УДК 004
DOI 10.31891/2307-5732-2020-285-3-10

В. І. ТОМЕНКО
Черкаський інститут пожежної безпеки імені Героїв Чорнобиля Національного університету цивільного захисту України
Н. В. САЧАНЮК-КАВЕЦЬКА
Вінницький національний технічний університет
О. О. ДМИТРІЄНКО
Полтавський національний педагогічний університет імені В.Г. Короленка

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Інтелектуальний аналіз даних, в інформатиці - процес виявлення цікавих та корисних закономірностей та взаємозв'язків у великих обсягах даних. Поле поєднує інструменти статистики та штучного інтелекту (*такі як нейронні мережі та машинне навчання*) *та управління базами даних для аналізу великих цифрових колекцій, відомих як набори даних. Видобуток даних широко використовується у бізнесі* (*страхування, банківська справа, роздрібна торгівля*), *наукових дослідженнях* (*астрономія, медицина*) *та державній безпеці* (*виявлення злочинців та терористів*). *Поширення численних великих, а іноді і пов'язаних між собою державних та приватних баз даних призвело до постанов, що гарантують точність та захист окремих записів від несанкціонованого перегляду та підробки. Більшість типів аналізу даних спрямовані на встановлення загальних знань про групу, а не на знання про конкретних осіб - супермаркет менше турбується продажем ще однієї речі одній людині, ніж продажем багатьох предметів багатьом людям, хоча аналіз шаблонів також може бути використаний для того, щоб розпізнати аномальну поведінку особистості, таку як шахрайство чи інша злочинна діяльність. Одним із підходів до підвищення надійності є спочатку групування осіб, які мають подібні схеми закупівель, оскільки групові моделі менш чутливі до незначних аномалій. Наприклад, група "частих ділових мандрівників", швидше за все, матиме шаблон, який включає безпрецедентні покупки в різних місцях, але учасники цієї групи можуть бути позначені для інших транзакцій, таких як покупки в каталозі, які не відповідають профілю цієї групи*

Ключові слова: комп'ютерні програми, аналіз баз даних, інтелектуальний аналіз, електротехніка, математична модель, аналіз даних, контроль доступу.

V. TOMENKO
Cherkasy Institute of Fire Safety named after the Heroes of Chernobyl of the National University of Civil Defense of Ukraine
N. SACHANIUK-KAVETS'KA
Vinnytsia National Technical University
O. DMYTRIIENKO
Poltava V.G. Korolenko National Pedagogical University

DATA MINING

As computer storage increased during the 1980s, many companies began to store more transaction data. The resulting collection of records, often referred to as data warehouses, was too large to be analyzed using traditional statistical approaches. Several computer science conferences and seminars have been held to consider how the latest advances in artificial intelligence (*such as the discovery of expert systems, genetic algorithms, machine learning, and neural networks*) *can be adapted to discover knowledge (the best term in the computer science community). When studying consumer buying behavior, a typical pattern usually becomes apparent; purchases made outside of this pattern may be marked for further investigation or for denial of the transaction. However, the wide variety of normal behaviors makes this a difficult task; no difference between normal and fraudulent behavior works for everyone or constantly.*

Data mining, in computer science - the process of identifying interesting and useful patterns and relationships in large amounts of data. The field combines statistics and artificial intelligence tools (*such as neural networks and machine learning*) *and database management to analyze large digital collections known as datasets. Data mining is widely used in business* (*insurance, banking, retail*), *research (astronomy, medicine) and national security* (*detection of criminals and terrorists*). *The proliferation of numerous large and sometimes interconnected public and private databases has resulted in regulations guaranteeing the accuracy and protection of individual records from unauthorized inspection and forgery. Most types of data analysis focus on establishing general group knowledge rather than specific knowledge - a supermarket is less concerned with selling another item to one person than selling many items to many people, although pattern analysis can also be used to identify abnormalities. personal behavior, such as fraud or other criminal activity. One approach to improving reliability is to first group individuals who have similar procurement schemes, as group models are less sensitive to minor anomalies. For example, the "frequent business travelers" group is likely to have a template that includes unprecedented purchases in different locations, but members of this group may be marked for other transactions, such as catalog purchases that do not match the group's profile.*

Keywords: computer programs, database analysis, intelligent analysis, electrical engineering, mathematical model, data analysis, access control.

Методика. У міру збільшення обсягу комп'ютерного сховища протягом 1980-х років багато компаній почали зберігати більше даних про транзакції. Отримані в результаті колекції записів, які часто називають сховищами даних, були занадто великими, щоб їх можна було аналізувати за допомогою традиційних статистичних підходів. Було проведено декілька конференцій та семінарів з інформатики, щоб розглянути, як останні досягнення у галузі штучного інтелекту (такі як відкриття експертних систем, генетичних алгоритмів, машинного навчання та нейронних мереж) можуть бути адаптовані для відкриття знань (кращий термін у спільноті інформатики). Вивчаючи купівельну поведінку споживача, зазвичай стає очевидним типовий зразок; покупки, зроблені поза цим зразком, можуть бути позначені для подальшого розслідування або для відмови у трансакції. Однак широке розмаїття нормальної поведінки робить це складним завданням; жодна різниця між нормальною та шахрайською поведінкою не працює для всіх або постійно.

Аналіз останніх досліджень і публікацій (Literature review). У 1995 році до Першої міжнародної конференції з питань виявлення знань та видобутку даних, що відбулася в Монреалі, та запуску в 1997 році журналу Data Mining and Knowledge Discovery. Це був також період, коли було створено багато компаній з раннього видобутку даних і представлено продукцію. Одним із найперших успішних застосувань інтелектуального аналізу даних, який, можливо, поступається лише маркетинговим дослідженням, було виявлення шахрайства з кредитними картками.

Результати. Приблизно за останнє десятиліття сотня виробників комп'ютерного програмного забезпечення стрибала на обробку даних. Основні статистичні програмні пакети, такі як SAS, S-PLUS, SPSS та STATISTICA тощо, продаються як інструменти видобутку даних, а не як статистичні інструменти. Більше того, більшість майнерів даних та статистиків продовжують саркастично критикувати один одного. Це шкодить обом дисциплінам. На жаль, антистатистичне ставлення не дозволить видобутку даних досягти свого реального потенціалу – видобуток даних може навчитися зі статистики.

Мета роботи. Метою аналізу є встановлення наступних залежностей: якщо в транзакції зустрівся деякий набір елементів X, то на підставі цього можна зробити висновок про те, що інший набір елементів Y також має з'явитись в цій транзакції. Встановлення таких залежностей дає нам можливість знаходити дуже прості і інтуїтивно зрозумілі правила. Кожна людина, швидше за все, зробить деякі покупки, що відрізняються від тих, які він робив раніше, тому, покладаючись на те, що є нормальним для окремої людини.

Постановка проблеми (Introduction). Повний процес видобутку даних включає декілька етапів - від розуміння цілей проекту та наявних даних до впровадження змін процесу на основі остаточного аналізу. Три ключові обчислювальні етапи – процес навчання моделі, оцінка моделі та використання моделі. Цей розділ найбільш чіткий з класифікацією даних. Навчання моделі відбувається, коли один алгоритм застосовується до даних, про які відомий атрибут групи (або класу), щоб створити класифікатор, або алгоритм, засвоєний з даних. Потім класифікатор тестується за допомогою незалежного набору оцінок, що містить дані з відомими атрибутами. Тоді, наскільки класифікації моделі узгоджуються з відомим класом для цільового атрибута, тоді можна використовувати для визначення очікуваної точності моделі. Якщо модель досить точна, її можна використовувати для класифікації даних, для яких цільовий атрибут невідомий.

