УДК 004.62
Є. С. ТИХОНОВ, аспірант,
Державний університет телекомунікацій, Київ

DATA MINING і ПРОБЛЕМА ВИКОРИСТАННЯ «БРУДНИХ ДАНИХ»

У сучасному світі інтелектуальний аналіз даних набув широкого визнання як потужний і універсальний інструмент аналізу даних не лише в інформаційних технологіях, а й у багатьох інших галузях, передусім у клінічній медицині, соціології, фізиці. Обчислювальний процес аналізу великих обсягів даних має на меті вилучення корисної інформації. У цій статті буде розглянуто методи боротьби з так званими брудними даними, які значно сповільнюють пошук цінних даних.

Ключові слова: отримання даних (Data Mining); аналітична обробка даних у реальному часі (OLAP); якість даних; пропущені значення; брудні дані.

Вступ

Data Mining поєднує в собі технології та засоби, використовувані з метою пошуку прихованих залежностей у різнорідних масивах даних. Дані — це результат фіксування деякої інформації. Самі дані, у свою чергу, можуть виступати як джерело нової корисної інформації. Засоби Data Mining дозволяють здобувати таку інформацію. По-справжньому цінна інформація має задовольняти такі вимоги [1]:

- бути досі невідома;
- містити нетривіальні відомості;
- мати практичну користь;
- бути доступна для інтерпретації.

Ці вимоги визначають сутність методів Data Mining. Потреба в таких методах очевидна, оскільки знання відповідних закономірностей дозволяє зменшити витрати на пошук цінної інформації. Саме тому застосування засобів Data Mining стає дедалі популярнішим серед великих компаній, зацікавлених у залученні нових клієнтів. Утім слід пам'ятати, що впровадження Data Mining на виробництві — занадто високовартісний і трудомісткий процес.

Основна частина

Згідно з класичним визначенням технології Data Mining ідеться про виявлення серед початкових («сировинних») даних певної сукупності раніше невідомих, нетривіальних, практично корисних і доступних для інтерпретації знань. Скажімо, показник середнього обсягу продажу деякого товару за певний період не відповідає вимогам щодо новизни та нетривіальності. Знання мають описувати нові зв'язки між досліджуваними властивостями, передбачати значення та характеристики одних параметрів на основі інших.

Головна мета застосування засобів і методів Data Mining полягає у виявленні прихованих правил і закономірностей у великих масивах даних. На відміну від аналітичної обробки даних у реальному часі (OLAP), найважливіше завдання Data Mining — формулювати гіпотези та виявляти незвичайні шаблони (закономірності), тобто йдеться про функції, перекладені з людини на комп'ютер.

Методи Data Mining дозволяють виявляти стандартні закономірності:

- асоціативні зв'язки подій (наприклад, при купівлі пива дуже часто купують і чіпси чи горішки);
- послідовність подій у часі (наприклад, нова квартира — нові меблі);
- кластеризацію подій (на відміну від класифікації групи заздалегідь не створюються), використовувану для сегментації ринку і замовників;
- прогнозування (базою слугує історична інформація), що спирається на побудову математичних моделей.

Існують два способи впровадження нової інформаційної технології в локальні інформаційні структури:

1) пристосування її до організаційної структури підприємства;

2) модернізування організаційної структури з метою найбільш ефективного використання нової інформаційної технології.

Перший спосіб дешевший і не вимагає великих змін в організації діяльності підприємства. Проте ефект від його впровадження може бути незначний. Другий спосіб вимагає більших капіталовкладень, але забезпечує якісно новий рівень діяльності підприємства чи організації.

Наведені далі приклади з різних галузей економіки розкривають головну перевагу методів Data Mining — здатність виявляти нові знання, які неможливо здобути методами статистичного регресивного аналізу або економетрії.

