів можна віднести наступні аспекти:

- простота реалізації;
- універсальність в сенсі незалежності від специфіки конкретної проблемної області;
- гарантоване отримання кращого з можливих рішень. Як недоліки методу найближчого сусіда можна виділити наступне:
- параметри об'єкта можуть брати тільки числові значення, інакше вони повинні бути перетворені (наприклад, дискретні значення так / ні будуть інтерпретуватися як 1/0);
- складність вибору метрики для визначення ступеня подібності та істотна залежність результату від обраної метрики;
- пряма залежність необхідних обчислювальних ресурсів від розміру використовуваної БП;

- неефективність при роботі з неповними і зашумленими вихідними

даними.

На практиці застосовуються різні модифікації зазначеного методу [2]. Зазвичай рішення вибирається на основі декількох найближчих точок (сусідів), а не однієї (метод k-NN). Такий метод більш стійкий, оскільки дозволяє згладити окремі викиди, випадковий шум, завжди присутній в даних. У методі найближчого сусіда можливе використання знань про предметну область, які можуть вказувати на відносну близькість один до одного значень параметрів, а не просто збіг значень цих параметрів. Цей підхід отримав назву - методу найближчого сусіда, заснованого на знаннях. Слід зазначити, що метод визначення найближчого сусіда може використовуватися спільно з іншими методами отримання прецедентів, коли первісна вибірка прецедентів здійснюється за допомогою іншого методу, а на другому етапі з отриманої вибірки, порівнюючи попарно прецеденти з поточною ситуацією, методом найближчого сусіда витягується найбільш близький прецедент.

2.2.2 Метод вилучення прецедентів на основі дерев рішень

Метод вилучення прецедентів на основі дерев рішень базується на знаходженні необхідних прецедентів шляхом дозволу вершин дерева рішень. Кожна вершина дерева вказує, з якої її гілки слід здійснювати подальший пошук рішення. Вибір гілки виробляється на основі інформації про поточну проблемної ситуації. Необхідно дістатися до кінцевої вершини, яка відповідає одному або декільком прецедентів. Якщо кінцева вершина пов'язана з деяким підмножиною прецедентів, то тоді для вибору, найбільш підходящого з них може використовуватися метод найближчого сусіда. Такий підхід рекомендується застосовувати для великих БП, тому що основна частина роботи по вилученню прецедентів виконується заздалегідь на етапі побудови дерева рішень, що значно скорочує час пошуку рішення.

2.2.3 Метод вилучення прецедентів на основі знань

Метод вилучення прецедентів на основі знань на відміну від попередніх методів дозволяє врахувати знання експертів (ЛПР) по конкретній предметній області (коефіцієнти важливості параметрів, виявлення залежності і т.д.) при вилученні прецедентів. Метод реалізує підхід, заснований на індексації прецедентів спеціальним чином (семантичної індексації). При визначенні прецедентів враховуються коефіцієнти, що відображають важливість параметрів прецедентів, задані експертом або ЛПР, і інша інформація, що дозволяє врахувати знання про конкретну предметну область. За рахунок цього значно скорочується час пошуку рішення, що є істотною перевагою даного методу. Процес виконання індексації ускладнюється з ростом числа прецедентів в БП і необхідністю проводити індексацію динамічно. Для реалізації методу потрібно передбачити в структурі прецедентів і БП можливості представлення і збереження семантичної інформації, а також додаткові витрати на підтримку БП для обліку знань про конкретну предметну область. Метод може успішно застосовуватися спільно з іншими методами отримання прецедентів, особливо коли БП має великі розміри, і предметна область є відкритою і динамічною.

2.2.4 Метод вилучення прецедентів з урахуванням їх застосовності

У більшості систем, що використовують механізми міркувань на основі прецедентів, передбачається, що найбільш схожі з поточної проблемною ситуацією прецеденти є і найбільш застосовними у цій ситуації. Однак це не завжди так. В основі методів вилучення на основі застосування прецедентів лежить той факт, що витяг прецедентів базується не тільки на їх схожості з поточної проблемною ситуацією, але і на те, наскільки хорошу для бажаного результату модель вони собою являють. Таким чином, на вибір видобутих прецедентів впливає можливість їх успішного застосування (адаптації) в

конкретній ситуації, тобто наявність відомостей про їх застосовності в ситуації, що склалася. У деяких системах ця проблема вирішується шляхом збереження прецедентів разом з коментарями щодо їх застосування. Використання даного методу дозволяє зробити пошук рішення більш ефективним, заздалегідь відкидаючи частину свідомо неперспективних прецедентів. Крім розглянутих методів вилучення прецедентів можуть успішно застосовуватися й інші методи (наприклад, апарат ШНМ). Безумовно, добре навчена ШНМ здатна успішно і досить швидко вирішувати завдання класифікації, кластеризації та визначення схожих прецедентів, але проблеми з ШНМ полягають в необхідності використання представницької навчальної вибірки для навчання мережі із заданою точністю і істотних витрат часу на навчання ШНМ. Крім того, виникає проблема, пов'язана з розробкою спеціальної топології ШНМ, орієнтованої на конкретну проблемну область і рішення складних багатопараметричних задач. Вибір методу отримання прецедентів безпосередньо пов'язаний зі способом подання прецедентів і відповідно зі способом організації БП.

