Тема 7. Задачі Data Mining. Прогнозування та візуалізація

План

1. Задачі прогнозування.
2. Прогнозування і часові ряди.
3. Тренд, сезонність і цикл.
4. Види помилок та прогнозів.
5. Візуалізація інструментів Data Mining.
6. Методи візуалізації.
7. Принципи компонування візуальних засобів.
8. Основні тенденції в області візуалізації.

Мета вивчення теми: вивчити задачі прогнозування; засвоїти особливості візуалізації даних.

Перелік ключових слів та понять із теми

Прогнозування, часовий ряд, ряд динаміки, прогноз, прогностика, похибки прогнозу, тренд, сезонність, цикл, візуалізація

Теоретичні відомості з теми

1.Задачі прогнозування

Задачі прогнозування розв'язуються в найрізноманітніших сферах людської діяльності, таких як наука, економіка, виробництво й безліч інших сфер. Прогнозування є важливим елементом організації управління як окремими господарюючими суб'єктами, так і економікою в цілому.

Розвиток методів прогнозування безпосередньо пов'язаний із розвитком інформаційних технологій, зокрема, із зростанням обсягів збережених даних і ускладненням методів і алгоритмів прогнозування, реалізованих в інструментах Data Mining.

Задача прогнозування, мабуть, може вважатися однією з найбільш складних задач Data Mining, вона вимагає ретельного дослідження вихідного набору даних і методів, що задовольняють аналізу.

Прогнозування (від грецького Prognosis), у широкому розумінні цього слова, визначається як випереджаюче відображення майбутнього**.**

Метою прогнозування є передбачення майбутніх подій.

Прогнозування (forecasting) є однією з задач Data Mining і одночасно одним із ключових моментів при прийнятті рішень.

Прогностика (prognostics) – теорія й практика прогнозування.

Прогнозування спрямоване на визначення тенденцій динаміки конкретного об'єкта або події на основі ретроспективних даних, тобто аналізу його стану колись і тепер. Отже, розв'язок задачі прогнозування вимагає деякої навчальної вибірки даних.

Прогнозування – установлення функціональної залежності між залежними й незалежними змінними.

Прогнозування є розповсюдженим і затребуваним завданням у багатьох сферах людської діяльності. У результаті прогнозування зменшується ризик прийняття невірних, необґрунтованих або суб'єктивних рішень.

Приклади задач прогнозування: прогноз руху грошових коштів, прогнозування врожайності агрокультури, прогнозування фінансової стабільності підприємства.

Крім економічної й фінансової сфери, задачі прогнозування постають в медицині, фармакології; популярним зараз стає політичне прогнозування.

Загалом розв'язок задачі прогнозування зводиться до розв'язку таких підзадач:

- вибір моделі прогнозування;
- аналіз адекватності й точності побудованого прогнозу.

Порівняння задач прогнозування і класифікації.

Прогнозування подібне із задачею класифікації.

Багато методів Data Mining використовуються для розв'язку задач класифікації і прогнозування. Це, наприклад, лінійна регресія, нейронні мережі, дерева рішень (які, іноді, так і називають – дерева прогнозування й класифікації).

Задачі класифікації й прогнозування мають подібності й відмінності.

Так у чому ж подібність задач прогнозування й класифікації? При розв'язку обох задач використовується двоетапний процес побудови моделі на основі навчального набору та її використання для прогнозування невідомих значень залежної змінної.

Відмінність задач класифікації й прогнозування полягає в тому, що в першій задачі передбачається клас залежної змінної, а в другій – числові значення залежної змінної, пропущені або невідомі (які відносяться до майбутнього).

Повертаючись до прикладу про туристичне агентство, розглянутого у попередній лекції, ми можемо сказати, що визначення класу клієнта є розв'язком задачі класифікації, а прогнозування доходу, який принесе цей клієнт наступного року, буде розв'язком задачі прогнозування.

2. Прогнозування і часові ряди

Прогнозування і часові ряди. Основою для прогнозування служить історична інформація, що зберігається в базі даних у вигляді **часових рядів**.

Існує поняття Data Mining **часових рядів** (Time-Series Data Mining).

На основі ретроспективної інформації у вигляді часових рядів можливий розв'язок різних задач Data Mining.

На рис. 7.1 представлені результати опитування відносно Data Mining часових рядів. Як бачимо, найбільший відсоток (23%) серед розв'язуваних задач займає прогнозування. Далі йдуть класифікація і кластеризація (по 14%), сегментація й виявлення аномалій (по 9%), виявлення правил (8%). На інші задачі доводиться менш, ніж по 6%.

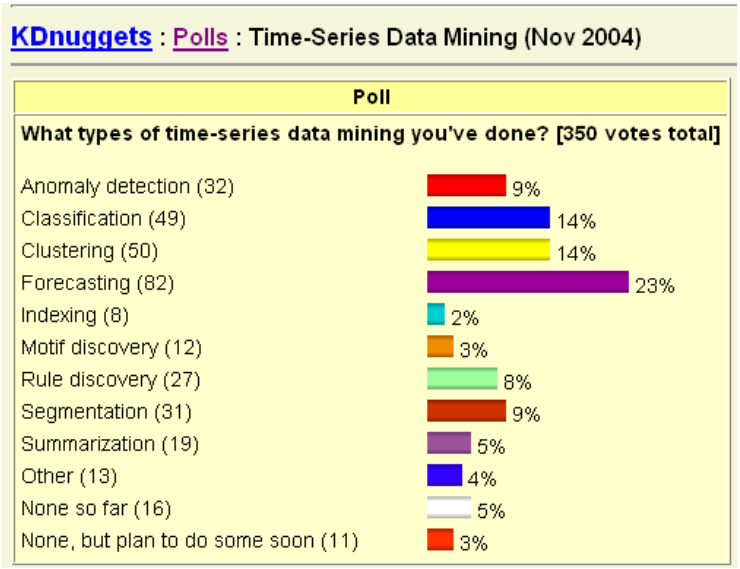

Рисунок 7.1 – Data Mining часових рядів

Однак, щоб зосередитися на понятті прогнозування, ми будемо розглядати часові ряди лише в рамках розв'язку задачі прогнозування.