Приклад 1. Клієнтів компанії за допомогою одного з інструментів Data Mining було об'єднано в сегменти зі схожими ознаками. Це дозволило здійснювати гнучку маркетингову політику, будуючи окремі моделі поведінки для кожного сегмента. Найважливішими чинниками сигментації виступили віддаленість регіону клієнта від центрального офісу компанії, сфера діяльності, середньорічні суми операцій, кількість операцій за тиждень.

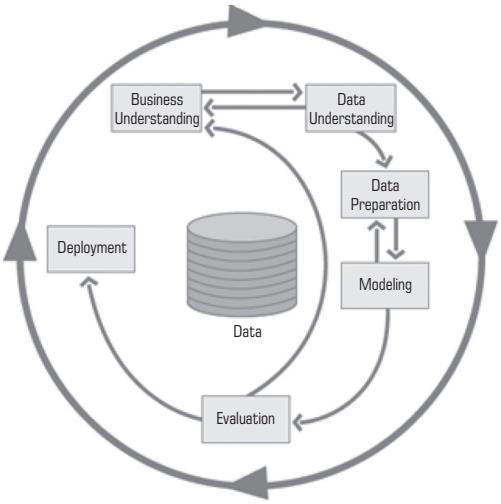
Приклад 2. Автоматичний аналіз банківської бази даних кредитних операцій фізичних осіб виявив правила, за якими позичальникам відмовляли у видачі кредиту. Вирішальними чинниками, виявились: термін кредиту, середньомісячний дохід і витрати позичальника. Надалі це враховувалося при експрес-кредитуванні.

Приклад 3. При аналізі бази даних клієнтів страхової компанії було встановлено соціальний портрет особи, що страхує життя. Це виявився чоловік 35–50 років, що має двох і більше дітей і середньомісячний дохід понад $2000.

Цікаві в тому чи іншому плані підгрупи є потужним і неодмінним компонентом інтелектуального аналізу даних, оскільки вони забезпечують інтерфейс між фактичними даними в базі даних і залежностями більш високого рівня, що описують дані. Деякі алгоритми інтелектуального аналізу даних призначено для виявлення таких цікавих підгруп. Проте цікаві підгрупи є обмеженими засобами здобуття знань про базу даних з огляду на те, що вони, за визначенням, описують тільки частину бази даних. Тому більшість алгоритмів розглядатимуть цікаві підгрупи не як кінцевий продукт, а як прості будівельні блоки для всеосяжного опису існуючих закономірностей. Структури, які є метою застосування зазначених алгоритмів, відомі як модель і процес огляду підгруп. При цьому побудову повної картини даних часто називають моделюванням.

Ми можемо розглядати базу даних як зібрання вихідних вимірювань щодо конкретного домена. Кожна людина є втіленням правил, що регулюють відповідну область. Модель, індукована з необробленими даними, є стислим поданням роботи домена, з абстрагуванням щодо відомостей про фізичних осіб.

Варто наголосити, що інтелектуальний аналіз даних часто застосовується для того, аби отримати прогностичні моделі (див. рисунок).



Міжгалузевий стандартний процес, характерний для інтелектуального аналізу даних

Існує чимало нетривіальних проблем, для розв'язання яких потрібен індивідуальний підхід до кожного конкретного завдання. Одна з таких проблем полягає в підтриманні високого рівня якості даних, на основі яких формуватиметься модель пошуку взаємозв'язків досліджуваних даних.

Якість даних (*Data Quality*) — це критерій, який визначає повноту, точність, своєчасність і можливість інтерпретації даних [2]. Дані можуть бути високої і низької якості. Останні — це так звані брудні, або погані, дані. Для підвищення якості даних їх очищають.

Під очищенням даних (*Data Cleaning, Data Cleansing або Data Scrubbing*) розуміється виявлення і вилучення помилок, невідповідностей і конфліктів серед даних. Очищення необхідне для поліпшення якості даних, що, у свою чергу, підвищує швидкість і якість аналізу даних методами Data Mining.