2.2.5 Метод k-NN

Метод k-NN широко поширений метод для вирішення задач класифікації, кластеризації, дискримінантного аналізу, регресії і розпізнавання образів. Суть методу полягає у визначенні заданого числа k найближчих сусідів (прецедентів) до нової ситуації, що склалася в просторі ознак (параметрів).

Що стосується виконання завдання класифікації визначається, до якого діагностичного класу належить більшість найближчих сусідів (наприклад, може застосовуватися просте голосування) і до цього класу належить поточна проблемна ситуація. Вибір значення параметра k займає ключове місце в методі k-NN. Дійсно, параметр k є одним з найбільш значущих чинників, що впливають на якість одержуваного результату. Один з підходів

до оцінки необхідного числа найближчих сусідів - сприймати k як параметр згладжування. Збільшення k призводить до зменшення впливу випадкових похибок в даних, але при цьому поділ на класи стає менш чітким. Отже, для параметра k як і для будь-якого згладжує параметра, необхідно визначити оптимальне значення, при якому б досягався бажаний компроміс. Оптимальне число k можна підбирати експериментальним шляхом або використовувати спеціальні методи оцінки невідомих параметрів (наприклад, метод крос-перевірки).

Як і у випадку з методом простого найближчого сусіда в методі k-NN необхідно вирішувати задачу вибору метрики для визначення близькості діагностованих ситуацій. Це завдання в умовах великої розмірності простору ознак надзвичайно складна і передбачає виконання аналізу багатовимірної структури експериментальних даних для мінімізації числа об'єктів (прецедентів), що представляють діагностичні класи.

Метод k-NN широко поширений метод для вирішення задач класифікації, кластеризації, дискримінантного аналізу, регресії і розпізнавання образів. Суть методу полягає у визначенні заданого числа k найближчих сусідів (прецедентів) до нової ситуації, що склалася в просторі ознак (параметрів).

На слайді наведено алгоритм вилучення прецедентів.

Вхідні дані:

- T - поточна ситуація;
- CB - непорожня безліч прецедентів (БП);
- m - кількість розглянутих прецедентів з БП;
- S (C, T) - задана метрика (міра схожості);
- H - порогове значення ступеня подібності.

Вихідні дані:

- безліч витягнутих прецедентів SC.

Проміжні дані: j.

На рисунку 2.3 приведено алгоритм вилцуення прецедентів.