Приведемо **дві принципові відмінності часового ряду від простої послідовності спостережень:**

- члени часового ряду, на відміну від елементів випадкової вибірки, не є статистично незалежними.
- члени часового ряду не є однаково розподіленими.

Часовий ряд – послідовність спостережуваних значень будь-якої ознаки, упорядкованих у невипадкові моменти часу.

Відмінністю аналізу часових рядів від аналізу випадкових вибірок є припущення про рівні проміжки часу між спостереженнями та їх хронологічний порядок. Прив'язка спостережень до часу відіграє тут ключову роль, тоді як при аналізі випадкової вибірки вона не має ніякого значення.

Типовий приклад часового ряду – дані біржових торгів.

Інформація, накопичена в різноманітних базах даних підприємства, є часовими рядами, якщо вона розташована в хронологічному порядку і зроблена в послідовні моменти часу.

Аналіз часового ряду здійснюється з метою:

- визначення природи ряду;
- прогнозування майбутніх значень ряду.

У процесі визначення структури й закономірностей часового ряду передбачається виявлення: шумів і викидів, тренду, сезонного компонента, циклічного компонента. Визначення природи часового ряду може бути використане як своєрідна «розвідка» даних. Знання аналітика про наявність сезонного компонента необхідне, наприклад, для визначення кількості записів вибірки, яка повинна брати участь у побудові прогнозу.

Аналіз часового ряду ускладнюють **шуми й викиди** (будуть докладно розглянуті в наступних темах курсу). Існують різні методи визначення й фільтрації викидів, що дають можливість виключити їх з метою більш якісного Data Mining.

3.Тренд, сезонність і цикл

Основними складовими часового ряду є тренд і сезонний компонент.

Тренд є систематичним компонентом часового ряду, який може змінюватися в часі.

Трендом називають невипадкову функцію, яка формується під дією загальних або довгочасних тенденцій, що впливають на часовий ряд.

Прикладом тенденції може виступати, наприклад, фактор зростання досліджуваного ринку.

Автоматичного способу виявлення трендів у часових рядах не існує. Але якщо часовий ряд включає монотонний тренд (тобто відзначене його стійке зростання або стійке спадання), аналізувати часовий ряд у більшості випадків неважко.

Існує велика різноманітність постановок задач прогнозування, які можна поділити на дві групи: прогнозування односерійних рядів і прогнозування мультисерійних, або взаємовпливаючих, рядів.

Група прогнозування односерійних рядів включає задачу побудови прогнозу однієї змінної за ретроспективними даними тільки цієї змінної, без врахування впливу інших змінних і факторів.

Група прогнозування мультисерійних, або взаємовпливаючих, рядів включає задачу аналізу, де необхідно враховувати взаємовпливаючі фактори на одну або декілька змінних.

Крім розподілу на класи по односерійності й багатосерійності, ряди також бувають сезонними й несезонними.

Останній розподіл має на увазі наявність або відсутність у часового ряду такої складової як сезонність, тобто включення **сезонного компонента**.

Сезонна складова часового ряду є періодично повторюваним компонентом часового ряду.

Властивість сезонності означає, що через приблизно рівні проміжки часу форма кривої, яка описує поведінку залежної змінної, повторює свої характерні обриси.

Властивість сезонності важлива при визначенні кількості ретроспективних даних, які будуть використовуватися для прогнозування.

Розглянемо простий приклад. На рис. 7.2 наведено фрагмент ряду, який ілюструє поведінку змінної «обсяги продажу товару X» за період, що становить один місяць. При вивченні кривої, наведеної на рисунку, аналітик не може зробити припущень щодо повторюваності форми кривої через рівні проміжки часу.

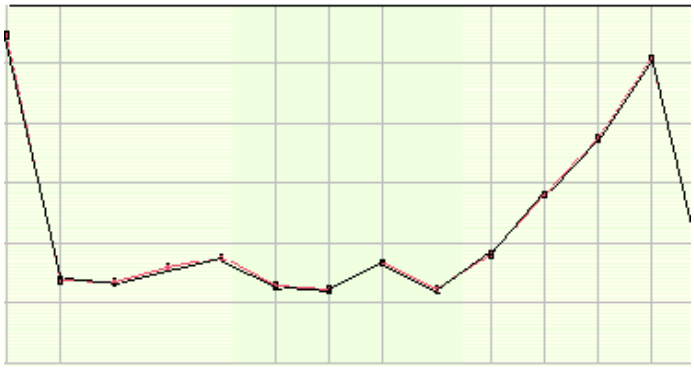

Рисунок 7.2 – Фрагмент часового ряду за сезонний період

Однак при розгляді більш тривалого ряду (за 12 місяців), зображеного на рис. 7.3, можна побачити явну наявність сезонного компонента. Отже, про сезонність продажів можна говорити тільки, коли розглядаються дані за кілька місяців.

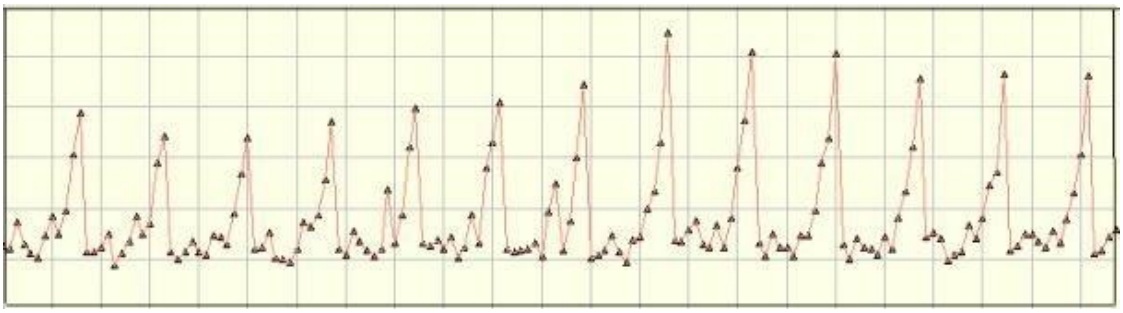

Рисунок 7.3 – Фрагмент часового ряду з 12-ти сезонних періодів

Отже, у процесі підготовки даних для прогнозування аналітикові слід визначити, чи має ряд, який він аналізує, властивість сезонності.