Виклад основного матеріалу. Існує багато типів інтелектуального аналізу даних, які поділяються на тип інформації (атрибутів), який відомий, і тип знань, які шукають у моделі інтелектуального аналізу даних.

Прогнозне моделювання застосовується, коли метою є оцінка значення певного цільового атрибута та існують зразки навчальних даних, для яких значення цього атрибуту відомі. Прикладом може служити класифікація, яка бере набір даних, вже розділених на заздалегідь визначені групи, і шукає закономірності в даних, що диференціюють ці групи. Потім ці виявлені закономірності можна використовувати для класифікації інших даних, де правильне позначення групи для цільового атрибуту невідомо (хоча інші атрибути можуть бути відомі).

Для досягнення найкращих результатів з обробки даних необхідний набір інструментів та методів. Деякі з найбільш часто використовуваних функцій включають:

– класифікація – це завдання прогнозування мітки або категорії нового спостереження (із набору міток чи категорій), враховуючи навчальний набір даних, що містять спостереження (або екземпляри), мітки яких вже відомі;

– кластеризація – це завдання групування спостережень (або екземплярів) у групи, відомі як кластери, на основі навчального набору даних, який містить спостереження. Мета полягає в тому, що екземпляри в одному кластері повинні бути більш схожими один на одного, ніж екземпляри в інших кластерах. На відміну від класифікації, ярлики не надаються заздалегідь;

– вимірювання є синонімом атрибута або атрибута. Приклад запису або екземпляра в наборі даних буде описаний набором вимірювань. Прикладами розмірів є зріст, стать та вік, або абсолютний поріг на одній частоті;

– домен – це метод високого рівня, де поняття ширше. Наприклад, спосіб життя людини може бути описаний в одному домені, а статус його слуху – в іншому. Кожен домен може вимірюватися кількома вимірами / ознаками, які можна згрупувати за різними методами;

– модальність – сукупність пов’язаних вимірів / ознак, що описують конкретний об’єкт чи концепцію. Наприклад, клінічна аудіограма зазвичай визначається пороговими значеннями на восьми різних частотах. Коли розміри разом описують єдине поняття, таке як аудіограма, ми називаємо це методом;

– регресія – завдання прогнозування безперервної реакції на вхідну змінну, враховуючи набір навчальних даних, що містить спостереження, безперервний відгук яких вже відомий. Це прогноз безперервної відповіді, на відміну від класифікації, де передбачається лише дискретна мітка або категорія;

– виявлення підгрупи – це завдання пошуку підмножини примірників у наборі даних, для яких встановлюється якесь підключення або залежність. Це особлива відмінність від класифікації, регресії та кластеризації, які забезпечують деяке прогнозування або опис усього набору даних.

Наприклад, виробник може розробити модель прогнозування, яка розрізняє деталі, які виходять з ладу в умовах сильної спеки, екстремального холоду чи інших умов на основі їх виробничого середовища, і ця модель може потім використовуватися для визначення відповідних застосувань для кожної деталі[1-7].

Іншим методом, що застосовується в прогнозному моделюванні, є регресійний аналіз, який можна використовувати, коли цільовим атрибутом є числове значення і метою є передбачення цього значення для нових даних.

Описове моделювання або кластеризація також поділяє дані на групи. Однак при кластеризації належні групи не відомі заздалегідь; закономірності, виявлені в результаті аналізу даних, використовуються для визначення груп. Наприклад, рекламодавець може проаналізувати загальну сукупність, щоб класифікувати потенційних споживачів у різні кластери, а потім розробити окремі рекламні кампанії, націлені на кожну групу. Виявлення шахрайства також використовує кластеризацію для виявлення груп осіб зі схожими моделями закупівель.

Видобуток шаблонів концентрується на визначенні правил, що описують конкретні закономірності в даних. Аналіз кошика ринку, який визначає предмети, які зазвичай трапляються разом в операціях закупівлі, був одним із перших застосувань для видобутку даних. Наприклад, супермаркети використовували аналіз ринкових кошиків для виявлення предметів, які часто купували разом – наприклад, магазин, де продавали рибу, також містив би запас соусу тартар.

Постановка проблеми. Незважаючи на те, що тестування таких асоціацій вже давно стало можливим, і його часто легко побачити в малих наборах даних, видобуток даних дозволив виявити менш очевидні асоціації у величезних наборах даних. Найбільший інтерес представляє відкриття несподіваних асоціацій, які можуть відкрити нові шляхи для маркетингу чи досліджень. Іншим важливим використанням вишукування шаблонів є відкриття послідовних шаблонів; наприклад, послідовності помилок або попереджень, що передують виходу з ладу обладнання, можуть бути використані для планування профілактичного обслуговування або можуть дати уявлення про недолік конструкції.

Виявлення аномалій можна розглядати як зворотну сторону кластеризації, тобто пошук екземплярів даних, які є незвичними та не відповідають жодному встановленому шаблону. Виявлення шахрайства є прикладом виявлення аномалій. Хоча виявлення шахрайства може розглядатися як проблема для прогнозного моделювання, відносна рідкість шахрайських транзакцій та швидкість, з якою злочинці розробляють нові типи шахрайства, означають, що будь-яка модель прогнозування, швидше за все, буде низькою точністю і швидко застаріє. Таким чином, виявлення аномалій натомість концентрується на моделюванні нормальної поведінки з метою виявлення незвичних транзакцій. Виявлення аномалій також використовується з різними системами моніторингу, наприклад, для виявлення вторгнень.

Розроблено безліч інших методів видобутку даних, включаючи виявлення закономірностей у даних часових рядів (наприклад, ціни на акції), потокових даних (наприклад, сенсорні мережі) та реляційне навчання (наприклад, соціальні мережі).

Багато людей турбує потенціал вторгнення в приватне життя за допомогою інтелектуального аналізу даних. Комерційні бази даних можуть містити детальні записи історії хвороби людей, купівельних операцій та використання телефону, серед інших аспектів їхнього життя.

Багато методів видобутку даних були винайдені статистиками або тепер інтегровані в статистичне програмне забезпечення; вони є розширенням стандартної статистики. Хоча майстри даних та статистики використовують подібні методи для вирішення подібних проблем, але підхід до аналізу даних відрізняється від стандартного статистичного підходу в декількох областях, таких як. Майнери даних припускають, що даних та обробної потужності достатньо. Без інтелектуального аналізу даних може бути важко розробити експерименти у діловому світі.

Це відмінності підходу, а не протилежності. Таким чином, вони проливають трохи світла на те, як бізнес-проблеми, що вирішуються майнерами даних, відрізняються від наукових проблем, які стимулювали розвиток статистики. Однією з основних відмінностей між бізнес-даними та науковими даними є те, що останні – це усічені / нецензуровані дані, а перші – усічені / цензуровані. Враховуючи методологію або алгоритм аналізу даних, часто дуже важко сказати, чи це «Статистика», чи «Видобуток даних». Незрозуміло, як слід наносити цей ярлик. Насправді, на практиці, вирішуючи реальні проблеми галузі, клієнти ніколи не запитують: «Ви майнер даних чи статистик?». Насправді їх основний інтерес полягає у вирішенні проблеми, яка стосується їх рівня задоволення, і не має значення, яку етикетку ми використовуємо. Як люди, що обслуговують клієнтів, нам (як майнеру даних або як статистику) потрібно випробувати ті статистичні прийоми чи алгоритми, які найкраще підходять для відповіді на запити замовника.

