ТЕМА №1 ОСНОВНІ ПОНЯТТЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

План

1.1. Сутність аналітичних технологій
1.2. Поняття інтелектуального аналізу даних
1.3. Етапи та методи знаходження нових знань
1.4. Основні моделі інтелектуальних обчислювань
1.5. Засоби програмної підтримки інтелектуального аналізу даних
1.6. Новітні напрямки застосування Data Mining

Література:

Основна

Черняк О. І., Захарченко П. В. Інтелектуальний аналіз даних: Підручник. Київ, 2014.

Додаткова

1. Барсегян А. А. Куприянов М. С. Степаненко В. В. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. Санкт-Петербург : БХВ–Петербург, 2008.

2. Барсегян А. А., Куприянов М. С., Степаненко В. В. Методы и модели анализа данных: OLAP и Data Mining. Санкт-Петербург : БХВ–Петербург, 2004.

3. Дюк В. А., Самойленко А. П. Data Mining: учебный курс. Санкт-Петербург : Питер, 2001.

4. Кулаичев А. П. Методы и средства комплексного анализаданных. Москва : ИНФРА-М, 2006.

5. Паклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям. Санкт-Петербург : Питер, 2009.

6. Brin S. et al. Dynamic Itemset Counting and Implication Rules for MarketBasket Data. New York : ACM Press, 1997.

7. Buntine W. A Theory of classification rules. John Wiley & Sons, 1992.

8. Siu Nin Lam. Discovering Association Rules in Data Mining. Department ofComputer Science, University of Illinois at Urbana-Champaign, 2009.

1.1 Сутність аналітичних технологій

Аналітичні технології почали застосовуватися людством досить давно. Простим прикладом аналітичної технології є теорема Піфагора, яка дозволяє визначити довжину гіпотенузи маючи відомі довжини катетів по відомій формулі $a^2+b^2=c^2$. Іншим прикладом аналітичної технології можна назвати алгоритм обробки інформації людським мозком. Навіть мозок дитини може виконувати задачі, непідвладні сучасним комп'ютерам, наприклад, розпізнавання знайомих облич в юрбі або ефективне управління декількома десятками м'язів при грі у футбол. Унікальністю мозку є здібність до

розв’язання нових задач – гри в шахи, водінню автомобіля і т.д. Але при цьому, мозок погано пристосований до обробки великих об'ємів числової інформації – людина не може перемножити два багатозначні числа, не використовуючи калькулятора або алгоритму обчислення в стовпчик. Реальні завдання з числами, набагато складніше, ніж множення і людині для вирішення таких завдань необхідні додаткові методики і інструменти.

Під ***аналітичною технологією*** будемо розуміти методики, які на основі певних моделей, алгоритмів, математичних теорем дозволяють за відомими даними оцінити значення невідомих характеристик і параметрів.

Аналітичні технології потрібні в першу чергу людям, що ухвалюють важливі рішення, – керівникам, аналітикам, експертам, консультантам. Дохід компанії більшою мірою визначається якістю цих рішень – точністю прогнозів, оптимальністю вибраних стратегій. І від якості цих рішень залежить розвиток компанії. За допомогою аналітичних технологій можна вирішувати проблеми прогнозування, наприклад, курсів валют, цін на сировині, попиту, доходу компанії, рівня безробіття і оптимізації, наприклад, плану закупівель, плану інвестицій, стратегії розвитку. Слід також зазначити, що для реальних задач бізнесу і виробництва не існує чітких алгоритмів їх розв’язання. Тому керівники і експерти знаходять рішення таких задачі тільки на основі особистого досвіду. Часто класичні методики виявляються малоефективними для багатьох практичних завдань, оскільки неможливо точно описати реальність за допомогою невеликого числа параметрів моделі, або розрахунок моделі займає дуже багато часу і обчислювальних ресурсів. Аналітичні технології дозволяють створювати моделі, що істотним чином підвищують ефективність рішень.

Серед класичних підходів до аналізу даних на практиці найбільш поширеними виявилися детерміновані технології та імовірнісні технології.