Визначення наявності компоненти сезонності необхідне для того, щоб вхідна інформація мала властивість репрезентативності.

Ряд можна вважати несезонним, якщо при розгляді його зовнішнього вигляду не можна зробити припущень про повторюваність форми кривої через рівні проміжки часу.

Іноді по зовнішньому вигляду кривої ряду не можна визначити, є він сезонним чи ні.

Існує поняття сезонного мультиряду. У ньому кожний ряд описує поведінку факторів, які впливають на залежну (цільову) змінну.

Приклад такого ряду – ряди продажів декількох товарів, що піддаються сезонним коливанням.

При зборі даних і виборі факторів для розв’язку задачі прогнозування в таких випадках слід ураховувати, що вплив обсягів продажів товарів один на одного тут набагато менше, ніж вплив фактору сезонності.

Важливо не плутати поняття сезонного компонента ряду й сезонів природи. Незважаючи на близькість їх звучання, ці поняття відрізняються. Так, наприклад, обсяги продажів морозива влітку набагато більше, ніж в інші сезони, однак це є тенденцією попиту на даний товар.

Дуже часто тренд і сезонність присутні в часовому ряді одночасно.

Приклад. Прибуток фірми зростає протягом декількох років (тобто в часовому ряді присутній тренд); ряд також містить сезонний компонент.

Відмінності циклічного компонента від сезонного:

1. Тривалість циклу, як правило, більше, ніж один сезонний період.
2. Цикли, на відміну від сезонних періодів, не мають певної тривалості.

При виконанні яких-небудь перетворень зрозуміти природу часового ряду значно простіше, такими перетвореннями можуть бути, наприклад, видалення тренда й згладжування ряду.

Перед початком прогнозування необхідно відповісти на такі питання:

1. Що потрібно прогнозувати?
2. У яких часових елементах (параметрах)?
3. З якою точністю прогнозу?

При відповіді на перше питання, ми визначаємо змінні, які будуть прогнозуватися. Це може бути, наприклад, рівень проведення конкретного виду продукції в наступному кварталі, прогноз суми продажу цієї продукції і т.д.

При виборі змінних слід ураховувати доступність ретроспективних даних, переваги осіб, що ухвалюють рішення, остаточну вартість Data Mining.

Часто при розв'язку задач прогнозування виникає необхідність прогнозування не самої змінної, а зміни її значень.

Друге питання при розв'язку задачі прогнозування – визначення таких параметрів:

- періоду прогнозування;
- горизонту прогнозування;
- інтервалу прогнозування.

Період прогнозування – основна одиниця часу, на яку робиться прогноз.

Наприклад, ми прагнемо довідатися дохід компанії через місяць. Період прогнозування для цієї задачі – місяць.

Горизонт прогнозування – це число періодів у майбутньому, які покриває прогноз.

Якщо ми прагнемо дізнатися прогноз на 12 місяців уперед, із даними по кожному місяцю, то період прогнозування в цьому завданні – місяць, горизонт прогнозування – 12 місяців.

Інтервал прогнозування – частота, з якою робиться новий прогноз. Інтервал прогнозування може збігатися з періодом прогнозування.

Рекомендації з вибору параметрів прогнозування. При виборі параметрів необхідно враховувати, що горизонт прогнозування повинен бути не менше, ніж час, який необхідний для реалізації розв'язку, прийнятого на основі цього прогнозу. Тільки в цьому випадку прогнозування буде мати сенс.

Зі збільшенням горизонту прогнозування точність прогнозу, як правило, знижується, а зі зменшенням горизонту – підвищується.

Ми можемо поліпшити якість прогнозування, зменшуючи час, необхідний на реалізацію розв'язку, для якого реалізується прогноз, і, отже, зменшити при цьому горизонт і помилку прогнозування.

При виборі інтервалу прогнозування слід вибирати між двома ризиками: вчасно не визначити зміни в аналізованому процесі й високою вартістю прогнозу. При тривалому інтервалі прогнозування виникає ризик не ідентифікувати зміни, які відбуваються в процесі, при короткому – зростають витрати на прогнозування.

При виборі інтервалу необхідно також ураховувати стабільність аналізованого процесу й вартість проведення прогнозу.

Точність прогнозу, що необхідна для розв'язку конкретної задачі, дуже впливає на прогнозуючу систему. Помилка прогнозу залежить від використовуваної системи прогнозу.

Чим більше ресурсів має така система, тим більше шансів одержати більш точний прогноз. Однак прогнозування не може повністю усунути ризики при прийнятті рішень. Тому завжди враховується можлива помилка прогнозування.

4. Види помилок та прогнозів

Точність прогнозу характеризується помилкою прогнозу.

Найпоширеніші види помилок:

- **Середня помилка (СП).** Вона обчислюється простим усередненням помилок на кожному кроці. Недолік цього виду помилки – позитивні й негативні помилки анулюють одна одну.
- **Середня абсолютна помилка (САП)**. Вона розраховується як середнє абсолютних помилок. Якщо вона дорівнює нулю, то ми маємо досконалий прогноз. У порівнянні із середньою квадратичною помилкою, цей захід «не надає занадто великого значення» викидам.
- **Сума квадратів помилок (SSE)**, середньоквадратична помилка. Вона обчислюється як сума (або середнє) квадратів помилок. Це найбільше часто використовувана оцінка точності прогнозу.
- **Відносна помилка (ВП).** Попередні міри використовували дійсні значення помилок. Відносна помилка виражає якість припасування в термінах відносних помилок.

Види прогнозів. Прогноз може бути короткостроковим, середньостроковим і довгостроковим.

Короткостроковий прогноз являє собою прогноз на кілька кроків уперед, тобто здійснюється побудова прогнозу не більше ніж на 3% від обсягу спостережень або на 1-3 кроку вперед.

Середньостроковий прогноз – це прогноз на 3-5% від обсягу спостережень, але не більш 7-12 кроків уперед; також під цим типом прогнозу розуміють прогноз на один або половину сезонного циклу. Для

побудови короткострокових і середньострокових прогнозів цілком підходять статистичні методи.

Довгостроковий прогноз – це прогноз більш ніж на 5% від обсягу спостережень.

При побудові даного типу прогнозів статистичні методи практично не використовуються, крім випадків дуже «гарних» рядів, для яких прогноз можна просто «намалювати».