Немає сумнівів, що аналіз даних має силу трансформувати підприємства; однак впровадження рішення, яке відповідає потребам усіх зацікавлених сторін, часто може зупинити вибір платформи. Широкий спектр доступних для аналітиків варіантів, включаючи мови з відкритим кодом, такі як R та Python, та такі знайомі інструменти, як Excel, у поєднанні з різноманітністю та складністю інструментів та алгоритмів можуть ще більше ускладнити процес.

Підприємства, які отримують найбільшу цінність із видобутку даних, зазвичай вибирають платформу, яка:

– включає найкращі практики для своєї галузі або типу проекту. Наприклад, організації охорони здоров'я мають інші потреби, ніж компанії електронної комерції;

– управляє усім життєвим циклом видобутку даних, починаючи від дослідження даних і закінчуючи виробництвом;

– вирівнюється з корпоративними додатками, включаючи BI-системи, CRM, ERP, фінансове та інше корпоративне програмне забезпечення, з яким воно має взаємодіяти для максимальної віддачі інвестицій;

– інтегрується з провідними мовами з відкритим кодом, надаючи розробникам та науковцям даних гнучкість та інструменти співпраці для створення інноваційних програм;

– відповідає потребам IT, науковців даних та аналітиків, одночасно обслуговуючи потреби бізнес-користувачів у звітності та візуалізації;

– платформа великих даних Talend надає повний набір можливостей управління даними та інтеграції даних, щоб допомогти командам з обробки даних швидше реагувати на потреби свого бізнесу.

Засноване на відкритій, масштабованій архітектурі та з інструментами для реляційних баз даних, плоских файлів, хмарних додатків та платформ, це рішення доповнює вашу платформу видобутку даних, надаючи більше даних для роботи за менший час – що перетворюється на швидший час для розуміння та конкурентні переваги[1–3].

Застосовуючи передові аналітичні методи в Data Mining, підприємства збільшують доходи, максимізують операційну ефективність, скорочують витрати та покращують задоволеність споживачів. Тоді як статистика дає можливість будувати прогнозні моделі або розробляти класифікації, які впливають на ваш підсумок. Без статистики немає ефективного аналізу.

Без ефективного аналізу не існує бізнес-аналітики. Без бізнес-аналітики, як ви можете сподіватися засвоїти гігабайти даних і послідовно приймати рішення, які будуть випереджати вашу конкуренцію. За допомогою статистики ви можете перетворити свої дані на знання про свої бізнес-процеси. Використання статистичних даних у процесі аналізу даних може суттєво вплинути на всі сфери вашої організації. На сьогоднішній день статистичне програмне забезпечення може покращити вашу конкурентоспроможність від цеху до торгового виконання.

Аналіз останніх досліджень і публікацій. На сьогоднішній діловій арені постійним завданням є не відставати від ринкових тенденцій та прогнозувати майбутні результати. Щоб збільшити частку ринку та ефективно працювати, ви не можете дозволити собі не використовувати статистику в Data Mining.

Якщо ви не видобуваєте свої дані настільки, наскільки це варте, ви винні у недостатньому використанні одного з найбільших активів вашої компанії. Хоча в статистиці є субдисципліна, яка стосується опису, огляд у будь-якому загальному тексті статистики продемонструє, що головним питанням є те, як робити заяви про сукупність, коли спостерігається лише вибірка.

Однак проблема видобутку даних часто має у своєму розпорядженні всю сукупність даних, наприклад деталі всієї робочої сили корпорації тощо. У таких випадках поняття перевірки статистичної значущості втрачає значення. З іншого боку, основною метою видобутку даних є відкриття, це не стосується тих областей статистики, які передбачають, як найкраще зібрати дані, перш за все, щоб відповісти на конкретне питання, таке як експериментальний дизайн та дизайн обстеження.

Видобуток даних по суті припускає, що дані вже зібрані, і стосується того, як відкрити їх секрети. Для подальшого вивчення схожості та відмінності між статистикою та засобами збору даних читачі можуть звернутися до Hand (1999a, 1999b).

Громадянські лібертаріанці вважають деякі бази даних, що зберігаються підприємствами та урядами, не обгрунтованим вторгненням та запрошенням до зловживань. Наприклад, Американський союз громадянських свобод подав позов до Агентства національної безпеки США (NSA) за звинуваченням у беззастережному шпигунстві за американськими громадянами шляхом придбання записів дзвінків у деяких американських телекомунікаційних компаній.

Програма, розпочата в 2001 р., була відкрита громадськістю лише в 2006 р., коли інформація почала просочуватися. Часто ризик полягає не в самому видобутку даних (який, як правило, спрямований на отримання загальних знань, а не на вивчення інформації про конкретні проблеми), а внаслідок неправильного використання або неналежного розкриття інформації в цих базах даних.

У Сполучених Штатах зараз багато федеральних відомств зобов'язані складати щорічні звіти, які конкретно розглядають наслідки конфіденційності їх проектів з видобутку даних. Законодавство США, яке вимагає звітів про конфіденційність від федеральних агентств, визначає видобуток даних досить строго, як «... аналіз для виявлення або виявлення прогнозного зразка або аномалії, що свідчить про терористичну чи злочинну діяльність з боку будь-якої фізичної особи».

Оскільки різні місцеві, національні та міжнародні правоохоронні органи почали обмінюватися або інтегрувати свої бази даних, потенціал зловживань або порушень безпеки змусив уряди співпрацювати з промисловістю над розробкою більш захищених комп'ютерів та мереж. Зокрема, було проведено дослідження методів захисту даних, які зберігають конфіденційність, які працюють на спотворених, перетворених або зашифрованих даних, щоб зменшити ризик розкриття даних будь-якої особи.

Інтелектуальний аналіз даних розвивається, комерційним прикладом цього стала премія Netflix у розмірі 1 мільйон доларів. Американська компанія Netflix, яка пропонує прокат фільмів, доставлених поштою або в Інтернеті, розпочала конкурс у 2006 році, щоб перевірити, чи зможе хтось покращити на 10 % свою систему рекомендацій – алгоритм прогнозування переваг кінофільмів людини на основі попередніх даних прокату. Приз був вручений 21 вересня 2009 року «Прагматичному хаосу» BellKor – команді з семи математиків, інформатиків та інженерів із США, Канади, Австрії та Ізраїлю, які досягли 10-відсоткової мети 26 червня 2009 р., і завершили свою перемогу вдосконаленим алгоритмом через 30 днів. Трирічний відкритий конкурс спонукав багатьох розумних нововведень щодо інтелектуального аналізу даних у учасників. Наприклад, на конференціях 2007 та 2008 рр. З питань виявлення знань та видобутку даних були проведені семінари, присвячені премії Netflix, на яких були представлені наукові роботи на теми, починаючи від нових методів спільної фільтрації і закінчуючи більш швидкою факторизацією матриць

(ключовим компонентом багатьох систем рекомендацій). Побоювання щодо конфіденційності таких даних також призвели до прогресу в розумінні конфіденційності та анонімності [3–7].

Виділення невирішених раніше частин загальної проблеми. Однак аналіз даних не є панацеєю, і результати слід розглядати з такою ж обережністю, як і будь-який статистичний аналіз. Однією з сильних сторін інтелектуального аналізу даних є можливість аналізу кількості даних, які було б недоцільно аналізувати вручну, а знайдені закономірності можуть бути складними та важкими для розуміння людьми; ця складність вимагає обережності при оцінці закономірностей. Тим не менше, методи статистичної оцінки можуть призвести до знань, які не мають людських упереджень, а великий обсяг даних може зменшити упередження, властиві меншим вибіркам. При правильному використанні інтелектуальний аналіз даних дає цінну інформацію про великі масиви даних, які в іншому випадку не було б практично або можливо отримати. «Mining» англійською означає «видобуток корисних копалин», а пошук закономірностей у величезній кількості даних дійсно схожий на цей процес.