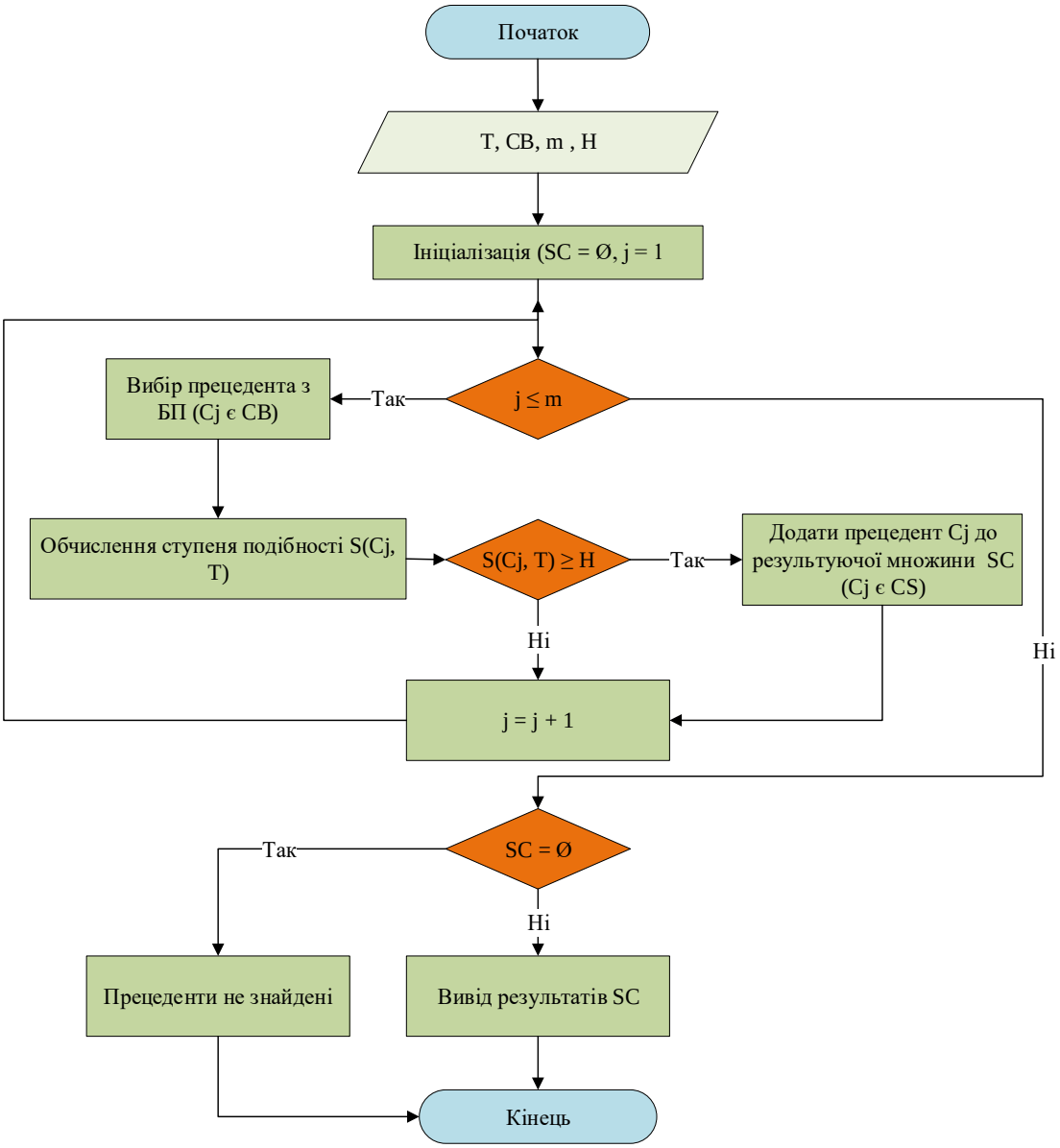

Рисунок 2.4 – Блок-схема алгоритму для витягу прецедентів

В роботі запропонована модифікація (рисунок 2.4), яка полягає в тому, що k будуть змінюватися в залежності від зміни розміру бази прецедентів (БП). Чим більше прецедентів в БП, тим більше значення можна вибрати для k (від 1 до k_{MAX}). k_{MAX} відповідає кількості елементів, що належать класу з максимальним числом прецедентів.

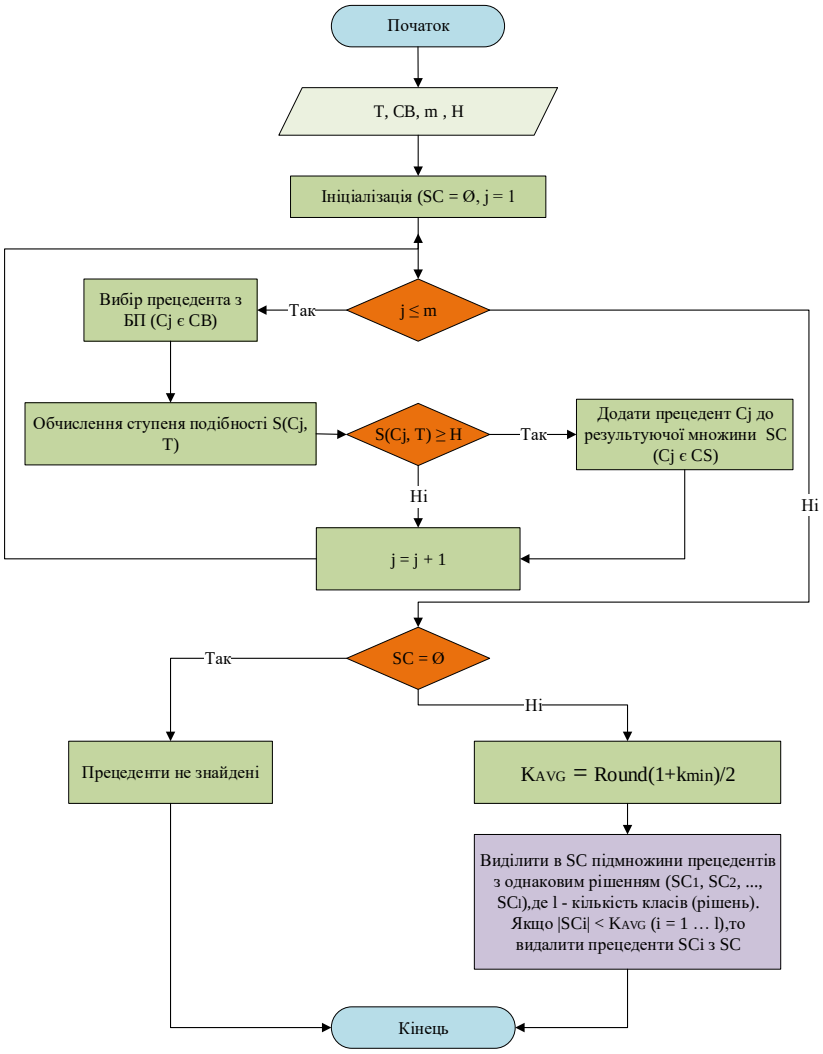

Рисунок 2.5 – Блок-схема модифікації алгоритму для витягу прецедентів на основі k-ближчих сусідів

