Тема 5. Задачі Data Mining та їх класифікація. Інформація та знання

План

1. Задачі Data Mining.
2. Класифікація задач інтелектуального аналізу даних.
3. Рівні аналізу.
4. Інформація. Властивості інформації.

Мета вивчення теми: вивчити задачі інтелектуального аналізу даних; засвоїти рівні аналізу Data Mining; засвоїти поняття інформації та вивчити її властивості.

Перелік ключових слів та понять із теми

Data Mining, інформація, класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків

Теоретичні відомості з теми

1. Задачі Data Mining

В основу технології Data Mining покладена концепція шаблонів, що представляють собою закономірності. У результаті виявлення цих, прихованих від неозброєного ока закономірностей вирішуються завдання інтелектуального аналізу даних

Задачі (tasks) Data Mining іноді називають закономірностями (regularity) або техніками (techniques).

Єдиної думки щодо того, які задачі слід відносити до Data Mining, немає.

Більшість авторитетних джерел перераховують такі задачі: класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків.

Класифікація (Classification). Найбільш проста і поширена задача Data Mining. У результаті розв'язання задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних – класи; за цими ознаками новий об'єкт можна віднести до того чи іншого класу.

Методи розв'язання. Для розв'язання задачі класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor), *k*-найближчого сусіда (*k*-Nearest Neighbor); байєсовські мережі (Bayesian Networks); індукція дерев рішень; нейронні мережі (neural networks).

Кластеризація (Clustering) є логічним продовженням ідеї класифікації. Ця задача більш складна. Особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи.

Приклад методу розв'язання задачі кластеризації: навчання «без вчителя» особливого виду нейронних мереж – самоорганізованих карт Кохонена.

Асоціація (Associations). У ході розв'язання задачі пошуку асоціативних правил відшукуються закономірності між пов'язаними подіями в наборі даних.

Відмінність асоціації від двох попередніх задач Data Mining полягає в тому, що пошук закономірностей здійснюється не на основі властивостей аналізованого об'єкта, а між кількома подіями, які відбуваються одночасно.

Найбільш відомий алгоритм розв'язання задачі пошуку асоціативних правил – алгоритм Apriori.

Послідовність (Sequence), або послідовна асоціація (sequential association). Послідовність дозволяє знайти тимчасові закономірності між транзакціями. Задача послідовності подібна асоціації, але її метою є встановлення закономірностей не між подіями, що наступають одночасно, а між подіями, які пов'язаними в часі (тобто відбуваються з деяким певним інтервалом у часі). Іншими словами послідовність визначається високою ймовірністю ланцюжка пов'язаних у часі подій.

Фактично, асоціація є окремим випадком послідовності з тимчасовим лагом, рівним нулю. Цю задачу Data Mining також називають задачею знаходження послідовних шаблонів (sequential pattern).

Правило послідовності: після події X через певний час відбудеться подія Y.

Приклад. Після покупки квартири мешканці в 60% випадків протягом двох тижнів купують холодильник, а протягом двох місяців в 50% випадків купується телевізор. Розв'язок цієї задачі широко застосовується в маркетингу і менеджменті, наприклад, при управлінні циклом роботи з клієнтом (управління життєвим циклом клієнта).

Прогнозування (Forecasting). У результаті розв'язання задачі прогнозування на основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників.

Для розв'язання таких задач широко застосовуються методи математичної статистики, нейронні мережі тощо.

Визначення відхилень або викидів (Deviation Detection), аналіз відхилень або викидів. Мета розв'язання цієї задачі – виявлення та аналіз даних, що найбільш відрізняються від загальної множини даних, виявлення так званих нехарактерних шаблонів.

Оцінювання (оцінка). Задача оцінювання зводиться до передбачення неперервних значень ознаки.

Аналіз зв'язків (Link Analysis) – задача знаходження залежностей в наборі даних.

Візуалізація (Visualization, Graph Mining). У результаті візуалізації створюється графічний образ аналізованих даних. Для розв'язання задачі візуалізації використовуються графічні методи, що показують наявність закономірностей у даних.

Приклад методу візуалізації – подання даних у 2-D і 3-D вимірах.

Підведення підсумків (Summarization) – задача, мета якої – опис конкретних груп об'єктів з аналізованого набору даних.

2. Класифікація задач інтелектуального аналізу даних

Згідно класифікації за стратегіями, задачі Data Mining поділяються на такі групи:

- навчання з учителем;
- навчання без вчителя;
- інші.

Категорія «навчання з учителем» представлена такими задачами Data Mining: класифікація, оцінка, прогнозування.