Перш ніж використовувати технологію Data Mining, необхідно ретельно проаналізувати її проблеми:

– Data Mining не може замінити аналітика;
– не може складати розробки і експлуатації додатку Data Mining;
– потрібна підвищена кваліфікація користувача;
– витягання корисних відомостей неможливе без доброго розуміння суті даних;
– складність підготовки даних;
– висока вартість;
– вимога наявності достатньої кількості репрезентативних даних.

Data Mining тісно пов’язана з різними дисциплінами, що засновані на інформаційних технологіях та математичних методах обробки інформації (рис. 1).

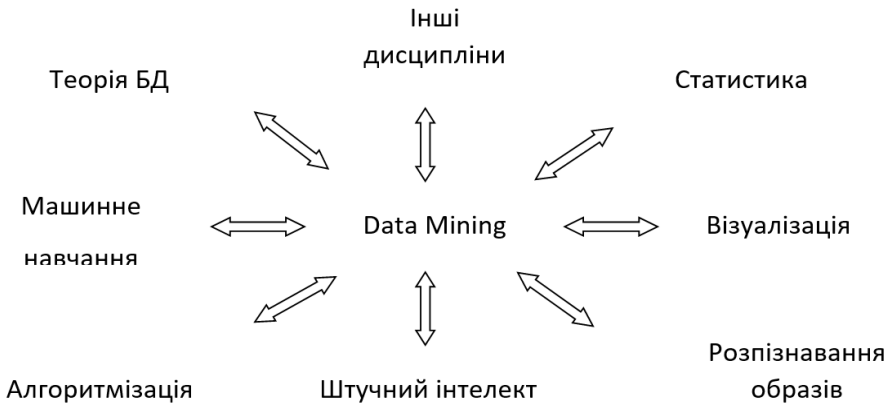

Рис. 1. Методи обробки інформації

Насамперед, обсяги продажів, запаси цін і відсоток відображення продукту дуже важливо прогнозувати, тому що вони могли залежати від дій кількох попередніх. Наступні, більше складні методи (наприклад, логістична регресія, рішення дерев або нейронні мережі) можуть бути необхідними для прогнозування майбутніх цінностей. Одні і ті типові моделі часто можна використовувати як для регресії, так і для класифікації. Наприклад, корзина (Класифікація Rvyhod) алгоритм рішень може бути використаний для побудови шпалерів класифікованих дерев (для класифікації категорійних відкритих) та регресійних дерев (для прогнозування певних безперервних вікліків). Згенеровані функції можна класифікувати за функціями часу та функціями дії, як зведено в таблиці 1. Було створено чотири функції часу: T_time, A_time, S_time та E_time, що вказує загальний час відгуку, час дії, проведений у процесі, час початку, витрачений на першу дію , і час закінчення, витрачений на останню дію, відповідно.

Передбачалося, що студенти з різними рівнями здібностей можуть відрізнятися за часом, коли вони читають запитання (час початку, витрачений на першу дію), час, який вони витрачають під час відповіді (час дії, проведений в процесі), і час, який вони використали, щоб зробити остаточний рішення (час закінчення, витрачений на останню дію). Різні дослідники пропонували різні підходи спільного моделювання як для точності відповіді, так і для часу відгуку, які пояснюють взаємозв'язок між ними (наприклад, van der Linden, 2007; Bolsinova et al., 2017). Таким чином, очікується, що загальний час відгуку також відрізнятиметься. Видобуток даних означає статистику чи більше, ніж статистику?

Формування цілі статті. Насправді, видобуток даних – це термін, що є синонімом поняття «поглиблення даних» або «вилов даних» і використовувався для опису процесу тралення через дані з надією виявити закономірності. Дані, які є не просто однорідними, мають відмінності, які можна інтерпретувати як закономірності. Проблема полягає в тому, що багато з цих "моделей" будуть просто продуктом випадкових коливань і не представлятимуть жодної базової структури. Тоді для статистиків термін «видобуток даних» передає відчуття наївної надії, марно бореться з холодними реаліями випадковості. Однак для інших дослідників цей термін розглядається більш позитивно.

Поверхово, звичайно, те, що тут описується, є не що інше, як дослідницький аналіз даних, діяльність, яка проводилася з моменту першого аналізу даних і яка досягла більшої респектабельності. Але різниця є, і саме ця різниця пояснює, чому статистики повільно зациклювались на можливостях. Ця різниця полягає у великому розмірі наборів даних, які зараз доступні. Статистичні спеціалісти зазвичай не займаються наборами даних, що містять багато мільйонів або навіть мільярдів записів.

Більше того, спеціальні методи зберігання та маніпулювання, необхідні для їх роботи, були розроблені абсолютно різними інтелектуальними спільнотами, ніж статистики. Можливо, не буде перебільшенням сказати, що більшість статистиків стурбовані первинним аналізом даних. З іншого боку, видобуток даних повністю займається вторинним аналізом даних. Насправді ми могли б визначити «видобуток даних» як процес вторинного аналізу великих баз даних, спрямованих на пошук не підозрюваних відносин, які представляють інтерес або цінність для власників баз даних. З цього ми бачимо, що «Видобуток даних» є в основному індуктивною вправою, на відміну від гіпотетико-дедуктивного підходу, який часто розглядається як парадигма прогресу сучасної науки.

Однак у цьому дослідженні особливості дії були створені шляхом кодування різної довжини сусідніх послідовностей дій разом. Таким чином, це дослідження створило 12 функцій дії, що складаються лише з однієї дії (уніграми), 18 функцій дії, що містять дві впорядковані суміжні дії (біграми), та 2 особливості дії, створені з чотирьох послідовних дій. Крім того, передбачалося, що всі створені послідовності дій мають однакову важливість.

Загальна закономірність полягає в тому, що нова ідея буде висунута дослідниками з якоїсь іншої дисципліни, приверне значний інтерес, і лише тоді статистик залучиться. Існує реальна небезпека того, що статистика та статистики будуть сприйматися як незначна недоречність і як така, що не відіграє фундаментальної ролі у науковому та широкому житті, яка належним чином виконується [4–6].

Статистиці необхідно терміново взяти участь у обробці даних. В основному класична статистика має справу з числовими даними. Але в наш час бази даних містять дані інших видів. Чотири очевидні приклади – це дані про зображення, звукові дані, текстові дані та географічні дані. Основне питання інтелектуального аналізу даних полягає у пошуку цікавих закономірностей та структур у цих базах даних. Звичайно, не можна просто попросити комп'ютер «шукати цікаві закономірності» або «перевіряти, чи є в даних якась структура».

Перш ніж це зробити, потрібно визначити, що вони мають на увазі під шаблонами чи структурою. І перед тим, як це зробити, потрібно вирішити, що означає «цікаве». Загалом, звичайно, те, що цікавить, буде дуже залежати від домену програми. Під час пошуку закономірностей або структур потрібно зробити компроміс між конкретним та загальним. Суть інтелектуального аналізу даних полягає в тому, що людина не знає, яку саме структуру шукає, тому доцільним буде досить загальне визначення. З іншого боку, занадто загальне визначення дасть занадто багато моделей кандидатів. Оскільки пошук шаблонів видасть велику кількість шаблонів-кандидатів, то існує велика ймовірність того, що помилкові конфігурації даних будуть визначені як шаблони.