2.3 Повторне використання прецедентів

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 видобутого прецеденту можна перенести на поточну ситуацію. У простих задачах класифікації відмінності просто ігноруються і клас рішення витягнутого прецеденту переноситися на клас рішення нового прецеденту. Це найпростіший спосіб повторного використання прецедентів. Багато системи,

однак, враховують відмінності між знайденим і наявним прецедентом, і тому рішення витягнутого прецеденту не може бути безпосередньо перенесено 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). При трансформаційному використанні рішення знайденого прецеденту не є безпосереднім рішенням поточної ситуації, однак існують деякі знання у вигляді трансформаційних операторів, застосування яких до старого рішенням перетворює його в рішення нового прецеденту. Можна впорядкувати ці оператори, проиндексировав їх щодо помічених відмінностей між витягнутим прецедентом і поточної проблемною ситуацією.

При трансформаційному використанні увага спрямована не так на те, як вирішується проблема, а на схожість двох рішень, тому системі потрібні глибокі знання про специфіку предметної області в формі трансформаційних операторів і режим управління для організації операторів в додатку.

При деривaційному використанні прецедентів увага приділяється тому, як було знайдено рішення в витягнутому прецедент. Знайдений прецедент містить інформацію про те, який метод використовувався при пошуку рішення, обгрунтованість використання операторів, передбачувані подцели, можливі альтернативи, вдалі і невдалі шляху пошуку рішення і т.д.

При деривaційному використанні до нової проблеми застосовується витягнутий метод і план пошуку рішення, тільки з урахуванням контексту нового завдання. При пошуку рішення вдалі гілки пошуку рішення, оператори, і шляхи вирішення застосовуються в першу чергу, а невдалі гілки пошуку не розглядаються. Нові мети формуються на основі старих і для їх досягнення будується новий план на основі знайденого раніше.

Таким чином, деривaційне використання дозволяє знаходити рішення нової проблемної ситуації, але з мінімальними витратами часу і ресурсів.

2.4 Збереження прецедентів

Збереження прецедентів - це процес відбору інформації з вирішення нової проблемної ситуації для збереження і подальшого використання. Збереження прецедентів включає в себе відбір інформації, яку потрібно зберегти, форму в якій вона буде збережена, індексацію для майбутнього вилучення в разі виникнення схожої ситуації і те, яким чином слід включити новий прецедент в склад БП. Якщо рішення було взято з витягнутого прецеденту, то може бути сформований новий прецедент або старий прецедент може бути узагальнено, покриваючи нову ситуацію. Якщо рішення було отримано іншими методами, наприклад, рішення знайдено самим ЛПР, то формується абсолютно новий прецедент. У будь-якому випадку має бути прийняте рішення, як поповнювати БП.

Крім опису параметрів проблемної ситуації і її вирішення, до складу прецеденту може бути включено пояснення, чому дане рішення є рішенням проблемної ситуації. У системах, в яких застосовується деривaційне використання прецедентів, зберігається метод пошуку рішення (план виведення, можливі альтернативи та т.і.). Інформація про невдалий застосуванні отриманих рішень так само може бути збережена у вигляді окремих прецедентів. Тоді в разі невдачі, система може знайти прецедент, який зберігає інформацію про схожий невдалому випадку і використовувати її для виправлення поточної ситуації.

2.5 Етапи розробки CBR-систем

Існують такі етапи:

- Вилучення знань для формування первинної безлічі прецедентів (збір інформації про ситуації і випадки, що мали місце в минулому) може здійснюватися на основі аналізу архіву системи і мали місце аварійні ситуацій, вивчення досвіду, накопиченого експертом (ЛПР), оперативних інструкцій, технологічного регламенту і т .і.

- Визначення структури для організації БП. Структура БП безпосередньо залежить від способу представлення прецедентів і специфіки проблемної області.

- Вибір методу отримання прецедентів. Метод вилучення залежить від предметної області і типу розв'язуваних завдань, а також від способу представлення прецедентів в БП і її структури.

Визначення способу адаптації прецедентів. Існує два типи адаптації прецедентів. Структурна адаптація застосовується для погано обумовлених прецедентів, а адаптація по відхиленню - для добре обумовлених. - Визначення інших аспектів сбя системи, обумовлених особливостями предметної області або типом розв'язуваної задачі.

3 РЕАЛІЗАЦІЯ ПІДСИСТЕМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Архітектура прототипу CBR системи складається з наступних основних компонентів (рисунок 3.1):

Інтерфейс – інтерфейс для взаємодії з ЛПР, експертом або користувачем і відображення результатів роботи.

Блок вилучення прецедентів – в системі реалізований ряд методів для отримання прецедентів з БП (алгоритм NN, k-NN і його модифікація).