Дотепер ми розглядали аспекти прогнозування, так чи інакше пов'язані із процесом ухвалення рішення. Існують і інші фактори, які необхідно враховувати при прогнозуванні.

Задача 1. Відомо, що аналізований процес відносно стабільний у часі, зміни відбуваються повільно, процес не залежить від зовнішніх факторів.

Задача 2. Аналізований процес нестабільний і дуже сильно залежить від зовнішніх факторів.

Розв'язок першої задачі повинен бути зосереджений на використанні великої кількості ретроспективних даних. **При розв'язку другої задачі** особливу увагу слід звернути на оцінки фахівця в предметній області, експерта, щоб мати можливість відобразити в прогнозуючій моделі всі необхідні зовнішні фактори, а також приділити час для збору даних по цих факторах (збір зовнішніх даних часто набагато складніший збору внутрішніх даних інформаційної системи). Доступність даних, на основі яких буде здійснюватися прогнозування, – важливий фактор побудови прогнозної моделі. Для можливості виконання якісного прогнозу дані повинні бути представницькими, точними й достовірними.

Методи прогнозування. Серед розповсюджених методів Data Mining, використовуваних для прогнозування, відзначимо **нейронні мережі** й **лінійну регресію**.

Вибір методу прогнозування залежить від багатьох факторів, у тому числі від параметрів прогнозування. Вибір методу слід провадити з обліком усіх специфічних особливостей набору ретроспективних даних і цілей, заради яких він будується.

Програмне забезпечення Data Mining, використовуване для прогнозування, повинно забезпечувати користувача точним і достовірним прогнозом. Однак одержання такого прогнозу залежить не тільки від програмного забезпечення й методів, закладених у його основу, але також і від інших факторів, серед яких повнота й вірогідність вихідних даних, своєчасність і оперативність їх поповнення, кваліфікація користувача.

5. Візуалізація інструментів Data Mining

Візуалізація – це спосіб, який дозволяє побачити кінцевий результат обчислень, організувати керування обчислювальним процесом і навіть повернутися назад до вихідних даних, щоб визначити найбільш раціональний напрямок подальшого руху.

У результаті використання візуалізації створюється графічний образ

даних. Застосування візуалізації допомагає в процесі аналізу даних побачити аномалії, структури, тренди. При розгляді задачі прогнозування ми використовували графічне представлення часового ряду й побачили, що в ньому присутній сезонний компонент. У попередній лекції розглянуто задачі **класифікації й кластеризації**, і для ілюстрації розподілу об'єктів у двомірному просторі також використана візуалізацію.

Можна говорити про те, що застосування візуалізації є більш економічним: лінія тренду або скупчення точок на діаграмі розсіювання дозволяє аналітикові набагато швидше визначити закономірності й прийти до потрібного розв'язку. Отже, тут йдеться про використання в Data Mining не символів, а образів.

Головна перевага візуалізації – практично повна відсутність необхідності в спеціальній підготовці користувача. За допомогою візуалізації ознайомитися з інформацією дуже легко, досить лише на неї подивитися.

Хоча найпростіші види візуалізації з'явилися досить давно, її використання зараз тільки набирає популярність. Візуалізація не спрямована винятково на вдосконалення техніки аналізу – за словами Скотта Лейбса, у деяких випадках візуалізація може навіть замінити її.

Візуалізація даних може бути представлена у вигляді: графіків, схем, гістограм, діаграм тощо.

Коротко роль візуалізації можна описати такими її можливостями:

- підтримка інтерактивного й погодженого дослідження;
- допомога в показі результатів;
- використання очей (зору), щоб створювати зорові образи й осмислювати їх.

Погана візуалізація. Результати візуалізації іноді можуть вводити користувача в оману. Приведемо простий приклад поганої візуалізації. Допустимо, ми маємо базу «Прибуток компанії А» за період з 2000 по 2017 рік, вона представлена в таблиці 7.1.

Таблиця 7.1 − Прибуток компанії А

Рік	Прибуток
2000	1100
2001	1101
2002	1104
2003	1105
2004	1106
2017	1007

Побудуємо гістограмму в Excel за цими даними. Гістограмма являє собою візуальне зображення розподілу даних.

Ця інформація відображається за допомогою серії прямокутників або смуг однакової ширини, висота яких указує кількість даних у кожному класі.

Використовуючи всі значення побудови графіка, прийняті за замовчуванням, одержуємо гістограмму, наведену на рис. 7.4.

Рисунок 7.4 – Гістограма прибуткукомпанії
(мінімальне значення вісі *Y* дорівнює 940)

Цей рисунок демонструє значне зростання прибутку компанії А за період з 2000 по 2017 року. Однак, якщо ми звернули увагу на вісь *Y*, що показує величину прибутку, то побачимо, що ця вісь перетинає вісь *X* у значенні, рівному 1096. Фактично, вісь *Y* зі значеннями від 1096 до 1108 вводить користувача в оману. Змінивши значення параметрів, відповідальних за формат вісі *Y*, одержуємо графік, наведений на рис. 7.5.

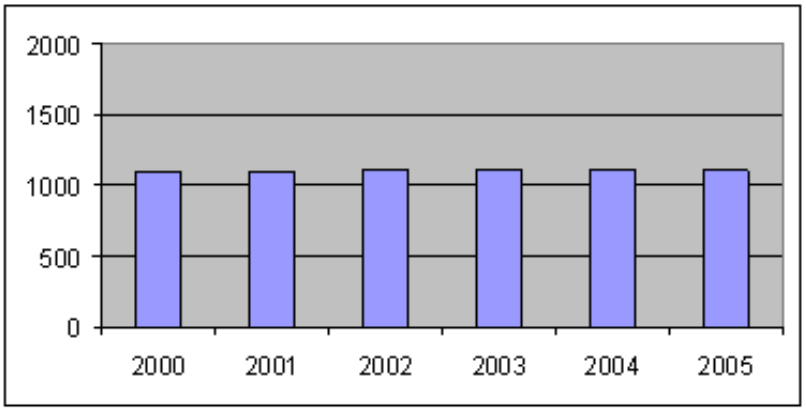

Рисунок 7.5 – Гістограма прибутку компанії
(мінімальне значення вісі Y дорівнює 0)

Вісь *Y* зі значеннями від 0 до 2000 дає користувачеві правильну інформацію про незначну зміну прибутків компанії.