У останні роки відбувається бурхливий розвиток аналітичних систем нового типу. У їх основі – технології штучного інтелекту, що імітують природні процеси, наприклад, діяльність нейронів мозку або процес природного відбору.

При розробці сучасних аналітичних технологій враховується їх здатність:

– розуміння задачі, загального процесу і знання можливостей інших систем і людей, що беруть участь у взаємодії;

– зв'язок з користувачами за допомогою розуміння природної мови, малюнків, зображень, і знаків;

– знання, засновані на здоровому глузді;

– координування ухвалення рішень, планування і дії;

– навчання на попередньому досвіді і адаптація поведінки.

1.2. Поняття інтелектуального аналізу даних

Термін *Data Mining* отримав свою назву з двох понять: пошуку цінної інформації у великій базі даних (*Data*) і видобутку гірської руди (*Mining*). Обидва процеси вимагають або просіювання величезної кількості сирого матеріалу, або розумного дослідження і пошуку корисних цінностей. Найчастіше *Data Mining* перекладається як видобуток даних, витягання інформації, розкопка даних, інтелектуальний аналіз даних, засоби пошуку закономірностей, витягання знань, аналіз шаблонів, «витягання зерен знань з гір даних», розкопка знань в базах даних, інформаційна проходка даних, «промивання» даних. Поняття «виявлення знань в базах даних» (*Knowledge Discovery in Databases, KDD*) можна вважати синонімом *Data Mining*.

У основу сучасної технології *Data Mining* покладена концепція шаблонів (паттернів), що відображають фрагменти багатоаспектних взаємин в даних. Ці шаблони є закономірностями, властивими підвибіркам даних, які можуть бути компактно виражені в зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру вибірки і виду розподілів значень аналізованих показників.

Суть і мету технології *Data Mining* можна охарактеризувати так: це технологія, яка призначена для пошуку у великих об'ємах даних неочевидних, об'єктивних і корисних на практиці закономірностей.

Неочевидних – це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом.

Об'єктивних – це означає, що виявлені закономірності повністю відповідатимуть дійсності, на відміну від експертної думки, яка завжди є суб'єктивною.

Практично корисних – це означає, що висновки мають конкретне значення, якому можна знайти практичне застосування.

Для успішного проведення процесу знаходження нового знання необхідною умовою є наявність сховища даних. ***Сховище даних*** – це предметно-орієнтований, інтегрований, прив'язаний до часу, незмінний збір даних для підтримки процесу ухвалення рішень. ***Предметна орієнтація*** означає, що дані об'єднані в категорії і зберігаються відповідно областям, що вони описують, а не до застосувань, що їх використовують. ***Інтегрованість*** означає, що дані задовольняють вимогам всього підприємства, а не одній функції бізнесу. Цим, сховище даних гарантує, що однакові звіти, що згенерували для різних аналітиків, міститимуть однакові результати. ***Прив'язка до часу*** означає, що сховище можна розглядати як сукупність «історичних» даних тобто можна відновити картину на будь-який момент часу. Атрибут часу завжди явно присутній в структурах сховища даних. ***Незмінність*** означає, що, потрапивши один раз в сховище, дані там зберігаються і не змінюються. У сховищі дані лише додаються. Для організації і експлуатації інформаційного сховища створюється

спеціалізоване програмне забезпечення, яке забезпечує ефективну взаємодію з користувачем.

Data Mining має досить суттєві відмінності від інших методів аналізу даних. Традиційні методи аналізу даних (статистичні методи) і *OLAP* в основному орієнтовані на перевірку наперед сформульованих гіпотез (*verification-driven Data Mining*) і на «грубий» розвідувальний аналіз, що становить основу оперативної аналітичної обробки даних (*Online Analytical Processing, OLAP*), тоді як одне з основних положень ***Data Mining*** – пошук неочевидних закономірностей. Інструменти *Data Mining* можуть знаходити такі закономірності самостійно і також самостійно будувати гіпотези про взаємозв'язки. Оскільки саме формулювання гіпотези щодо залежностей є найскладнішим завданням, перевага *Data Mining* в порівнянні з іншими методами аналізу є очевидною. Більшість статистичних методів для виявлення взаємозв'язків в даних використовують концепцію усереднювання по вибірці, що приводить до операцій над неіснуючими величинами, тоді як *Data Mining* оперує реальними значеннями. *OLAP* більше підходить для розуміння ретроспективних даних, *Data Mining* спирається на ретроспективні дані для отримання відповідей на питання про майбутньому.