База знань – містить базу вдалих прецедентів (БП) і базу невдалих прецедентів (БНП).

Набір тестових вибірок – експертна інформація, що містить дані (приклади) з коректними (експертними) рішеннями для їх врахування при добуванні прецедентів і для формування БП і БНП.

Модуль оптимізації БП – призначений для виконання скорочення кількості прецедентів в БП з використанням різних класифікаційних і кластерних алгоритмів (наприклад, скорочення кількості прецедентів в БП на основі алгоритму NN або скорочення кількості прецедентів в БП на основі кластерного алгоритму к-середніх).

Програмна реалізація прототипу CBR системи виконана на мові C # в середовищі програмування MS Visual Studio 2010 року (прилож. 1) для СУБД Microsoft SQL Server 2008 з використанням SQL Server Analysis Services і аналітичної платформи Deductor 5.3.

Прототип CBR системи реалізований в середовищі Microsoft Visual Studio 2010 із використанням наступних технологій і аналітичних платформ:
- Windows Forms;
- ADO.NET Entity Framework;
- аналітична платформа Deductor 5.3;
- SQL Server Analysis Services.

Windows Forms – це технологія інтелектуальних клієнтів для .NET Framework. Вона являє собою набір керованих бібліотек, що спрощують виконання стандартних завдань, таких як читання з файлової системи і запис в неї. За допомогою середовища розробки типу Visual Studio можна створювати інтелектуальні клієнтські програми Windows Forms, які відображають інформацію, запитують введення даних від користувачів і обмінюються даними з віддаленими комп'ютерами по мережі.

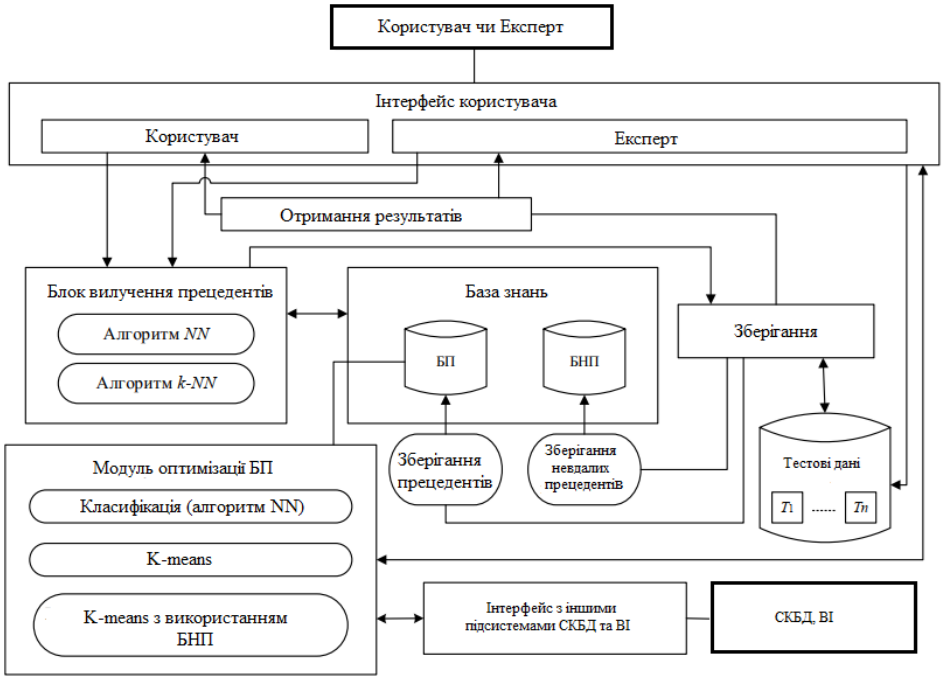

Рисунок 3.1 – Архітектура прототипу CBR-системи

ADO.NET Entity Framework (EF) – об'єктно-орієнтована технологія доступу до даних, є object-relational mapping (ORM) рішенням для .NET Framework від Microsoft. Надає можливість взаємодії з об'єктами як за допомогою LINQ у вигляді LINQ to Entities, так і з використанням Entity SQL. Для полегшення побудови web-рішень використовується як ADO.NET Data Services (Astoria), так і зв'язка з Windows Communication Foundation і Windows Presentation Foundation, що дозволяє будувати багаторівневі додатки, реалізуючи один з шаблонів проектування MVC, MVP або MVVM.

Платформа ADO.NET Entity Framework дозволяє розробникам створювати додатки для доступу до даних, що працюють з концептуальною моделлю додатки, а не безпосередньо з реляційної схемою зберігання. Мета полягає в зменшенні обсягу коду і зниженні витрат на супровід додатків, орієнтованих на обробку даних. ORM – технологія програмування, яка зв'язує бази даних з концепціями об'єктно-орієнтованих мов програмування, створюючи «віртуальну об'єктну базу даних».