Можливо, рішення буде знайдено лише шляхом виходу за рамки загальноприйнятої імовірнісної статистичної системи, яка використовує правила скорингу замість імовірнісних інтерпретацій. Проблема подібна до проблеми надмірної підгонки статистичних моделей, що викликало новий інтерес завдяки розробці надзвичайно гнучких моделей, таких як нейронні мережі. Статистичні дані, викладені в традиційній навчальній програмі, можуть бути описані як такі, що характеризуються невеликими, чистими, статичними та випадковими вибірками, і часто збираються для відповіді на конкретне питання. Жодне з них не застосовується в контексті аналізу даних. Оскільки для класичного статистика набір даних з кількома тисячами спостережень може бути великим, але для майнера даних це мало.

Можна стверджувати, що, хоча між аналізом даних та статистикою є дуже багато спільного, вони мають свої унікальні ідентичності. Ми можемо також стверджувати, що особливості проблеми, з якою вони вирішують, а також характер та обмеження методів, які вони використовують, можуть призвести до плідного синергізму. Насправді існують глибокі теоретичні проблеми, що виникають із проблем видобутку даних, які мали б користь із статистичної точки зору та розуміння. Здається, що видобуток даних можна поставити в контекст більшої статистики, яку можна визначити, принаймні вільно, як «все, що пов'язане з навчанням на даних». Більша статистика має тенденцію бути всеохоплюючою, еклектичною щодо методології, тісно пов'язаною з іншими дисциплінами, і практикується багатьма за межами традиційної професійної статистики та за межами наукових кіл.

Насправді головне питання полягає в тому, що якщо студент, який навчається за програмою статистики, хоче зробити кар'єру в галузі прикладної статистики в галузі, чому ми повинні навчити його чи її? Як найкраще ми можемо забезпечити його / її ефективність у майбутній ролі? Поряд із регулярними курсами статистики мають існувати курси, пов'язані з різними алгоритмами інтелектуального аналізу даних, та дослідження щодо використання програмного забезпечення, що включає реалізацію цих алгоритмів. Ми повинні піддавати студентів статистики різним базам даних, різним типам наборів даних з різних доменів. Студенти статистики також повинні знати про проблеми зберігання, доступу та маніпулювання великими обсягами даних з ефективною візуалізацією та поданням.

Немає сумнівів, що між статистиками та майнерами даних існує взаємне невігластво. Однією з причин такої взаємної невігластва є консервативність статистики проти ставлення до ризиків у обчислювальній техніці. Зараз широко визнано, що прогрес у видобутку даних вимагатиме злиття думок спеціалістів з обчислювальної техніки та статистиків. Це щось на зразок звинувачення статистичної професії, що мало хто

із статистиків зайнявся глибоким процесом з видобутку даних. Статистики мають чому навчити видобувачів даних, тоді як обробники даних мають багато захоплюючих та нових проблем, на які статистики навіть не почали дивитись. Існує можливість надзвичайно корисної синергії між статистиками та майнерами даних. Однак більшість майнерів даних, як правило, не знають статистики та домену клієнта; статистики, як правило, не знають про аналіз даних та домен клієнта; а клієнти, як правило, не знають про аналіз даних та статистику. На жаль, вони також мають тенденцію стримуватися різними точками зору; інформатики зосереджуються на маніпуляціях з базами даних та алгоритмах обробки; статистики зосереджуються на виявленні та усуненні невизначеностей: а клієнти - на інтеграції знань у своїй область [7–9].

Більшість майнерів даних, як видається, мають відносно невелику офіційну статистичну експертизу. Отже, вони іноді роблять помилки, яких кваліфіковані статистики уникали б такими очевидними. Це означає, що майнери даних повинні брати до уваги статистичну інформацію щодо потенційних можливостей для помилкових асоціацій та суттєвих аспектів проти статистичної значущості, що призводить до вимог щодо підготовки майнерів даних із статистики або випускників статистики в галузі аналізу даних. Цей підхід повинен бути практичним та базуватися на прикладах, і бажано переорієнтувати традиційну навчальну програму статистики, наголошуючи на змінах у зборі та аналізі даних, які виникли за останні п'ятнадцять років.

Видобуток даних та статистика неминуче будуть зростати один до одного найближчим часом, оскільки видобуток даних не стане відкриттям знань без статистичного мислення, статистика не зможе досягти успіху на масивних та складних наборах даних без підходів до аналізу даних. Пам’ятайте, що відкриття знань ґрунтується на трьох збалансованих ногах інформатики, статистики та знань клієнтів; воно не буде стояти ні на одній нозі, ні на двох ногах, ані навіть на трьох неврівноважених ногах. Отже, успішне відкриття знань потребує значної співпраці цих трьох спільнот[8].

Усі сторони повинні розширити свою увагу, доки справжня співпраця і видобуток золота не стануть реальністю. Проблема зрілості полягає в тому, щоб майнери даних, статистики та клієнти визнавали свою залежність один від одного і щоб усі вони розширювали свою увагу, доки справжня співпраця не стане реальністю. Найважливішим викликом для нас усіх є розгляд викликів як можливостей для нашого спільного успіху. Усі сторони повинні розширити свою увагу, доки справжня співпраця і видобуток золота не стануть реальністю.

Було б непогано подумати, що дослідники з цих дисциплін об’єднаються, щоб об’єднати свої різні точки зору та підходи, щоб вирішити справді важливі проблеми, з якими ми стикаємось у цьому сучасному світі, багатому на дані. Немає сумнівів, що аналіз даних є «статистично інтелектуальним». Для подальшого обговорення цього конкретного питання читачі можуть проконсультуватися з Ганеш (2002) та Куоненом (2004). Статистичне дослідження видобутку даних активно проводилося в останні роки за допомогою дослідницьких статей, що з'являються в статистичних, а також нестатистичних журналах.

З точки зору розглядуваної проблеми, це має прямий вплив, як узагальнення результатів, якість рекомендацій щодо можливого вдосконалення процесу тощо. Такі речі, як правило, не розглядаються з великою глибиною в будь-якому нестатістичному курсі. Важливо розуміти, що не всі набори даних є масивними. Багато разів він чудово вписується в пам’ять їх місцевого робочого столу чи ноутбука. З огляду на це, для аналітиків даних тим більше важливо більше зосереджуватись на оволодінні наукою та мистецтвом моделювання та аналізу. Отже, курси з видобутку даних, орієнтованих на статистику, повинні бути передбачені в навчальній програмі «статистики». Однак дискусійним питанням найближчого майбутнього є «Чи повинна статистика як поле охоплювати видобуток даних як субдисципліну або залишати це за комп'ютерними вченими?».

Оскільки організації продовжують засипатися величезними обсягами внутрішніх та зовнішніх даних, їм потрібно переганяти цю сировину до ефективних знань із швидкістю, необхідною для їхнього бізнесу.

Підприємства в будь-якій галузі покладаються на Talend, щоб допомогти їм пришвидшити аналіз даних. Наша сучасна платформа для інтеграції даних дає можливість користувачам працювати розумніше і швидше в командах, що дозволяє їм розробляти та розгортати наскрізні завдання інтеграції даних у десять разів швидше, ніж ручне кодування.

Нарешті, звертаючись до питання «Чи слід включати видобуток даних до навчальної програми статистики?», Відповідь неминуче повинна бути захопленим «Так». Зрештою, видобуток даних – це, по суті, статистичний аналіз даних. Оскільки досвід свідчить нам, що статистика як дисципліна має погані результати для своєчасного визнання важливих ідей, тому, безумовно, є шанс поліпшити репутацію цієї дисципліни. Це було б великою втратою для репутації статистики як дисципліни, а також для окремих статистиків, якби ці можливості не були використані [2].