Категорія «навчання без вчителя» представлена задачею кластеризації.

До категорії «інші» належать задачі, не включені в попередні дві стратегії.

Задачі інтелектуального аналізу даних, залежно від використовуваних моделей, можуть бути **дескриптивними** і **прогнозуючими.**

Відповідно до цієї класифікації, **задачі Data Mining** представлені **групами описових і прогнозуючих задач.**

У результаті розв'язання описових (descriptive) задач аналітик отримує шаблони, що описують дані, які піддаються інтерпретації.

Ці задачі описують загальну концепцію аналізованих даних, визначають інформативні, підсумкові, відмінні особливості даних. Концепція описових задач передбачає характеристику і порівняння наборів даних.

Характеристика набору даних забезпечує короткий і стислий опис деякого набору даних.

Порівняння забезпечує порівняльний опис двох або більше наборів даних.

Прогнозуючі (predictive) задачі ґрунтуються на аналізі даних, створенні моделі, передбаченні тенденцій або властивостей нових або невідомих даних.

Досить близьким до вищезгаданої класифікації є розділення задач Data Mining на такі:

a) дослідження та відкриття;
b) прогнозування та класифікація;
c) пояснення й опис.

Автоматичне дослідження і відкриття (вільний пошук). *Приклад задачі: виявлення нових сегментів ринку.*

Для розв'язання цього класу задач використовуються методи кластерного аналізу прогнозування та класифікація.

Приклад задачі: передбачення зростання обсягів продажів на основі поточних значень.

Методи: регресія, нейронні мережі, генетичні алгоритми, дерева рішень.

Задачі класифікації та прогнозування становлять групу так званого індуктивного моделювання, в результаті якого забезпечується вивчення

аналізованого об'єкта або системи. У процесі вирішення цих завдань на основі набору даних розробляється загальна модель або гіпотеза.

Пояснення й опис. *Приклад задачі: характеристика клієнтів за демографічними даними і історіями покупок.*

Методи: дерева рішень, системи правил, правила асоціації, аналіз зв'язків.

Якщо дохід клієнта більше, ніж 50 умовних одиниць, і його вік – понад 30 років, тоді клас клієнта – перший.

В інтерпретації узагальненої моделі аналітик отримує нове знання. Групування об'єктів відбувається на основі їх подібності.

Нагадаємо, що головна цінність Data Mining – це практична спрямованість даної технології, шлях від сирих даних до конкретного знання, від постановки завдання до готового додатку, за підтримки якого можна приймати рішення.

Велика кількість понять, які об'єдналися в Data Mining, а також різноманітність методів, що підтримують дану технологію, починаючому аналітику можуть нагадати мозаїку, частини якої мало пов'язані між собою.

Як же ми можемо зв'язати в одне ціле задачі, методи, дії, закономірності, додатки, дані, інформацію, рішення?

Розглянемо два потоки:

1. Дані – інформація – знання і рішення.
2. Завдання – дії і методи розв'язання – програми.

Ці потоки є «двома сторонами однієї медалі», відображенням одного процесу, результатом якого має бути знання і прийняття рішення.

Від даних до рішень. Для початку розглянемо перший потік. На рис. 5.1 показано зв'язок понять «дані», «інформація» і «рішення», яка виникає в процесі прийняття рішень.

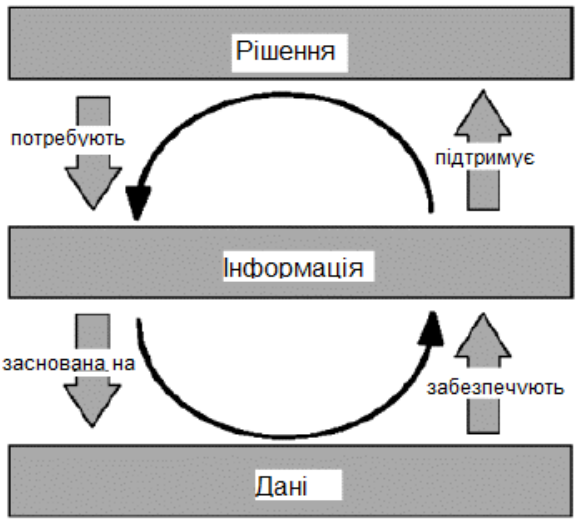

Рисунок 5.1 – Рішення, інформація і дані

Як видно з рисунку, цей процес є циклічним. Прийняття рішень потребує інформації, яка заснована на даних. Дані забезпечують інформацію, яка підтримує рішення і т.д.