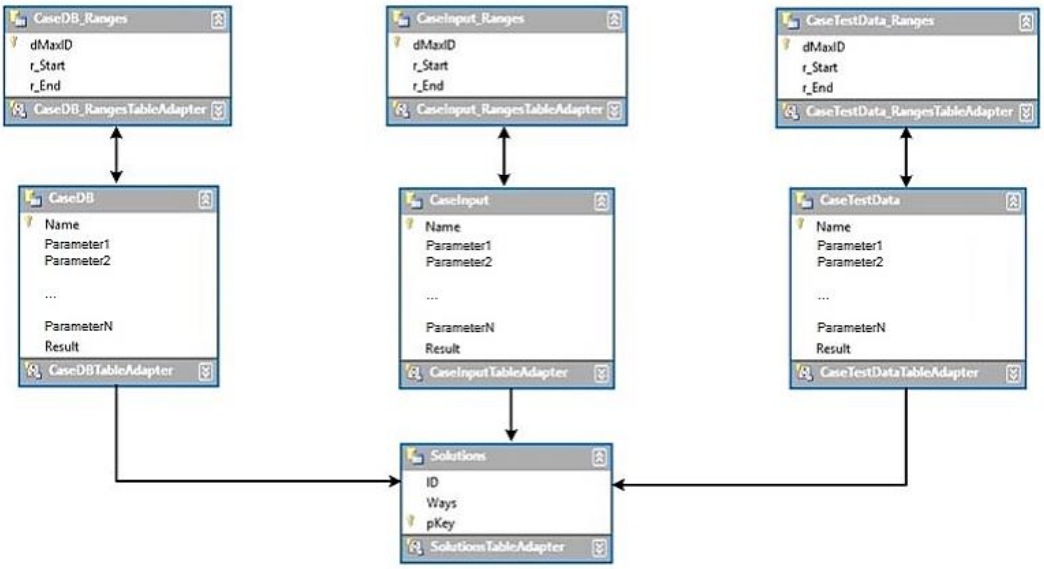

Рисунок 3.2 – Схема БД для CBR-системи

Додатки Entity Framework надають такі переваги.

- Програми можуть працювати концептуальною моделлю в термінах предметної області - в тому числі з успадкованими типами, складними елементами і зв'язками.

- Програми звільняються від жорстких залежностей від конкретного ядра СКБД або схеми зберігання.

- Зіставлення між концептуальною моделлю і схемою, специфічною для конкретного сховища, можуть змінюватися без зміни коду програми.

Deductor є аналітичною платформою BaseGroup Labs, розглянутої в огляді систем ІАД в попередньому розділі і концентрує багаторічний досвід

компанії і увібрав в себе найбільш вдалі архітектурні ідеї і сучасний математичний апарат. Deductor є платформою, на базі якої створюються закінчені аналітичні рішення. Платформа орієнтована на застосування експертами в різних предметних областях, дозволяє обробляти будь-яку структуровану табличну інформацію. Це доступна за ціною і проста у використанні система з чудовими аналітичними можливостями.

Deductor надає аналітикам інструментальні засоби, необхідні для вирішення найрізноманітніших аналітичних завдань: корпоративна звітність, прогнозування, сегментація, пошук закономірностей - ці та 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 в комплексі, так і окремо для вирішення широкого спектра бізнес-проблем.

3.1 Програмна реалізація

Робота прототипу була розглянута на прикладі наборів даних зі сховищ UCI Machine Learning Repository Каліфорнійського університету. БД з інформацією про рівень знань учнів (студентів)

На першому кроці необхідно завантажити початкову БП для даного прикладу (рисунок 3.4). Початкова БП була сформована на основі перших 20 записів з БД UCI Machine Learning Repository (рисунок 3.3), а також можна вибрати тестову і навчальну вибірки. БД з репозиторію включає 258 прикладів, що характеризуються 5 атрибутами (параметрами) і належать одному з 4 рішень (класів): 1 – дуже низький (very low), 2 – низький (low), 3 – середній (middle) і 4 – високий (high).

	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 3.3 – Приклад даних з БД *UCI Machine Learning Repository*

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

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

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

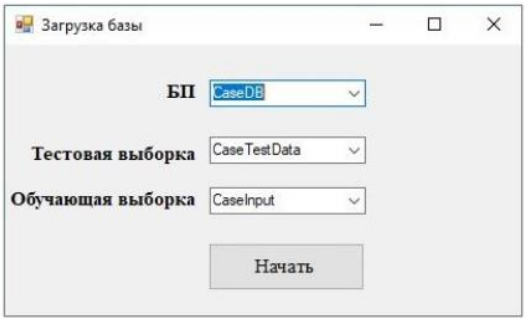

Рисунок 3.4 – Інтерфейс ПЗ

Далі можна задати нову ситуацію для пошуку рішення на основі прецедентів з БП (рисунок 3.4). Для прикладу була взята наступна ситуація:

- STG = 0.24;
- SCG = 0.1;
- STR = 0.25;
- LPR = 0.06;
- PEG = 0.9.