Якщо мова йде про велику розмірність і складності вихідних даних, кошти візуалізації забезпечують їхнє різке зменшення, конденсуючи, можливо, мільйони записів даних у прості, легкі для розуміння й маніпулювання показники. Такі показники називають візуальним або графічним способом показу інформації. Візуалізацію можна вважати

ключовим фактором у дослідженні даних, отриманих за допомогою інструментів Data Mining. У таких випадках говорять про візуальний Data Mining.

Зі зростанням кількості даних, що накопичуються, навіть при використанні як завгодно потужних і різносторонніх алгоритмів Data Mining, стає усе складніше «переварювати» і інтерпретувати отримані результати. А, як відомо, одне з положень Data Mining – пошук практично корисних закономірностей. Закономірність може стати практично корисною, тільки якщо її можна осмислити й зрозуміти.

У 1987 році з ініціативи ACM SIGGRAPH IEEE Computer Society Technical Committee of Computer Graphics, у зв'язку з необхідністю використання нових методів, засобів і технологій даних, були сформульовані відповідні завдання напрямку візуалізації.

До способів візуального або графічного представлення даних відносять графіки, діаграми, таблиці, звіти, списки, структурні схеми, карти тощо.

Візуалізація традиційно розглядалася як допоміжний засіб при аналізі даних, однак зараз усе більше досліджень говорить про її самостійну роль.

Традиційні методи візуалізації можуть знаходити таке застосування:

- представляти користувачеві інформацію в наочному вигляді;
- компактно описувати закономірності, властиві вихідному набору даних;
- знижувати розмірність або стискати інформацію;
- відновлювати пробіли в наборі даних;
- знаходити шуми й викиди в наборі даних.

Візуалізація інструментів Data Mining. Кожний з алгоритмів Data Mining використовує певний підхід до візуалізації. У попередніх лекціях ми розглянули ряд методів Data Mining. У ході використання кожного з методів, а точніше, його програмної реалізації, ми одержували якісь візуалізатори, за допомогою яких нам вдавалося інтерпретувати результати, отримані в результаті роботи відповідних методів і алгоритмів.

Для дерев рішень це візуалізатор дерева рішень, список правил, таблиця спряженості.

Для нейронних мереж залежно від інструмента це може бути топологія мережі, графік зміни величини помилки, що демонструє процес навчання.

Для карт Кохонена: карти входів, виходів, інші специфічні карти.

Для лінійної регресії як візуалізатор виступає лінія регресії.

Для кластеризації: дендрограми, діаграми розсіювання.

Діаграми й графіки розсіювання часто використовуються для оцінки якості роботи того або іншого методу.

Усі ці способи візуального представлення або відображення даних можуть виконувати одну з функцій:

- є ілюстрацією побудови моделі (наприклад, представлення структури (графа) нейронної мережі);
- допомагають інтерпретувати отриманий результат;

- є засобом оцінки якості побудованої моделі;
- поєднують перераховані вище функції (дерево розв'язків, дендрограма).

Існує багато різних способів представлення моделей, але графічне їх представлення дає користувачеві максимальну «цінність».

Користувач, у більшості випадків, не є фахівцем у моделюванні, найчастіше він експерт у своїй предметній області. Тому модель Data Mining повинна бути представлена на найбільш природній для нього мові або, хоча б, містити мінімальну кількість різних математичних і технічних елементів.

Отже, доступність є однією з основних характеристик моделі Data Mining. Незважаючи на це, існує й такий розповсюджений і найбільш простий спосіб показу моделі, як «чорний ящик». У цьому випадку користувач не розуміє поведінки тієї моделі, якою користується. Однак, незважаючи на нерозуміння, він одержує результат – виявлені закономірності. Класичним прикладом такої моделі є модель нейронної мережі.

Інший спосіб представлення моделі – представлення її в інтуїтивному, зрозумілому виді. У цьому випадку користувач дійсно може розуміти те, що відбувається «усередині» моделі. Таким чином можна забезпечити його особисту участь у процесі.

Такі моделі забезпечують користувачеві можливість обговорювати її логіку з колегами, клієнтами й іншими користувачами, або пояснювати її.

Розуміння моделі веде до розуміння її змісту. У результаті розуміння зростає довіра до моделі. Класичним прикладом є дерево рішень. Побудоване дерево рішень дійсно поліпшує розуміння моделі, тобто використовуваного інструмента Data Mining.

Крім розуміння, такі моделі забезпечують користувача можливістю взаємодіяти з моделлю, задавати їй питання й одержувати відповіді. Прикладом такої взаємодії є засіб «що, якщо». За допомогою діалогу «система-користувач» користувач може одержати розуміння моделі.

Тепер перейдемо до функцій, які допомагають інтерпретувати й оцінити результати побудови Data Mining моделей. Це всілякі графіки, діаграми, таблиці, списки тощо.

Прикладами засобів візуалізації, за допомогою яких можна оцінити якість моделі, є діаграма розсіювання, таблиця спряженості, графік зміни величини помилки.

Діаграма розсіювання являє собою графік відхилення значень, прогнозованих за допомогою моделі, від реальних. Ці діаграми використовують для безперервних величин. Візуальна оцінка якості побудованої моделі можлива тільки по закінченню процесу побудови моделі.

Таблиця спряженості використовується для оцінки результатів класифікації. Такі таблиці застосовуються для різних методів класифікації. Оцінка якості побудованої моделі тут також можлива тільки по закінченню процесу побудови моделі.

Графік зміни величини помилки. Графік демонструє зміну величини помилки в процесі роботи моделі. Наприклад, у процесі роботи нейронних мереж користувач може спостерігати за зміною помилки на навчальній й тестовій множинах і зупинити навчання для недопущення «перенавчання» мережі. Тут оцінка якості моделі і його зміни може оцінюватися безпосередньо в процесі побудови моделі.

Прикладами засобів візуалізації, які допомагають інтерпретувати результат, є: лінія тренду в лінійній регресії, карти Кохонена, діаграма розсіювання в кластерному аналізі.