Виклад основного матеріалу та отриманих наукових результатів. Комп'ютерні вчені обіграли статистиків, пропонуючи курси з видобутку даних, з часів значного вдосконалення обчислювальної потужності в 1990-х роках. Однак більшість, якщо не всі ці пропозиції, зосереджені на реалізації ефективних алгоритмів з точки зору машинного навчання. Незважаючи на те, що на третинному рівні пропонуються декілька курсів статистичних методів статистичних досліджень, орієнтованих на послуги, часто через статистичні підрозділи або консультаційні центри, курси зі статистично орієнтованого аналізу даних не були широко доступні в меню.

Ці курси повинні забезпечити широку статистичну перспективу видобутку даних на бакалаврському рівні та орієнтовані на студентів, що спеціалізуються на статистиці, та на тих, хто спеціалізується в таких областях, як інформатика, управління базами даних та бізнес-дослідження. Такі курси повинні

забезпечувати широке висвітлення методів, які часто класифікуються як контрольовані чи неконтрольовані, та передбачати описове або прогнозне введення в моделювання.

Типовий рецепт може мати форму: «… вступ до додатків для інтелектуального аналізу даних, включаючи підготовку та зберігання даних; запити, асоціації, ринкові кошики та методи індукції правил; прогнозування за допомогою регресії, дерев рішень та нейронних мереж; кластеризація з використанням ієрархічних методів та самоврядування – організація карт; класифікація за допомогою дерев та нейронних мереж; візуальний підхід з реальними прикладами та тематичними дослідженнями; використання провідних програмних засобів для обробки даних;…». Просунутий курс статистичного видобутку даних, скажімо на рівні аспірантури, може включати статистичну підтримку загальновживаних методів, а також широкий спектр нових розробок, таких як генетичні алгоритми, видобуток тексту, алгоритми створення мішків, ударів та посилення, а також байєсівські мережі переконань [6].

В принципі, статистична експертна система буде втілювати велику базу розумного розуміння процесу аналізу даних, яку вона може автоматично застосовувати до відносно невеликого набору даних. Тоді як система аналізу даних, яка втілює невелику базу розумного розуміння, але яка застосовує її до великого набору даних.

В обох випадках заявка є автоматичною, хоча в обох випадках взаємодія з дослідником є принциповою. У статистичній системі експертів програма керує аналізом, дотримуючись статистичної стратегії, оскільки користувач не має достатньої статистичної експертизи для цього. Тоді як у програмі інтелектуального аналізу даних програма керує аналізом, оскільки користувач не має достатньо ресурсів для ручного дослідження мільярдів записів та сотень тисяч потенційних моделей. Для детального уявлення про видобуток даних у порівнянні зі статистикою читачі можуть звернутися до Hand (1998). Враховуючи ці подібності між двома підприємствами, розумно запитати, чи є уроки, які спільнота з видобутку даних може отримати з досвіду роботи статистичної експертної системи. Відповідь, безумовно, «Так».

Сан-Педро та ін. (2012) показали, що створені функції повинні бути теоретично важливими для конструкції для досягнення кращої інтерпретації та ефективності. За їх пропозицією, функції були сформовані як показники здатності вирішувати проблеми, виміряні цим елементом, що підтримується рубрикою оцінювання. Наприклад, одна послідовність дій складалася з чотирьох дій, яка була закодована як "city_con_daily_cancel", що має вирішальне значення для оцінки. Якщо студент спочатку вибрав "city_subway" для екскурсії містом, потім скористався тарифом на концесію студента («концесія»), подивився на ціну щоденного пропуску («щодня») далі і, нарешті, він / вона натиснув «Скасувати», щоб побачити інший варіант, ця послідовність дій необхідна, але недостатня для повного зарахування.

Остаточний перекодований набір даних для аналізу складається з 426 учнів у вигляді рядків та 36 об’єктів (у т.ч. 32 функції послідовності дій та 4 функції часу) у вигляді стовпців. Бали для кожного студента служили відомими позначками при застосуванні контрольованих методів навчання. Частоту кожної генерованої функції дії розраховували для кожного студента. Вибір ознак повинен базуватися як на теоретичній основі, так і на використаних алгоритмах. Оскільки в даному дослідженні особливості були сформовані з чисто теоретичної точки зору, такого вибору не потрібно.

Ще дві проблеми, які потребують розгляду, – це зайві змінні та змінні з невеликою дисперсією. Деревовидні методи добре справляються з цими двома проблемами та мають вбудовані механізми вибору функцій. Важливість особливостей, вказана методами, заснованими на деревах, показана на рис. 3. І в випадковому посиленні лісу, і в градієнті найбільш важливим є "city_con_daily_cancel". Наступним важливим є "other_buy", що означає, що студент не вибрав trip_4 до дії "Buy".

Важливість ознак, вказана методами, заснованими на дереві, особливо корисна, коли вибір потрібно робити серед сотень ознак. Це може допомогти звузити кількість функцій для відстеження, аналізу та інтерпретації. Точність класифікації машини опорного вектора (SVM) знижується через надлишкові змінні. Однак, враховуючи кількість ознак (36) у цьому дослідженні порівняно невелика, видалення висококорельованих змінних ($\rho \geq 0{,}8$) не покращило точність класифікації для SVM.

Це дослідження демонструє, як за допомогою методів видобутку даних зіставити обрані особливості (як дії, так і час) із результатами роботи учнів за цим предметом для вирішення проблем у 2012 PISA. З огляду на те, що результати балів студентів доступні у файлі даних, алгоритми контрольованого навчання можна навчити, щоб допомогти класифікувати студентів на основі їх відомих показників результатів (тобто, бальної категорії) у наборі даних тренінгу, тоді як неконтрольовані алгоритми навчання класифікують учнів на групи на основі вхідних змінних не знаючи їх продуктивності. Жодних припущень щодо розподілу даних щодо цих методів видобутку даних не робиться[5-9].

Для розробки класифікаторів досліджуються чотири контрольовані методи навчання: Дерево класифікації та регресії (CART), посилення градієнта, випадковий ліс та SVM, тоді як для подальшого вивчення використовуються два некеровані методи навчання – Карта самоорганізації (SOM) та k-засоби. різні стратегії, що використовуються студентами як в одній і тій же категорії, так і в різних балах.

CART був обраний, оскільки він ефективно працював у попередньому дослідженні (DiCerbo and Kidwai, 2013) і відомий своїми швидкими обчисленнями та простою інтерпретацією. Однак він може не мати оптимальних показників порівняно з іншими методами. Крім того, невеликі зміни даних можуть кардинально змінити структуру дерева (Kuhn, 2013). Таким чином, посилення градієнта та випадковий ліс, які можуть покращити продуктивність дерев за допомогою ансамблевих методів, також використовувались для порівняння. Хоча SVM ще не застосовувався багато в аналізі даних процесу, він застосовувався як одна

з найпопулярніших та гнучких контрольованих методик навчання для інших психометричних аналізів, таких як автоматичний бал (Vapnik, 1995). Два алгоритми кластеризації, SOM і k-середні, були застосовані при аналізі даних процесу в файлах журналів (Stevens and Casillas, 2006; Fossey, 2017). Дослідники пропонують використовувати кілька методів кластеризації для перевірки рішень кластеризації (Xu et al., 2013). Всі аналізи проводились у програмній програмі Rstudio (RStudio Team, 2017).