Розглянуті поняття є складовою частиною так званої інформаційної піраміди, в основі якої знаходяться дані, наступний рівень – це інформація, потім йде рішення, завершує піраміду рівень знання. У міру просування вгору по інформаційній піраміді обсяги даних переходять у цінність рішень, тобто цінність для бізнесу. А, як відомо, метою Business Intelligence є перетворення обсягів даних у цінність бізнесу.

Тепер підійдемо до цього ж процесу з іншого боку. Розглянемо рис. 5.2.

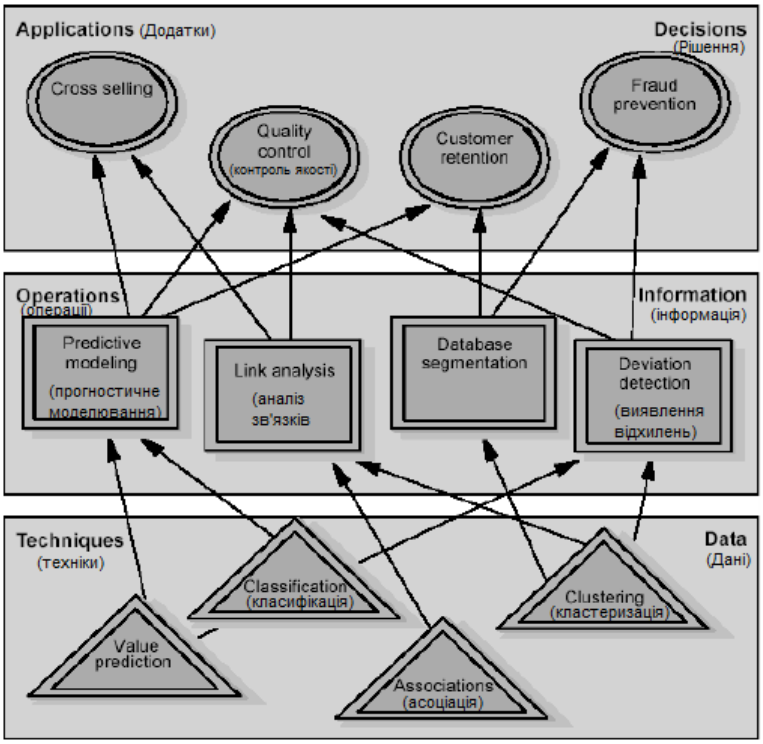

Рисунок 5.2 – Задачі, дії, додатки

Слід зазначити, що рівні аналізу (дані, інформація, знання) практично відповідають етапам еволюції аналізу даних, яка відбувалася протягом останніх років.

3. Рівні аналізу

Верхній – рівень додатків – є рівнем бізнесу (якщо ми маємо справу із завданням бізнесу), на ньому менеджери приймають рішення. Наведені приклади додатків: перехресні продажі, контроль якості, утримування клієнтів.

Середній – рівень дій – за своєю суттю є рівнем інформації, саме на ньому виконуються дії Data Mining; на рисунку наведені такі дії: прогностичне моделювання, аналіз зв'язків, сегментація даних та інші.

Нижній – рівень визначення задачі інтелектуального аналізу даних, яку необхідно розв'язати стосовно даних, що є в наявності, на рисунку наведені завдання передбачення числових значень, класифікація, кластеризація, асоціація.

Розглянемо таблицю, що демонструє зв'язок цих понять.

Таблиця 5.1 − Рівні Data Mining

Рівень 3	Додатки	Утримання клієнтів	Знання DM	Результат
Рівень 2	Дії	Прогностичне моделювання	Інформація	Метод аналізу
Рівень 1	Задачі	Класифікація	Дані	Запити

Нагадаємо, що для розв'язання задачі класифікації результати роботи першої стадії (індукції правил) використовуються для віднесення нового об'єкта з певною впевненістю до одного з відомих, визначених класів на підставі відомих значень.

Розглянемо задачі утримання клієнтів (визначення надійності клієнтів фірми).

Дані – база даних за клієнтами. Є дані про клієнта (вік, стать, професія, дохід). Певна частина клієнтів, скориставшись продуктом фірми, залишилася їй вірною; інші клієнти більше не купували продукти фірми. На цьому рівні визначаємо тип задачі – це задача класифікації.

На другому рівні визначаємо дію – прогностичне моделювання. За допомогою прогностичного моделювання ми з певною частиною впевненості можемо віднести новий об'єкт, у цьому випадку, нового клієнта, до одного з відомих класів – постійний клієнт, або це, швидше за все, його разова покупка.

На третьому рівні ми можемо скористатися додатком для прийняття рішення. У результаті придбання знань, фірма може істотно знизити витрати, наприклад, на рекламу, знаючи заздалегідь, яким із клієнтів слід активно розсилати рекламні матеріали.