6. Методи візуалізації

Методи візуалізації, залежно від кількості використовуваних вимірів, прийнято класифікувати на дві групи:

- представлення даних в одному, двох і трьох вимірах;
- представлення даних у чотирьох і більше вимірах.

Представлення даних в одному, двох і трьох вимірах. До цієї групи методів ставляться добре відомі способи відображення інформації, які доступні для сприйняття людською уявою. Практично будь-який сучасний інструмент Data Mining включає способи візуального представлення із цієї групи.

Відповідно до кількості вимірів представлення це можуть бути такі способи:

- одномірне (univariate) або 1-D;
- двовимірне (bivariate) або 2-D;
- тривимірне, проекційне (projection) або 3-D.

Слід відзначити, що найбільше природно людське око сприймає двомірні представлення інформації.

При використанні двох- і тривимірного представлення інформації користувач має можливість побачити закономірності набору даних:

- його кластерну структуру й розподіл об’єктів на класи (наприклад, на діаграмі розсіювання);
- топологічні особливості;
- наявність трендів;
- інформацію про взаємне розташування даних;
- існування інших залежностей, властивих досліджуваному набору даних.

Якщо набір даних має більше трьох вимірів, то можливі такі варіанти:

- використання багатомірних методів представлення інформації (вони розглянуті нижче);
- зниження розмірності до одно-, двох- або тривимірного представлення. Існують різні способи зниження розмірності, один з них – факторний аналіз – був розглянутий в одній з попередніх лекцій. Для зниження розмірності й одночасного візуального представлення інформації на двовимірних картах використовуються карти, що самоорганізуються.

Представлення даних в чотирьох і більше вимірах. Представлення інформації в чотиривимірному й більш вимірах недоступні для людського сприйняття. Однак розроблені спеціальні методи для можливості відображення й сприйняття людиною такої інформації.

Найбільш відомі способи багатомірного представлення інформації:

- паралельні координати;
- «особи Чернова»;
- пелюсткові діаграми.

Паралельні координати. У паралельних координатах змінні кодуються по горизонталі, вертикальна лінія визначає значення змінної. Приклад набору даних, представленого в декартових координатах і паралельних координатах, подано на рис. 7.6. Цей метод представлення багатомірних даних був винайдений Альфредом Інселбергом (Alfred Inselberg) в 1985 році.

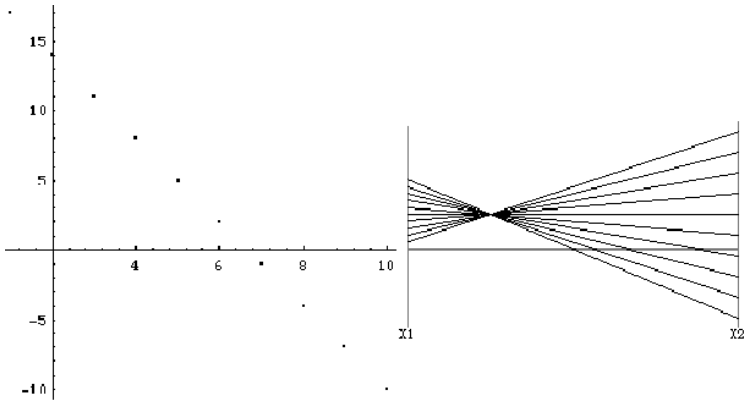

Рисунок 7.6 – Набір даних у декартових координатах
і в паралельних координатах

«Особи Чернова». Основна ідея представлення інформації в «особах Чернова» полягає в кодуванні значень різних змінних у характеристиках або рисах людської особи. Приклад такої «особи» наведено на рис 7.7.

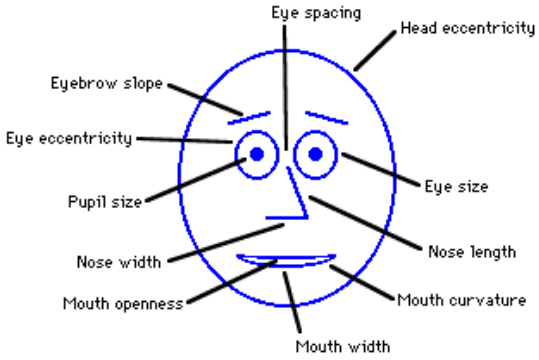

Рисунок 7.7 – «Особа Чернова»

Для кожного спостереження зображується окрема «особа». На кожній «особі» відносні значення змінних представлені як форми й розміри окремих

рис особи (наприклад, довжина й ширина носа, розмір очей, розмір зіниці, кут між бровами).

Аналіз інформації за допомогою такого способу відображення заснований на здатності людини інтуїтивно знаходити подібності й відмінності в рисах особи.

На рис. 7.8 представлений набір даних, кожний запис якого виражений у вигляді «Особи Чернова»**.**

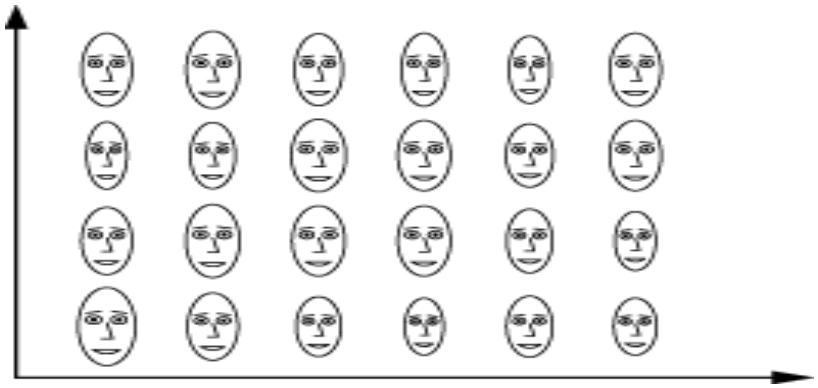

Рисунок 7.8 – Приклад багатомірного зображення даних за допомогою «осіб Чернова»

Перед використанням методів візуалізації необхідно:

- проаналізувати, чи варто зображувати всі дані або ж лише якусь їхню частину;
- вибрати розміри, пропорції й масштаб зображення;
- вибрати метод, який може найбільше яскаво відобразити закономірності, властиві набору даних.