По-перше, для демонстрації аналізу даних процесу, використовуючи як контрольовані, так і неконтрольовані методи, у поточному дослідженні були представлені конкретні кроки у формуванні ознак, виборі ознак, розробці класифікатора та оцінці результатів. Серед усіх етапів генерація ознак була найважливішою, оскільки якість ознак значною мірою визначає результати класифікації.

Хороші особливості слід створювати на основі глибокого розуміння процедури підрахунку предметів та конструкції. Ключові послідовності дій, що дозволяють розрізнити правильні та неправильні відповіді, слугували особливостями з хорошою продуктивністю. Несподівано характеристики часу, включаючи загальний час відгуку та його частини, не виявились важливими ознаками для класифікації. Це означає, що значні розбіжності у часі відповіді існували в кожній групі балів, і різниця у розподілі часу відповіді між групами була недостатньо великою, щоб чітко розмежувати групи.

Це дослідження породило особливості, засновані на теоретичних уявленнях про конструкцію, що вимірюється та використовується студентами як одиниця аналізу. Дані можуть бути структуровані іншими способами відповідно до різних дослідницьких питань. Наприклад, замість використання студентів як одиниці аналізу, зроблені студентами спроби можуть бути використані як рядки, а дії – як стовпці, тоді спроби можна класифікувати замість людей. Фоссі (2017) включив детальний підручник з алгоритмів кластеризації з такою структурою даних в оцінку на основі гри.

По-друге, для оцінки узгодженості класифікації цих часто використовуваних методів видобутку даних, поточне дослідження порівнювало чотири контрольовані методики з різними властивостями, а саме CART, посилення градієнта, випадковий ліс та SVM. Всі чотири методи досягли задовільної точності класифікації на основі різних вимірювань результатів, при цьому посилення градієнта показало дещо кращу загальну точність та значення Каппи. Загалом, легка інтерпретація та графічна візуалізація – головні переваги дерев. Дерева також добре справляються з галасливими та неповними даними (James et al., 2013).

Однак на дерева легко впливати навіть незначні зміни даних через його ієрархічну структуру розбиття (Hastie et al., 2009). SVM, навпаки, добре узагальнює, оскільки як тільки гіперплан знайдено, незначні зміни даних не можуть сильно вплинути на гіперплан (James et al., 2013). З огляду на конкретний набір даних у поточному дослідженні, навіть метод CART працював дуже добре. Крім того, метод CART можна легко зрозуміти та надати достатньо інформації про детальну класифікацію між кожною категорією балів та всередині них.

Таким чином, на основі результатів поточного дослідження, методу CART достатньо для майбутніх досліджень на подібних наборах даних. Алгоритми навчання без нагляду, SOM та k-середні, також показали збіжні результати кластеризації на основі значень DBI та Kappa. В остаточному рішенні кластеризації студенти були згруповані в 9 кластерів, розкриваючи конкретні процеси вирішення проблем, які вони пройшли [2–5].

По-третє, контрольовані та неконтрольовані методи навчання служать для відповіді на різні дослідницькі питання. Методи навчання під контролем можуть бути використані для навчання алгоритму прогнозування членства в майбутніх даних, наприклад автоматичного підрахунку балів.

Методи, що не контролюються, можуть виявити закономірності стратегії вирішення проблем та ще більше розмежувати учнів в одній і тій же категорії балів. Це особливо корисно для цілей формування. Студентам можуть бути надані більш докладні та індивідуальні діагностичні звіти. Вчителі можуть краще зрозуміти сильні та слабкі сторони учнів і відповідно скорегувати інструкції в класі або забезпечити більш цілеспрямоване навчання для конкретних учнів.

Крім того, необхідно перевірити будь-які вказівки на шахрайство в помилково класифікованих або відхилених випадках від обох типів методів аналізу даних. Наприклад, студенти правильно відповіли на питання протягом надзвичайно короткого проміжку часу, що може означати компроміс.

Це дослідження має свої обмеження. Інші методи видобутку даних, такі як інші алгоритми дерев рішень та алгоритми кластеризації, варті вивчення. Однак процедуру, продемонстровану в цьому дослідженні, можна легко узагальнити до інших алгоритмів. Крім того, шість методів порівнювали на основі одного і того ж набору даних, а не даних за різних умов. Тому узагальнення поточного дослідження обмежене через такі фактори, як обсяг вибірки та кількість ознак.

У майбутніх дослідженнях можна використовувати більший обсяг вибірки та виділити більше функцій із більш складних сценаріїв оцінки. Нарешті, основне увагу приділено поточному дослідженню лише для одного предмета з дидактичною метою. У майбутньому дослідженні можна одночасно аналізувати дані процесу для більшої кількості предметів, щоб отримати всебічне уявлення про студентів.

Підводячи підсумок, вибір методів аналізу даних для аналізу даних процесу в процесі оцінки залежить від мети аналізу та структури даних. Контрольовані та неконтрольовані методи, по суті, служать різним цілям для видобування даних, причому перший – як підтверджуючий підхід, а другий – як дослідницький.

Кластеризація є загальним завданням під час аналізу дослідницьких даних. Навчальним завданням без нагляду є виявлення значущих групувань даних у класи, які заздалегідь не відомі («апріорі» чи

«попередні»), а навпаки, вивчаються з даних. При кластеризації точки в наборі даних групуються таким чином, що точки в межах даної групи в певному сенсі більш схожі між собою, ніж точки поза групою. Такий аналіз дослідницьких даних корисний у нашому контексті, оскільки може допомогти розкрити загальні профілі слуху чи способу життя. Наприклад, аудіограми можна кластеризувати, щоб аудіограми подібної форми віднести до однієї групи (Lee, Hwang, Hou, & Liu, 2010). Зведена статистика вивчених кластерів часто може надати більш інформативний, високорівневий погляд на склад користувачів і, можливо, навіть на етіологію.

Крім того, така кластеризація може сприяти вибору конкретного пристрою, пристосованого для кращого відповідності загальним профілям, наприклад, зворотного нахилу у порівнянні з аудіограмами пресбіакузису. З новим на ринку пристроєм, коли пов'язана база даних зростала, такий вибір можна було перевірити, порівнюючи показники результатів між пристроями. Поняття «хорошої» (як у розумній та надійній) групуванні може досить сильно відрізнятися, і тому в літературі з аналізу даних є численні алгоритми кластеризації.

Простим, але часто ефективним робочим способом для кластеризації є K-засоби (Wu & Kumar, 2009, с. 21–33). K-означає – це метод пошуку кількості кластерів K, такий, що сума дисперсії всіх кластерів зведена до мінімуму. Це робиться щодо деякого метричного простору, який зазвичай є евклідовим. Засіб багатовимірного інтелектуального аналізу даних повинен знаходити закономірності як в тих, що деталізуються, так і в агрегованих з різним ступенем узагальнення даних.

Аналіз багатовимірних даних повинен будуватися над гіперкубом спеціального вигляду, вічка якого містять не довільні чисельні значення (кількість подій, об'єм продажів, сума зібраних податків), а числа, що визначають вірогідність відповідного поєднання значень атрибутів.

Проекції такого гіперкуба (що виключають з розгляду окремі вимірювання) також повинні досліджуватися на предмет пошуку закономірностей. J. Han пропонує ще більш просту назву – "OLAP Mining" і висуває декілька варіантів інтеграції двох технологій:

– "Cubing then mining". Можливість виконання інтелектуального аналізу повинна забезпечуватися над будь-яким результатом запиту до багатовимірного концептуального уявлення, тобто над будь - яким фрагментом будь-якої проекції гіперкуба показників;

– "Mining then cubing". Подібно даним, витягнутим з сховища, результати інтелектуального аналізу повинні представлятися в гіперкубічній формі для подальшого багатовимірного аналізу;

– "Cubing while mining". Цей гнучкий спосіб інтеграції дозволяє автоматично активізувати однотипні механізми інтелектуальної обробки.