Отже, ми визначилися з поняттями «дані», «завдання», «методи», «дії».

4. Інформація. Властивості інформації

Інформація (лат. informātiō) –

1) будь-які повідомлення про що-небудь;

2) відомості, що є об'єктом зберігання, переробки і передачі (наприклад, генетична інформація);

3) у математиці (кібернетиці) – кількісна міра усунення невизначеності (ентропія), міра організації системи; в теорії інформації – розділ кібернетики, що вивчає кількісні закономірності, які пов'язані зі збором, передачею, перетворенням і обчисленням інформації.

Інформація – будь-які, невідомі раніше відомості про якусь подію, сутності, процеси і т.п., є об'єктом деяких операцій, для яких існує змістовна інтерпретація.

Під операціями тут мається на увазі сприйняття, передача, перетворення, зберігання і використання. Для сприйняття інформації необхідна деяка сприймаюча система, яка може інтерпретувати її, перетворювати, визначати відповідність певним правилам і т.п. Отже,

поняття інформації слід розглядати тільки при наявності джерела і одержувача інформації, а також каналу зв'язку між ними.

Властивості інформації:

• **повнота інформації.** Це властивість характеризує якість інформації і визначає достатність даних для прийняття рішень, тобто інформація повинна містити весь необхідний набір даних.

Приклад. «Продажі товару *А* почнуть скорочуватися». Ця інформація неповна, оскільки невідомо, коли саме вони почнуть скорочуватися.

Приклад повної інформації. «Починаючи з першого кварталу, продажі товару *А* почнуть скорочуватися». Цієї інформації достатньо для прийняття рішень;

• **достовірність інформації.** Інформація може бути **достовірною** і **недостовірною**. У недостовірній інформації присутній інформаційний шум, і чим він вищий, тим нижче достовірність інформації;

• **цінність інформації.** Цінність інформації не може бути абстрактною. Інформація повинна бути корисною і цінною для певної категорії користувачів;

• **адекватність інформації.** Ця властивість характеризує ступінь відповідності інформації реальному об'єктивному стану. Адекватна інформація – це повна і достовірна інформація;

• **актуальність інформації.** Інформація повинна бути актуальною, тобто НЕ застарілою. Ця властивість інформації характеризує ступінь відповідність інформації справжньому моменту часу;

• **ясність інформації.** Інформація повинна бути зрозуміла для того кола осіб, для якого вона призначена;

• **доступність інформації.** Доступність характеризує міру можливості отримати певну інформацію. На цю властивість інформації впливають одночасно доступність даних і доступність адекватних методів;

• **суб'єктивність інформації.** Інформація носить суб'єктивний характер, вона визначається ступенем сприйняття суб'єкта (одержувача інформації).

Вимоги, що пред'являються до інформації:

• **динамічний характер інформації.** Інформація існує тільки в момент взаємодії даних і методів, тобто в момент інформаційного процесу. Решту часу вона перебуває в стані даних;

• **адекватність використовуваних методів.** Інформація витягується з даних. Проте в результаті використання одних і тих даних може з'являтися різна інформація. Це залежить від адекватності вибраних методів обробки вихідних даних.

Дані, за своєю суттю, є **об'єктивними**. **Методи** є **суб'єктивними**, в основі методів лежать алгоритми, суб'єктивно складені та підготовлені. Отже, інформація виникає та існує в момент діалектичної взаємодії об'єктивних даних і суб'єктивних методів.

Для бізнесу інформація є вихідною складової прийняття рішень.

Всю інформацію, що виникає в процесі функціонування бізнесу та управління ним, можна класифікувати певним чином. Залежно від **джерела одержання, інформацію** поділяють на **внутрішню** і **зовнішню** (наприклад, інформація, що описує явища, які відбуваються за межами фірми, але мають до неї безпосереднє відношення).

Також **інформація** може бути класифікована на **фактичну** і **прогнозну**. До фактичної інформації про бізнес відноситься інформація, що характеризує доконаний факт, вона є точною. Прогнозна інформація розраховується або передбачається, тому її не можна вважати точною, вона може мати певну похибку.

Знання – сукупність фактів, закономірностей і евристичних правил, за допомогою яких вирішується поставлене завдання.

Отже, формування інформації відбувається в процесі збору та передачі, тобто обробки даних. Яким же чином із інформації отримують знання?

Усе частіше істинні знання утворюються на основі розподілених взаємозв'язків різнорідної інформації. Коли інформація зібрана і передана для отримання явно не визначеного заздалегідь результату, то ви отримуєте знання. Сама по собі інформація в чистому вигляді безглузда. Звідси випливає висновок, що інформація – це чиєсь тактичне знання, передане у вигляді символів і за допомогою будь-яких прикладних засобів.