Розглянемо деякі бізнес-приклади Data Mining.

Роздрібна торгівля. Підприємства роздрібної торгівлі сьогодні збирають докладну інформацію про кожну окрему покупку, використовуючи кредитні картки з маркою магазина і комп'ютеризовані системи контролю. Ось типові задачі, які можна вирішувати за допомогою *Data Mining* у сфері роздрібної торгівлі:

– аналіз купівельної корзини (аналіз схожості) призначений для виявлення товарів, яких покупці прагнуть придбати разом. Знання купівельної корзини необхідне для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки в торгових залах;

– дослідження часових шаблонів допомагає торговим підприємствам приймати рішення про створення товарних запасів. Воно дає відповіді на питання типу «Якщо сьогодні покупець придбав відеокамеру, то через який час він найімовірніше купить нові батареї і плівку?»;

– створення прогнозуючих моделей дає можливість торговим підприємствам дізнаватися характер потреб різних категорій клієнтів з певною поведінкою, наприклад, таких, що купують товари відомих дизайнерів або таких, що відвідують розпродажі. Ці знання потрібні для розробки точно направлених, економічних заходів щодо просування товарів.

Банківська справа. Досягнення технології *Data Mining* використовуються в банківській справі для вирішення наступних поширених завдань:

– виявлення шахрайства з кредитними картками. Шляхом аналізу минулих транзакцій, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства;

– сегментація клієнтів. Розбиваючи клієнтів на різні категорії, банки роблять свою маркетингову політику більш цілеспрямованою і результативною, пропонуючи різні види послуг різним групам клієнтів**;**

– прогнозування змін клієнтури. *Data Mining* допомагає банкам будувати прогнозні моделі цінності своїх клієнтів, і відповідним чином обслуговувати кожну категорію.

Телекомунікації. В області телекомунікацій методи *Data Mining* допомагають компаніям енергійніше просувати свої програми маркетингу і ціноутворення, щоб утримувати існуючих клієнтів і привертати нових. Серед типових заходів відзначимо наступні:

– аналіз записів про докладні характеристики викликів. Призначення такого аналізу – виявлення категорій клієнтів з схожими стереотипами користування їх послугами і розробка привабливих наборів цін і послуг;

– виявлення лояльності клієнтів. *Data Mining* можна використовувати для визначення характеристик клієнтів, які, один раз скориставшись послугами даної компанії, з великою часткою вірогідність залишиться їй вірними. У результаті засоби, що виділяються на маркетинг, можна витрачати там, де віддача більше всього.

Страхування. Страхові компанії протягом ряду років накопичують великі об'єми даних. Тут обширне поле діяльності для методів *Data Mining*:

– виявлення шахрайства. Страхові компанії можуть понизити рівень шахрайства, відшукуючи певні стереотипи в заявах про виплату страхового відшкодування, що характеризують взаємини між юристами, лікарями і заявниками;

– аналіз ризику. Шляхом виявлення поєднань чинників, пов'язаних із сплаченими заявами, страховики можуть зменшити свої втрати за зобов'язаннями. Відомий випадок, коли в США крупна страхова компанія виявила, що суми, виплачені за заявами людей, що є в браку, удвічі перевищує суми за заявами самотніх людей. Компанія відреагувала на це нове знання переглядом своєї загальної політики надання знижок сімейним клієнтам.