Багато сучасних засобів аналізу даних дозволяють будувати сотні типів різних графіків і діаграм. Тому вибір методу візуалізації, якщо він самостійно здійснюється користувачем, не такий простий і легкий, як може здатися на перший погляд. Наявність великої кількості засобів візуалізації, представлених в інструменті, який застосовує користувач, може навіть викликати розгубленість.

Ту саму інформацію можна представити за допомогою різних засобів. Для того, щоб засіб візуалізації міг виконувати своє основне призначення – представляти інформацію в простому й доступному для людського сприйняття вигляді – необхідно дотримуватися законів відповідності обраного розв’язку змісту відображуваної інформації і її функціональному призначенню. Іншими словами, потрібно зробити так, щоб при погляді на візуальне представлення інформації можна було відразу виявити закономірності у вихідних даних і приймати на їхній основі рішення.

Серед двомірних і тривимірних засобів найбільше широко відомі лінійні графіки, лінійні, стовпчикові, кругові секторні й векторні діаграми.

Приведемо рекомендації з використання цих найбільш простих і популярних засобів візуалізації.

За допомогою лінійного графіка можна відобразити тенденцію, передати зміни якої-небудь ознаки в часі. Для порівняння декількох рядів чисел такі графіки наносяться на ті самі осі координат.

Гістограму застосовують для порівняння значень протягом деякого періоду або ж співвідношення величин.

Кругові діаграми використовують, якщо необхідно відобразити співвідношення частин і цілого, тобто для аналізу складу або структури явищ. Складові частини цілого зображуються секторами круга. Сектори рекомендують розміщати за їхньою величиною: угорі – найбільший, інші – по рухові годинної стрілки в порядку зменшення їх величини. Кругові діаграми також застосовують для відображення результатів факторного аналізу, якщо дії всіх факторів є односпрямованими. При цьому кожний фактор відображається у вигляді одного із секторів кола.

Вибір того або іншого засобу візуалізації залежить від поставленого завдання (наприклад, потрібно визначити структуру даних або ж динаміку процесу) і від характеру набору даних.

Якість візуалізації. Сучасні аналітичні засоби, у тому числі й Data Mining, немислимі без якісної візуалізації. У результаті використання засобів візуалізації повинні бути отримані наочні й виразні, ясні й прості зображення, за рахунок використання різноманітних засобів: кольору, контрасту, границь, пропорцій, масштабу тощо.

У зв'язку із зростанням вимог до засобів візуалізації, а також необхідності порівняння їх між собою, в останні роки був сформований ряд принципів якісного візуального представлення інформації.

Принципи Тафта (Tufte's Principles) про графічне представлення даних високої якості говорить:

- надавайте користувачеві найбільшу кількість ідей, у найкоротший час, з найменшою кількістю чорнила на найменшому просторі;
- говоріть правду про дані.

7. Принципи компонування візуальних засобів

Основні принципи компонування візуальних засобів представлення інформації:

1. Принцип лаконічності.
2. Принцип узагальнення й уніфікації.
3. Принцип акценту на основних значимих елементах.
4. Принцип автономності.
5. Принцип структурності.
6. Принцип стадійності.
7. Принцип використання звичних асоціацій і стереотипів.

Принцип лаконічності говорить про те, що засіб візуалізації повинен містити лише ті елементи, які необхідні для повідомлення користувачеві істотної інформації, точного розуміння її значення або прийняття (з

імовірністю не нижче допустимої величини) відповідного оптимального розв’язку.

Крім позначених вище принципів, засіб візуалізації повинний мати високу надійність і швидкість, яка влаштує користувача, що приймає на основі цієї інформації рішення.

Представлення просторових характеристик. Окремим напрямком візуалізації є наочне представлення просторових характеристик об’єктів. У більшості випадків такі засоби виділяють на карті окремі регіони й позначають їхніми різними кольорами залежно від значення аналізованого показника.

На рис. 7.9 наведений приклад такої візуалізації в середовищі Mineset, що є, у цьому випадку, інструментом візуального Data Mining. Карта представлена у вигляді графічного інтерфейсу, що відображає дані у вигляді тривимірного ландшафту довільно визначених і позиціонованих форм (стовпчастих діаграм, кожна з індивідуальними висотою й кольором). Такий спосіб дозволяє наочно показувати кількісні й реляційні характеристики просторово-орієнтованих даних і швидко ідентифікувати в них тренди.

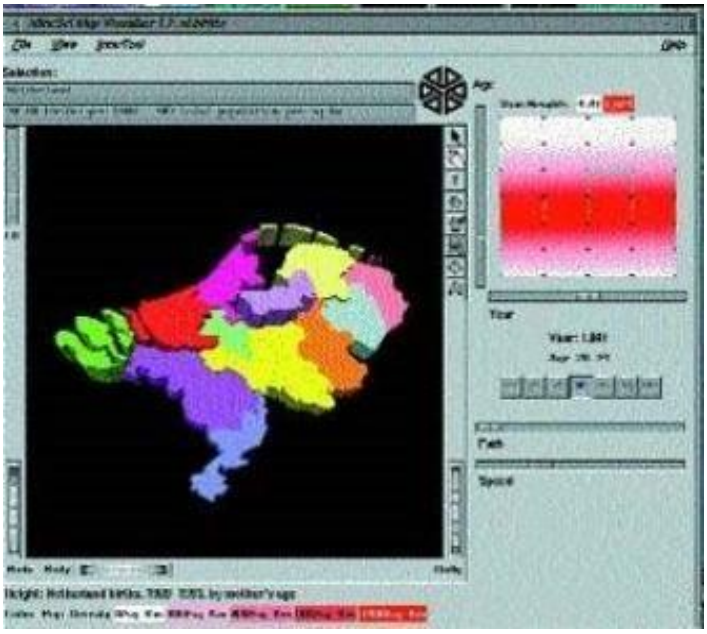

Рисунок 7.9 – Mineset. Ландшафтний візуалізатор

8. Основні тенденції в області візуалізації

Як ми вже відзначали, за допомогою засобів візуалізації підтримуються важливі завдання бізнесу, серед яких – процес прийняття рішень. У зв’язку із цим виникає необхідність переходу засобів візуалізації на більш якісний рівень, який характеризується появою абсолютно нових засобів візуалізації й поглядів на їх функції, а також розвитком ряду тенденцій у цій області.