За визначенням Денхема Грея, «знання – це абсолютне використання інформації і даних, спільно з потенціалом практичного досвіду людей, здібностями, ідеями, інтуїцією, переконаністю і мотиваціями».

Знання мають певні **властивості**, які відрізняють їх від інформації.

1. Структурованість. Знання повинні бути «розкладені по поличках».

2. Зручність доступу і засвоєння. Для людини – це здатність швидко зрозуміти і запам'ятати або, навпаки, згадати, для комп'ютерних знань – засоби доступу до знань.

3. Лаконічність. Лаконічність дозволяє швидко освоювати і переробляти знання і підвищує «коефіцієнт корисного змісту». У цей список лаконічність була додана через усім відому проблему шуму і сміттєвих документів, характерних саме для комп'ютерної інформації – Інтернету та електронного документообігу.

4. Несуперечливість. Знання не повинні суперечити одне одному.

5. Процедури обробки. Знання потрібні для того, щоб їх використовувати. Одна з головних властивостей знань – можливість їх передачі іншим і здатність робити висновки на їх основі. Для цього повинні існувати процедури обробки знань. Здатність робити висновки означає для машини наявність процедур обробки та виведення і підготовленість структур даних для такої обробки, тобто наявність спеціальних форматів знань.

Зіставлення і порівняння понять «інформація», «дані», «знання».

Для того, щоб впевнено оперувати поняттями «інформація», «дані», «знання», необхідно не тільки розуміти суть цих понять, а й відчути відмінності між ними. Однак, однієї інтуїтивної інтерпретації цих понять тут недостатньо. Складність розуміння відмінностей вищезазначених понять – в

їх уявній синонімічності. Згадаймо, що поняття Data Mining переводиться на українську мову за допомогою цих же трьох понять: як видобуток даних, вилучення інформації, розкопування знань.

Для початку зробимо спробу розібратися в цих термінах на простих прикладах.

1. Студент, який здає іспит, потребує даних.
2. Студент, який здає іспит, потребує інформації.
3. Студент, який здає іспит, потребує знань.

При розгляді першого варіанту – студент потребує даних – виникає думка, що студенту потрібні дані, наприклад, для обчислень. Інформацією в другому варіанті може виступати конспект або підручник. У результаті їх використання студент отримує лише інформацію, яка в певних випадках може перейти у знання. Третій варіант звучить найбільш логічно.

Інформація, на відміну від даних, має сенс.

Поняття «інформація» і «знання», з філософської точки зору, є поняттями більш високого рівня, ніж «дані», яке виникло відносно недавно.

Поняття «інформації» безпосередньо пов’язано із сутністю процесів усередині інформаційної системи, тоді як поняття «знання» швидше орієнтоване на якість процесів. Поняття «знання» тісно пов’язане з процесом прийняття рішень.

Незважаючи на відмінності, розглянуті поняття, як уже зазначалося раніше, не є розрізненими і непов’язаними. Вони є частиною одного потоку: біля витоку його знаходяться дані, у процесі передачі яких виникає інформація, і в результаті використання інформації, за певних умов, виникають знання.

У процесі руху вгору в інформаційній піраміді обсяги даних переходять у цінність знань. Однак великі обсяги даних зовсім не означають і, тим більше, не гарантують отримання знань. Існує певна залежність цінності отриманих знань від якості та потужності процедур обробки даних. Типовим прикладом інформації, яку не можна перетворити в знання, є текст іноземною мовою. За відсутності словника і перекладача ця інформація взагалі не має цінності, вона не може перейти в знання. За наявності словника процес переходу від інформації до знання можливий, але тривалий і трудомісткий. За наявності перекладача інформація дійсно переходить в знання.

Таким чином, для отримання цінних знань необхідні якісні процедури обробки. Процес переходу від даних до знань займає багато часу і коштує дорого. Тому очевидно, що технологія Data Mining з її потужними і різноманітними алгоритмами є інструментом, за допомогою якого, просуваючись вгору по інформаційній піраміді, ми можемо отримувати дійсно якісні та цінні знання.

Питання для самоконтролю

1. Дайте визначення класифікації (Classification).

2. Дайте визначення кластеризації (Clustering).
3. На які групи поділяють задачі Data Mining?
4. Який зв’язок між поняттями «дані», «інформація» і «рішення»? Намалюйте схему їх зв’язку.
5. Які рівні аналізу ви знаєте?
6. Які властивості мають знання?