Управління виробництвом, менеджмент якості. Шляхом аналізу даних автоматизованого виробництва і відхилень від нього можна ідентифікувати проблеми на етапах виробництва як з погляду якості, так і з погляду збереження темпу виробництва. На підставі такої встановленої інформації можна, наприклад, у виробничий процес ввести етап додаткового контролю, завдяки якому вже в процесі виробництва будуть виявлені розроблені вироби, які після закінчення виробничого процесу не пройдуть вихідний контроль.

Молекулярна генетика і генна інженерія. Мабуть найгостріше завдання виявлення закономірностей в експериментальних даних стоїть в молекулярній генетиці і генній інженерії. Тут вона формулюється як визначення так званих маркерів, під якими розуміють генетичні коди, контролюючі ті або інші фенотипічні ознаки живого організму. Такі коди

можуть містити сотні, тисячі і більш зв'язаних елементів. На розвиток генетичних досліджень виділяються великі кошти. Останнім часом в даній області виник особливий інтерес до застосування методів *Data Mining*. Відомо декілька крупних фірм, що спеціалізуються на застосуванні цих методів для розшифровки генома людини і рослин.

Медицина. Відомо багато експертних систем для постановки медичних діагнозів. Вони побудовані головним чином на основі правил, що описують поєднання різних симптомів різних захворювань. За допомогою таких правил дізнаються не тільки, на що хворий пацієнт, але і як потрібно його лікувати. Правила допомагають вибирати засоби медикаментозної дії, визначати свідчення – протипоказання, орієнтуватися в лікувальних процедурах, створювати умови найбільш ефективного лікування, передбачати результати призначеного курсу лікування і т.п. Технології *Data Mining* дозволяють виявляти в медичних даних шаблони, які складають основу вказаних правил.

Прикладна хімія. Методи *Data Mining* знаходять широке застосування в прикладній хімії (органічної і неорганічної). Тут нерідко виникає питання про з'ясування особливостей хімічної будови тих або інших з'єднань, що визначають їх властивості. Особливо актуальне таке завдання при аналізі складних хімічних сполук, опис яких включає сотні і тисячі структурних елементів і їх зв'язків.

Можна привести ще багато прикладів різних областей знання, де методи *Data Mining* грають провідну роль. Особливість цих областей полягає в їх складній системній організації. Вони відносяться головним чином до суперскладного рівня організації систем, закономірності якого не можуть бути достатньо точно описані на мові статистичних або інших аналітичних математичних моделей. Дані у вказаних областях неоднорідні, гетерогенні, нестаціонарні і часто відрізняються високою розмірністю.

1.3. Етапи та методи знаходження нових знань

Важливо розуміти, що побудова моделі інтелектуального аналізу даних є складовою частиною масштабнішого процесу, який включає всі етапи, починаючи з визначення базової проблеми, яку модель вирішуватиме, до розгортання моделі в робочому середовищі. Даний процес може бути заданий за допомогою наступних п’яти базових кроків:

1. Постановка задачі.

2. Підготовка та огляд даних:

– оцінювання даних;

– об'єднання і очищення даних;

– відбір даних;

– перетворення.

3. Побудова моделей:

– оцінка і інтерпретація;

– зовнішня перевірка.

4. Використання моделей.

5. Нагляд за моделлю.

Може виникнути необхідність в оновленні вже розгорнених моделей за рахунок нових даних, що поступили. Таким чином, важливо розуміти, що створення моделі інтелектуального аналізу даних є процесом і що кожен крок такого процесу може бути повторений стільки раз, скільки необхідно для створення ефективної моделі. Розглянемо докладніше кожний з цих етапів:

Визначення проблеми. Для того, щоб повніше використовувати всі переваги інтелектуальних технологій необхідно ясно представити мету майбутнього аналізу. Побудова моделі проводиться залежно від мети. Якщо необхідно збільшити прибуток торгової організації, то для цілей: «збільшення кількості продажів» і «збільшення ефективності реклами» необхідно будувати різні моделі. На цьому ж етапі визначаються способи оцінювання результатів майбутнього проекту і можливі витрати на його реалізацію.