Серед основних тенденцій в області візуалізації Філіп Рассом (Philip Russom) виділяє:

1. Розробка складних видів діаграм.
2. Підвищення рівня взаємодії з візуалізацією користувача.

3. Збільшення розмірів і складності структур даних, що представляються візуалізацією.

Розробка складних видів діаграм. Більшість візуалізацій даних побудовані на основі діаграм стандартного типу (секторні діаграми, графіки розсіювання тощо). Ці способи є одночасно найбільш елементарними й розповсюдженими. В останні роки перелік видів діаграм, підтримуваних інструментальними засобами візуалізації, суттєво розширився. Оскільки потреби користувачів досить різноманітні, інструменти візуалізації підтримують всілякі типи діаграм. Наприклад, відомо, що бізнес-користувачі віддають перевагу секторним діаграмам і гістограмам, тоді як вчених більше влаштовують візуалізації у вигляді графіків розсіювання й діаграм констеляції. Користувачі, що працюють із геопросторовими даними, більш зацікавлені в картах та інших тривимірних представленнях даних.

Електронні інструментальні панелі, своєю чергою, більш популярні серед керівників, що використовують бізнес-аналітичні технології для контролю над показниками роботи компанії. Такі користувачі потребують наочної візуалізації у вигляді «спідометрів», «термометрів» і «світлофорів».

Засоби створення діаграм і презентаційної графіки призначені головним чином для візуалізації даних. Однак можливості такої візуалізації звичайно вбудовані й у безліч різних інших програм і систем – в інструменти репортинга й OLAP, кошту для Text Mining і Data Mining, а також в Crm-Додатки й додатка для керування бізнесом. Для створення вбудованої візуалізації багато постачальників реалізують візуалізаційну функціональність у вигляді компонентів, що вбудовуються в різні інструменти, додатки, програми й web-сторінки (у тому числі інструментальні панелі й персоналізовані сторінки порталів).

Підвищення рівня взаємодії користувача з візуалізацією. Ще зовсім недавно більша частина коштів візуалізації являла собою статичні діаграми, призначені винятково для перегляду. Зараз широко використовуються динамічні діаграми, уже самі по собі, що є користувацьким інтерфейсом, у якому користувач може прямо й інтерактивно маніпулювати візуалізацією, підбираючи нову виставу інформації.

Наприклад, базова взаємодія дозволяє користувачеві обертати діаграму або змінювати її тип у пошуках найбільш повної представлення даних. Крім того, користувач може міняти візуальні властивості – приміром, шрифти, кольори й рамки. У візуалізаціях складного типу (графіках розсіювання або діаграмах констеляції) користувач може вибирати інформаційні крапки за допомогою миші й переміщати їх, полегшуючи тим самим розуміння представлення даних.

Більш досконалі методи візуалізації даних часто містять у собі діаграму або будь-яку іншу візуалізацію як складений рівень. Користувач може глибшатися (drill down) у візуалізацію, досліджуючи подробиці узагальнених нею даних, або глибшатися в OLAP, Data Mining або інші складні технології.

Складна взаємодія дозволяє користувачеві змінювати візуалізацію для знаходження альтернативних інтерпретацій даних. Взаємодія з візуалізацією

має на увазі мінімальний за своєю складністю користувацький інтерфейс, у якому користувач може управляти виставою даних, просто «кликая» на елементи візуалізації, перетаскуючи й поміщаючи представлення об'єктів даних або вибираючи пункти меню. Інструменти OLAP або Data Mining перетворюють безпосередню взаємодія з візуалізацією в один з етапів ітераційного аналізу даних. Кошту Text Mining або керування документами надають такій безпосередній взаємодії характер навігаційного механізму, що допомагає користувачеві досліджувати бібліотеки документів.

Візуальний запит є найбільш сучасною формою складної взаємодії користувача з даними. У ньому користувач може, наприклад, бачити крайні інформаційні крапки графіка розсіювання, вибирати їхньою мишкою й одержувати нові візуалізації, що представляють саме ці крапки. Додаток візуалізації даних генерує відповідна мова запиту, управляє прийняттям запиту базою даних і візуально представляє результуючу безліч. Користувач може сфокусуватися на аналізі, не відволікаючись на складання запиту.

Збільшення розмірів і складності структур даних, що представляються візуалізацією. Елементарна секторна діаграма або гістограма візуалізує прості послідовності числових інформаційних крапок. Однак нові вдосконалені типи діаграм здатні візуалізувати тисячі таких крапок і навіть складні структури даних – наприклад, нейронні мережі.

Скажемо, кошту OLAP (а також інструменти генерації запитів і випуску звітів) уже давно підтримують діаграми для своїх он-лайнових звітів. Нові візуалізаційні програми обновляють контент за рахунок періодично повторюваного зчитування даних. Візуалізаційні програми відслідковування лінійних процесів (коливання фондового ринку, показники роботи комп'ютерних систем, сейсмограми, сітки корисності) потребують завантаження даних у режимі реального часу або близькому до нього режимі для зручності користувачів.

Користувачі інструментів Data Mining звичайно аналізують дуже великі набори чисельних даних. Традиційні типи діаграм для бізнесу (секторні діаграми й гістограми) погано справляються з показом тисяч інформаційних точок. Тому інструменти Data Mining майже завжди підтримують якусь форму візуалізації даних, здатну відображати структури й закономірності досліджуваних наборів даних, відповідно до тих аналітичних підходів, які використовується в інструменті.

Крім того, що візуалізація підтримує обробку структурованих даних, вона також є ключовим засобом представлення схем так званих неструктурованих даних, наприклад текстових документів.

Text Mining. Зокрема, засоби Text Mining можуть здійснювати парсинг більших пакетів документів і формувати предметні покажчики понять і тем, освітлених у цих документах. Коли предметні покажчики створені за допомогою нейронної мережевої технології, користувачеві непросто продемонструвати їх без деякої форми візуалізації даних.

Питання для самоконтролю

1. Назвіть мету та визначення поняттю прогнозування.
2. Дайте визначення поняттю часовий ряд.
3. Які існують помилки прогнозу?
4. Які існують види прогнозів? Чим вони відрізняються?
5. Дайте визначення поняттю візуалізації. Для чого вона використовується?
6. Які існують способи багатомірного представлення інформації?