Видобуток даних може запропонувати великі перспективи для пошуку нових та складних взаємозв'язків у наборах даних, але через розмір наборів даних та кількість порівнянь, зроблених під час видобутку, багато з них можуть бути помилковими. Окрім статистичної впевненості, завжди потрібні інтерпретація та перевірка експертів для того, щоб надати контекст та витягти потенційну цінність із отриманих результатів. Несподівані висновки, якщо вони можуть призвести до генерування раціональних гіпотез, можуть викликати нові напрямки цілеспрямованих досліджень [1–3].

Для такого завдання, як класифікація або регресія, метою є прогнозована сила вивченої моделі на нових екземплярах даних, що викликає кілька запитань: (a) де щойно отриманий набір даних вписується в шаблони з історичних наборів даних, і, якщо цього не відбувається, (b) чи потребує оновлення модель, і нарешті, (c) як це впливає на наші рішення щодо управління пацієнтами? Модель, яка не узагальнює можливість отримувати розумні прогнози за новими даними, а моделює лише навчальні дані, називається «переобладнаною» і порівняно марною.

Висновки та перспективи подальшого розвитку у цьому напрямі. Отже, ефективність моделі слід оцінювати на основі даних, які є окремими від даних, що використовуються для навчання або оновлення моделі. Передбачувана ефективність моделі на основі навчальних даних буде надмірно впевненою, оскільки модель може бути адаптована відповідно до побачених даних і, отже, може перевершити дані. Це аналогічно забезпеченню студента відповідями перед іспитом, щоб він міг вивчити їх напам'ять і очікувати, що їх результати іспиту будуть неупередженим показником знань студента з загальної теми.

Статистики та майнери даних вирішують подібні проблеми. Однак через історичні відмінності та відмінності в характері проблем існують певні відмінності в підходах. Існує очевидний потенціал, можливості та навіть ажіотаж у видобутку даних для відкриття у великому наборі даних. Жвава дискусія щодо проблеми «Різниця між статистикою та видобутком даних» дійшла висновку, що неважливо те, що ви називаєте цим або видобуванням даних, або статистикою.

Оскільки «обчислення» відіграють важливу роль у процесі видобутку даних, то інформатики мають значні претензії щодо права власності на видобуток даних. Тим не менше, методи аналізу даних, як правило, мають статистичну базу; і статистики починають виявляти значний інтерес до цієї області, включаючи пропонування вищих курсів з видобутку статистичних даних. Подальші загальні читання щодо видобутку даних та статистики можна знайти в таких посиланнях, як Berry and Linoff (1997, 2000), Chatfield (1997), Friedman (1998), Hand (1999a, 1999b), Hastie et al. (2001).

Для ефективного використання наявних даних, отримуючи менш упереджену оцінку ефективності, ми застосовуємо процедуру, яка називається «N-кратна перехресна перевірка». Вибирається кількість складок, N (загальний вибір N дорівнює 10) для того, щоб розділити дані на N окремих підпроб за допомогою «складок», розділів набору даних. Для кожної з N моделей, що підлягають навчанню, для

тестування моделі зберігається одна підпроба, обмежена межами згину, тоді як решта N–1 підпроби використовуються для підготовки тієї самої моделі. Це дасть N неупереджених оцінок ефективності. Якщо оцінки ефективності, отримані з даних тренувань, високі, а оцінки тестування низькі, то модель переобладнала дані і не узагальнила добре.

Обидві технології можна розглядати як складові частини процесу підтримки ухвалення рішень. Проте ці технології як би рухаються у різних напрямах: OLAP зосереджує увагу виключно на забезпеченні доступу до багатовимірних даних, а методи Data Mining в більшості випадків працюють з плоскими одновимірними таблицями і реляційними даними.

Інтеграція технологій OLAP і Data Mining «збагатила» функціональність і однієї, і іншої технології. Ці два види аналізу повинні бути тісно з'єднано, щоб інтегрована технологія могла забезпечувати одночасно багатовимірний доступ і пошук закономірностей.

Література

1. Брайан Ларсон. Розробка бізнес-аналітики в Microsoft SQL Server 2005. – Санкт-Петербург: Пітер, 2008. – 688 с.

2. Каленік А. І. Використання нових функцій Microsoft SQL Server 2005. – Санкт-Петербург: Петро, 2006. – 334 с.

3. Барсегян А.А., Купріянов М.С., Степаненко В.В., Холод І.І. Методи та моделі аналізу даних: OLAP та Data Mining. – СПб: БХВ-Петербург, 2004. – 336 с.

4. Вступ до аналізу асоціативних правил – Доступно за посиланням: <http://www.basegroup.ru/library/analysis/association_rules/intro/>

5. Дата К. Дж. Вступ до систем баз даних, 8-е видання. – М.: Видавництво Вільямса, 2005. – 1328 с.

6. Основи баз даних: курс лекцій: підручник / С.Д. Кузнєцов. – Москва: Інтернет-університет Інформ. Технології, 2005. – 488с.

7. С. Я. Архипенко, Д. В. Голубєв, О. Б. Максименко. Сховища даних. Від концепції до реалізації. – М.: Діалог-МІФІ, 2002. – 528 с.

8. Брюс Еккель. Філософія JAVA, бібліотека програміста. 3-є вид. – СПб.: Питер, 2003. – 638 с.

9. Microsoft SQL Server 2005 Analysis Services.OLAP та багатовимірний аналіз даних. За редакцією А. Бергера, І. Горбача. – СПб.: БХВ-Петербург, 2007. – 908 с.

10. Maclennan J., Tang Z. Data Mining With SQL Server 2005. – Індіанаполіс.: Wiley, 2005. – 296 с.

11. Уллман Д., Вітер Д. Вступ до систем баз даних. – М.: Лорі, 2000. – 1328 с.

12. Гансен Г., Гансен Г. Бази даних: розробка та управління. – М.: БІНОМ, 1999. – 296 с.

Referencse

1. Berry, J. and Linoff, G. (1997), Data mining techniques-for marketing, sales and customer support, Williams Publishing House, New York: Wiley, USA.

2. Berry, J.A.M. and Linoff, G. (2000), Mastering data mining -the art and science of customer relationship management, Williams Publishing House, New York: Wiley, USA.

3. Chatfield, C. (1997), Royal statistical society news, Williams Publishing House, New York: Wiley, USA.

4. Friedman, J. (1998), Data mining and statistics-what's the connection, 29th Symposium on the interface, Williams Publishing House, Cape Town, South Africa.

5. Hand, D. (1998), Data mining-statistics and more, Williams Publishing House, New York: Wiley, USA.

6. Hand, D. (1999), Data mining-new challenges for statisticians, Williams Publishing House, New York: Wiley, USA.

7. Hand, D.J. (1999), Statistics and Data mining-intersecting disciplines, SIGKDD Explorations, New York: Wiley, USA.

8. Hastie, T. Tibshirani, R. and Friedman, J. (2001), Elements of statistical learning-data mining inference and prediction, Williams Publishing House, New York: Wiley, USA.

9. Kuonen, D. (2004), Data mining and Statistics: What is the connection, The Data Administrative Newsletter, Wiley, USA.

Надійшла / Paper received: 23.04.2020 Надрукована / Paper Printed : 04.06.2020