Підготовка та огляд даних. Це є найтриваліший етап, який може займати від 50% до 85% часу всього процесу знаходження нового знання. На цьому етапі необхідно визначити джерела отримання даних. Це можуть бути дані, накоплені самою організацією або зовнішні дані від загальнодоступних джерел (відомості про погоду або перепис населення) або приватних джерел (різні архівні дані, бази нотаріальних контор і ін.).

Оцінювання даних. При побудові моделі необхідно пам'ятати одне правило, що стосується коректності вхідних даних: «Якщо на вхід задачі поступає «сміття», то і результатом теж буде «сміття». Вхідні дані можуть знаходитися в одній базі або в декількох. Перед «завантаженням» даних в сховище необхідно врахувати, що різні джерела даних можуть бути спроектовані під певні задачі і, відповідно, виникають проблеми, пов'язані з об'єднанням даних: різні формати представлення числових даних (наприклад, цілі або вещественні); різне кодування даних (наприклад, різний формат дат); різні способи зберігання даних; різні одиниці вимірювання (дюйми і сантиметри); а також частота збору даних і дата останнього оновлення. Навіть, якщо дані знаходяться в одній базі, то все одно треба звертати увагу на пропущені значення і значення, нереальної величини, так звані «викиди». Аналітик винен завжди знати, як, де і за яких умов збираються дані, і бути упевненим, що всі дані, які використовуються для проведення аналізу зміряні однаковим способом.

Об'єднання і очищення даних. На цьому етапі відбувається побудова сховища даних для подальшої обробки, тобто, відбувається наповнення сховища або додавання до нього даних, відібраних на попередніх етапах. В цей же час відбувається очищення, тобто виправлення всіх виявлених помилок. Існують різні аспекти очищення даних. Всі вони направлені на знаходження і виправлення помилок, допущених на етапі збору інформації. Помилкою в даних можуть вважатися: пропущене значення, неможлива подія (невірно набране значення – «викид»). Корекція відбувається на основі

здорового глузду, використання правил або із залученням експерта, добре обізнаного з предметною областю. Запис в базі даних, в якому є помилка, повинен бути виправлений або, в спірних випадках, виключений з подальшого розгляду. Після перевірки даних, вони перетворюються і форматуються відповідно результатам оцінювання. Це робиться для більшої зручності спостереження за даними. Дані дискретних подій перетворюються на спеціально розроблену або стандартну форму, в якій відбивається час і опис подій. Якщо користувачі легко розбиратимуться в цій формі, вони зможуть швидко вивчити події, які були в основі побудови цієї форми. Може здатися, що цей крок дублює етап збору даних, але насправді це два зовсім різних етапа. На першому з них відбувається відбір даних для прискорення машинної обробки інформації без втрати якості, на другому дані доводяться до вигляду, зручному для візуального контролю користувача. Людина, яка проводить аналіз, може повніше уявити собі вхідні дані. Це необхідно для різного роду звітів, коли потрібно коротко охарактеризувати вхідні дані, вживані для аналізу.

Відбір даних. Якщо сховище сформоване і визначені типи моделей, які будуть побудовані для вирішення задач, відбувається відбір даних необхідних саме для цих моделей. Мається на увазі не тільки зменшення кількості записів в базі по певній умові, але також і зміну кількості полів, злиття різних таблиць в одну, або, навпаки, створення на основі однієї таблиці декілька. Тобто, перетворення відбувається в «трьох вимірюваннях»: по кількості записів, по кількості полів і по структурі.

Перетворення даних. Служить для збагачення отриманої бази, тобто додавання різних відносин на основі існуючих полів (не просто «ціна» і «кількість», а їх твір – «загальна сума», не борг і дохід, а відношення борга до доходу), додавання інтервалів (по номеру місяця можна поставити номер кварталу, а відсоток виконання плану можна доповнити характеристиками «добре», «задовільно»), додавання критичних значень (максимум, середнє, мінімум).

Побудова моделі є ітераційний процес, тобто, необхідно побудувати ряд моделей для знаходження однієї, що найбільш задовольняє поставленим цілям.