CaseTestData Далі натискаємо кнопку «Пошук» для отримання результатів (рисунок 3.4).

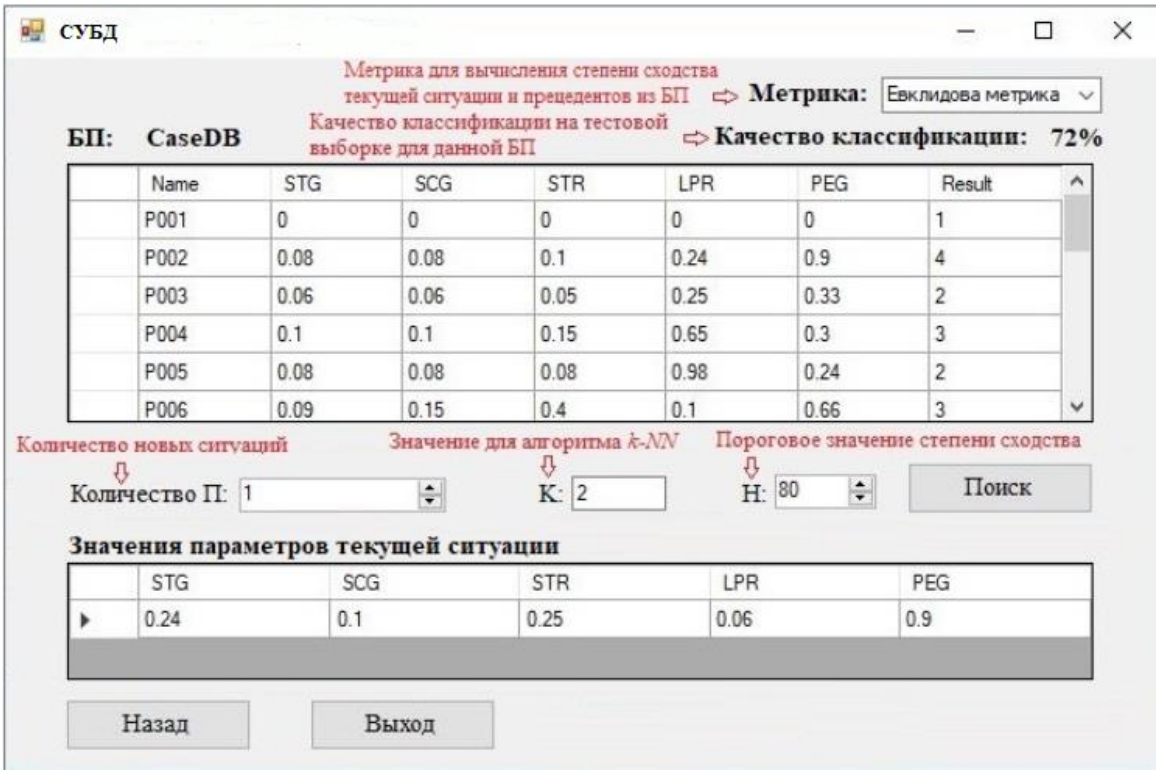

Рисунок 3.5 – Інтерфейс ПЗ

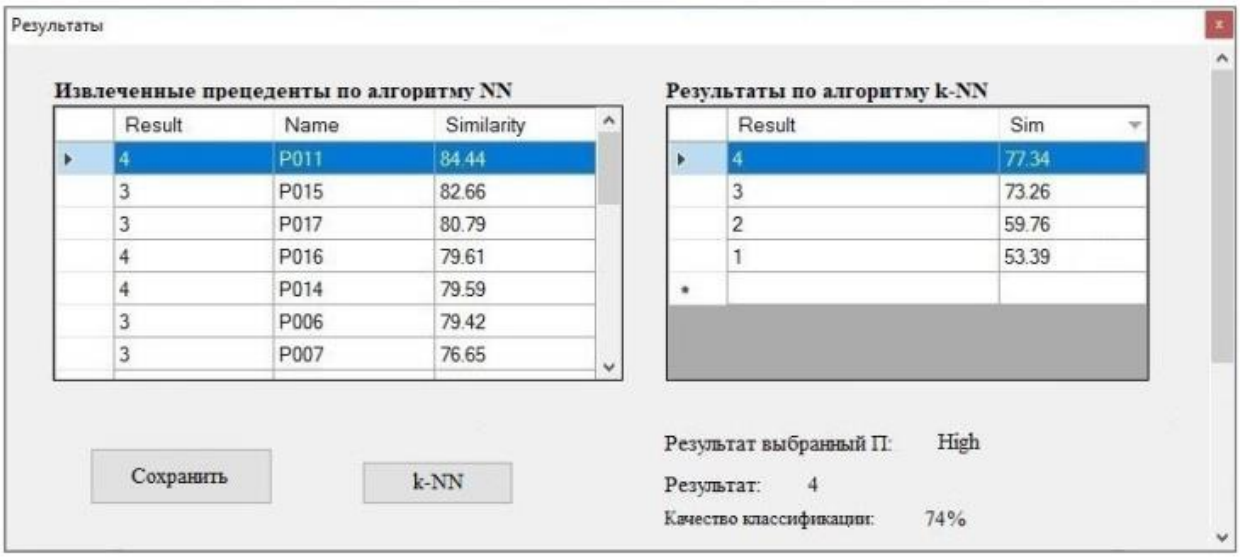

Рисунок 3.6 – Діалогове вікно з результатами

Для поточної ситуації за алгоритмом NN знайдений найближчий прецедент P011 зі ступенем схожості 84.44%, відповідно якої для поточної ситуації знайдено рішення «4» – високий (high) рівень знань того, хто навчається.

За допомогою натискання кнопки «k-NN» можна отримати результати за алгоритмом k-NN, які для даного прикладу також дають рішення «4» – високий (high) рівень знань того, хто навчається з усередненою оцінкою по найближчих прецедентів рівній 77.34%.

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

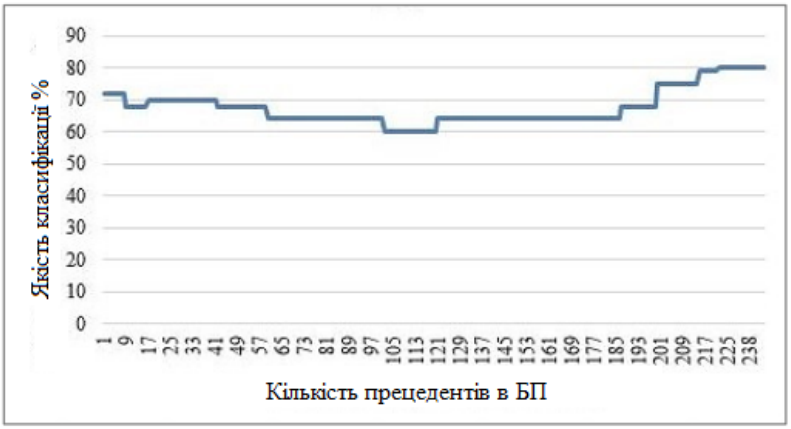

Рисунок 3.7 – Результати роботи

Наведені результати обчислювальних експериментів по оцінці якості класифікації для різних значень k.

ВИСНОВКИ

Проведено аналіз засобів та методів інтелектуального аналізу даних на основі прецедентів для систем керування базами даних. Проведено дослідження різних технологій, методів і програмних засобів ІАД, що включаються до складу сучасних СКБД. Однією з перспективних можливостей розширення засобів ІАД і аналітичних інструментів СКБД є використання прецедентного підходу. Розроблено модифікацію алгоритму вилучення прецедентів на основі k-NN для ІАД, яка полягає в зміні значення k в залежності від розміру БП. Дана модифікація дозволяє підвищити якість рішення задач ІАД, зокрема, підвищити якість класифікації даних з використанням CBR методу.

Запропонована архітектура прототипу CBR системи для ІАД, що включає в себе наступні основні компоненти: призначений для користувача інтерфейс, блок вилучення прецедентів, БЗ з БП і БНП, набір тестових вибірок і модуль оптимізації БП для скорочення кількості прецедентів в БП CBR