Моделі можна розділити на дві групи:

контрольовані (моделі класифікації, регресії, прогнозування часових послідовностей);

неконтрольовані (кластеризація, асоціація і послідовність).

Після того, як визначений тип моделі, необхідно вибрати алгоритм побудови моделі або технологію знаходження знання.

Оцінка і інтерпретація. Після побудови моделі необхідно оцінити результати і пояснити (інтерпретувати) їх значущість. При оцінці моделі обчислюється точність, але треба пам'ятати, що це значення вірно лише для даних, на яких модель побудована і бути готовим, що нові дані, до яких

надалі застосовуватиметься модель, можуть відрізнятися від результатних невідомим чином.

Зовнішня перевірка. Висока точність моделі не є гарантією того, що модель правильно відображає реальне середовище. Однією з причин, є існування так званих неявних припущень в моделі. Тобто, сам по собі коефіцієнт інфляції не може бути частиною моделі, що пояснює схильність покупців до покупки чи того іншого товару, але різка зміна цього коефіцієнта з 3% до 20% вже, напевно, може пояснити таку поведінку. Інша причина – це існування неминучих проблем з даними, що приводять до некоректності моделі, тому дуже важливо перевірити модель в реальному середовищі.

Використання моделі. Після побудови і оцінки моделі, її можна використовувати різними способами.

Спостереження за моделлю. Коли модель починає працювати в реальному середовищі, то необхідно вимірювати точність моделі на реальних даних.

Класифікація – найпоширеніша модель інтелектуального аналізу даних. З її допомогою виявляються ознаки, що характеризують групу, до якої належить той або інший об'єкт.

Регресійний аналіз використовується, коли стосунки між змінними можуть бути виражені кількісно у вигляді деякої комбінації цих змінних.

Прогнозування часових послідовностей. Основою для будь-яких систем прогнозування служить історична інформація, що зберігається в інформаційних сховищах у вигляді часових рядів.

Кластеризація відрізняється від класифікації тим, що класи заздалегідь не задані і за допомогою моделі кластеризації засоби інтелектуальних обчислень самостійно створюють однорідні групи даних.

Асоціація відноситься до аналізу структури і застосовується, коли декілька подій зв'язано між собою.

1.4. Основні моделі інтелектуальних обчислювань

Розглянемо основні види моделей, які використовуються для знаходження нового знання на основі даних інформаційного сховища. Метою інтелектуальних технологій є знаходження нового знання, яке користувач може надалі застосувати для поліпшення результатів своєї діяльності. Результат моделювання – це виявлення відносин в даних.

На практиці широке застосування знайшли такі види (алгоритми) інтелектуальних обчислень:

– нейронні мережі;
– дерева рішень;
– системи роздумів на основі аналогічних випадків;
– алгоритми визначення асоціацій і послідовностей;
– нечітка логіка;
– генетичні алгоритми;

– еволюційне програмування;
– візуалізація даних;
– комбіновані методи.

Нейронні мережі це системи з архітектурою, що умовно імітують роботу нейронів. Математична модель нейрона є деяким універсальним нелінійним елементом з можливістю широкої зміни і настроювання його характеристик. Нейронні мережі є сукупністю зв'язаних між собою прошарків нейронів, які отримують вхідні дані, здійснюють їх обробку і генерують на виході результат. Між вузлами видимих вхідного і вихідного прошарків може знаходитися певне число прихованих прошарків. Нейронні мережі реалізують непрозорий процес. Це означає, що побудована модель, як правило, не має чіткої інтерпретації. Багато пакетів, які реалізують алгоритми нейронних мереж, застосовуються у сфері обробки комерційної інформації, при розпізнаванні образів, розшифровки рукописного тексту, інтерпретації кардіограм. Апаратні або програмні реалізації алгоритмів нейромереж називаються нейрокомп'ютером.

Дерева рішень – метод, широко вживаний в області фінансів і бізнесу, де частіше зустрічаються задачі числового прогнозу. В результаті застосування цього методу, для навчальної вибірки даних створюється ієрархічна структура правил класифікації типу, «ЯКЩО... ТОДІ...», що мають вид дерева. Для того, щоб вирішити, до якого класу віднести деякий об'єкт або ситуацію, треба відповісти на питання, що стоїть у вузлах цього дерева, починаючи з його кореня. Питання можуть мати вигляд «Значення параметра А більше Х ?» або вигляду «Значення змінної В належить підмножині ознак З ?». Якщо відповідь позитивна, перехід до правого вузла наступного рівня, якщо негативний – то до лівого вузла; потім знову відповідь на питання, пов'язане з відповідним вузлом. Таким чином, врешті-решт, можна дійти до одного з кінцевих вузлів, де визначений клас об'єкту.

Системи роздумів на основі аналогічних випадків. Ідея алгоритму проста. Для того, щоб зробити прогноз майбутнього або вибрати правильне рішення, системи знаходять у минулому близькі аналоги наявної ситуації і вибирають ту ж відповідь, що була для них правильною. Тому, цей метод ще називають методом «найближчого сусіда». Системи роздумів на основі аналогічних випадків дають добрі результати в різних завданнях.

Алгоритми виявлення асоціацій знаходять правила про окремі предмети, які з'являються разом в одній транзакції, наприклад в одній покупці. Послідовність – ця теж асоціація, але залежна від часу. Асоціація записується як $А \rightarrow Б$, де $А$ називається передумовою, $Б$ – наслідком. Частота появи кожного окремого предмету або групи предметів, визначається дуже просто – підраховується кількість появи цього предмету у всіх подіях (покупках) і ділиться на загальну кількість подій. Ця величина вимірюється у відсотках і носить назву «поширеність». Низький рівень поширеності (менш одного тисячною відсотка) говорить про неістотність асоціації.

Нечітка логіка застосовується для наборів даних, де приналежність даних до якої-небудь групи є вірогідністю в інтервалі від 0 до 1. Чітка логіка маніпулює результатами, які можуть бути або істиною, або ложью. Нечітка логіка застосовується в тих випадках, коли існує «може бути» в доповненні до «так» чи ні». Областю впровадження алгоритмів нечіткої логіки є будь-які аналітичні системи.

Генетичні алгоритми є могутнім засобом рішення різних комбінаторних задач і проблем оптимізації. Проте, генетичні алгоритми увійшли зараз до стандартного інструментарію методів інтелектуальних обчислень. Цей метод названий так тому, що якоюсь мірою імітує процес природного відбору в природі. Хай нам треба знайти рішення задач, найбільш оптимальні з погляду деякого критерію, де кожне рішення цілком описується певним набором чисел або величин нечислової природи. Скажімо, якщо нам треба вибрати сукупність фіксованого числа параметрів ринку, що істотно впливають на його динаміку, це буде набір імен цих параметрів. Про цей набір можна говорити як про сукупність хромосом, що визначають якості індивіда, – даного рішення поставленої задачі. Значення параметрів, що визначають рішення, називаються генами. Пошук оптимального рішення при цьому схожий на еволюцію популяції індивідів, представлених наборами хромосом.

Еволюційне програмування наймолодша область інтелектуальних обчислень. Гіпотези про вид залежності цільової змінної від інших змінних формулюються системою у вигляді програм на деякій внутрішній мові програмування. Якщо це універсальна мова, то теоретично на ній можна виразити залежність будь-якого вигляду. Процес побудови таких програм будується як еволюція в світі програм (цим метод трохи схожий на генетичні алгоритми). Якщо система знаходить програму, яка точно виражає залежність, яка шукається, вона починає вносити до неї невеликі модифікації і відбирає серед побудованих таким чином дочірніх програм ті, які підвищують точність.

Програми візуалізації даних в певному значенні не є засобом аналізу інформації, оскільки вони тільки представляють її користувачеві.

Комбіновані методи. Часто виробники об’єднують вказані підходи. Об'єднання алгоритмів нейронних мереж і технології дерев рішень сприяє побудові точнішої моделі і підвищенню швидкості. Для вирішення кожної проблеми слід шукати свій оптимальний метод.