системи. Виконана програмна реалізація прототипу CBR системи для розширення можливостей засобів ІАД в КУБД 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 плаформе Deductor 5.3 і SQL Server Analysis Services. Розглянуто приклад використання розробленого прототипу системи для вирішення задачі класифікації даних зі сховищ UCI Machine Learning Repository.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

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. - 2006. - №3. С. 39-62.

3. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта, №3, 2004, С. 3-19.

4. W. Frawley, G. Piatetsky-Shapiro, C. Matheus Knowledge Discovery in Databases: An Overview. - AI Magazine. - 1992. pp. 213-228.

5. Kitchin Rob. The Data Revolution. United States: Sage. 2014, p. 6.

6. Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Databases. USA: MIT Press, 1991.

7. 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.

8. Fayyad U., Piatetsky-Shapiro G., Smyth P., Advances in Knowledge Discovery and Data Mining, (Chapter 1), AAAI/MIT Press, 1996.

9. 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.

10. 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.

11. Интеллектуальный анализ данных средствами MS SQL Server 2008 - [Электронный ресурс].

12. Data Mining – технология добычи данных – [Электронный ресурс]. URL: http://bourabai.ru/tpoi/datamining.htm: (дата обращения: 05.01.2017).

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

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

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

16. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977.

17. А.Н.Тихонов, В.Я.Арсенин. Методы решения некорректных задач. Наука, Москва, 1974.

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. Управленческий учет: учебник / под ред. А.Д. Шеремета. 4-е изд. – М.: ИНФРА-М, 2009.

21. Публикация.