1.5. Засоби програмної підтримки інтелектуального аналізу даних

Інструменти *Data Mining* можна оцінювати по різних критеріях. Оцінка програмних засобів *Data Mining* з погляду кінцевого користувача визначається шляхом оцінки набору його характеристик. Їх можна поділити на дві групи: бізнес-характеристики і технічні характеристики. Цей розподіл

є достатньо умовним, і деякі характеристики можуть потрапляти одночасно в обидві категорії.

Інтуїтивний інтерфейс. *Інтерфейс* – середовище передачі інформації між програмним середовищем і користувачем, діалогова система, яка дозволяє передати людині всі необхідні дані, отримані на етапі формалізації і обчислення. Він припускає розташування різних елементів, в т.ч. блоків меню, інформаційних полів, графічних блоків, блоків форм, на екранних формах. Для зручності роботи користувача необхідно, щоб інтерфейс був інтуїтивним.

Зручність експорту - імпорту даних. При роботі з інструментом *Data Mining* користувач часто застосовує різноманітні набори даних, працює з різними джерелами даних.

Наочність і різноманітність отримуваної звітності. Ця характеристика припускає отримання звітності в термінах предметної області, а також в якісно спроектованих вихідних формах в тій кількості, яка може надати користувачеві всю необхідну результативну інформацію.

Зручність і простота використання. Істотно полегшує роботу користувача можливість використовувати програми Майстер.

Можливості візуалізації. Наявність графічного представлення інформації істотно полегшує інтерпритованість отриманих результатів.

Наявність значень параметрів, заданих за умовчанням. Для користувачів, що починають, – це достатньо істотна характеристика, оскільки при виконанні багатьох алгоритмів від користувача потрібне завдання або вибір великого числа параметрів.

Кількість методів і алгоритмів. У багатьох інструментах *Data Mining* реалізоване відразу декілька методів, що дозволяють вирішувати одну або декілька задач.

Можливості пошуку, сортування, фільтрації. Така можливість корисна як для вхідних даних, так і для вихідної інформації. Застосовується сортування по різних критеріях (полям), з можливістю накладення умов. За умови фільтрації вхідних даних з'являється можливість побудови моделі Data Mining на одній з вибірок набору даних.

Захист, пароль. Дуже часто за допомогою *Data Mining* аналізується конфіденційна інформація, тому наявність пароля доступу в систему є бажаною характеристикою для інструменту.

Ринок інструментів *Data Mining* визначається широтою цієї технології і внаслідок цього – величезним різноманіттям програмного забезпечення. Найбільш популярна група інструментів містить наступні категорії:

– набори інструментів;

– класифікація даних;

– кластеризація і сегментація;

– інструменти статистичного аналізу;

– аналіз текстів (*Text Mining*), витягання відхилень (*Information Retrieval*);

– інструменти візуалізації.

1.6. Новітні напрямки застосування Data Mining

Серед новітних напрямків застосування технологій інтелектуального аналізу даних слід виділити Web-задачі, практика вирішення яких викликає поширений інтерес і набуває популярності.

Web Mining. *Web Mining* можна перевести як «видобуток даних в *Web*». *Web Intelligence* готовий «відкрити нову главу» в стрімкому розвитку електронного бізнесу.

Text Mining. *Text Mining* охоплює нові методи для виконання семантичного аналізу текстів, інформаційного пошуку і управління. Синонімом поняття *Text Mining* є *KDT* (*Knowledge Discovering in Text* – пошук або виявлення знань в тексті). На відміну від технології *Data Mining*, яка передбачає аналіз впорядкованої в деякі структури інформації, технологія Text Mining аналізує великі і надвеликі масиви неструктурованої інформації.

Call Mining. За словами аналітиків «видобуток дзвінків» може стати популярним інструментом корпоративних інформаційних систем. Технологія *Call Mining* об'єднує в собі розпізнавання мови, її аналіз і *Data Mining*. Її мета – спрощення пошуку в аудіо-архівах, що містять записи переговорів між операторами і клієнтами.