Можна виокремити чотири основні групи брудних даних:

1) брудні дані, які можуть бути автоматично виявлені і очищені;

2) дані, появу яких може бути припинено;

3) дані, непридатні для автоматичного виявлення і очищення;

4) дані, появі яких неможливо запобігти.

Важливо розуміти, що спеціальні засоби очищення можуть впоратися не з усіма видами брудних даних. До найбільш поширених видів брудних даних можна віднести [3]:

- дані з пропущеними значеннями атрибутів;
- суперечливі дані;
- дубльовані дані;
- шуми і викиди.

Пропущені значення (*Missing Values*) можуть виникнути з будь-якої причини. При цьому атрибуту об'єкта не присвоюється значення. Наприклад, при анкетуванні може бути не вказано вік. Так само деякі атрибути можуть бути незастосовні для об'єктів певного типу. Скажімо, немає сенсу вказувати атрибут «річний дохід» для дитини.

Прогнозування на основі таких даних здійснюється неякісно або зі значними обмеженнями. Цю проблему можна розв'язати кількома способами:

- виключити об'єкти з пропущеними значеннями з обробки;
- розрахувати нові значення для пропущених атрибутів;
- ігнорувати пропущені значення в процесі аналізу;
- замінити пропущені значення атрибутів на ймовірні значення.

Ще один вид брудних даних — ***суперечливі дані***. Суперечливість може виникнути, коли дані містяться в сховищі в незв'язаному вигляді. Це свідчить про те, що база даних погано спроек-

тована. Існує кілька варіантів поводження із суперечливими даними:

• об'єкти із суперечливими значеннями атрибутів вилучити з обробки;

• вибирати з множини можливих значень атрибутів єдине значення. Наприклад, обчислювати ймовірність появи кожної із взаємно суперечливих подій і вибирати найбільш імовірну з них.

Вилучення об'єктів — дуже грубий, але й найпростіший спосіб, що не вимагає застосування жодних додаткових алгоритмів. Другий спосіб хоча й складніший у реалізації, але більш правильний, бо дозволяє уникнути втрати даних.

Ще один вид брудних даних — ***дублікати*** (*Duplicate Data*). Дублікатами називаються записи з однаковими значеннями всіх атрибутів.

Дублікати іноді використовують, щоб штучно підвищити значущість певних записів, але здебільшого, наявність дублікатів негативно позначається на результатах аналізу. Тому перш ніж починати працювати з даними, потрібно обробити продубльовані записи.

Можливі два варіанти такої обробки.

• У першому випадку всі записи, які мають дублікати, вилучаються. Такий варіант використовується, коли наявність дублікатів повністю знецінює інформацію або викликає недовіру до неї.

• У другому випадку з групи продубльованих записів залишають тільки один.

Ідентифікація дублікатів так само становить проблему, оскільки значення деяких атрибутів потрібно вважати однаковими навіть тоді, коли вони не збігаються повністю. Прикладом може бути атрибут «Ім'я». Одне й те саме ім'я може бути записано в скороченій чи в повній формі, у вигляді ініціалів тощо.

Останній вид забруднених даних — це ***шуми*** і ***викиди***. Викидами називають об'єкти або спостереження, які різко виділяються з усього набору. Шумом називають сильні відхилення від середнього значення в наборі даних. Шум у даних не несе жодної корисної інформації, тому його намагаються мінімізувати. При аналізі даних шуми і викиди являють собою серйозну проблему, істотно знижуючи вірогідність результату аналізу. Викиди можуть бути поодинокі або включати в себе цілі групи об'єктів. Головне завдання аналітики — виявляти такі аномалії та оцінювати ступінь їх впливу на результати подальшого аналізу. Якщо викиди є інформативною частиною аналізованого набору даних, використовують робасні методи і процедури.

Добре зарекомендувала себе практика, коли аналіз здійснюється в два етапи.

На першому етапі виконується аналіз усього набору даних, включаючи викиди.

Другий етап — аналіз очищених від викидів даних.

Здобуті результати порівнюються і на їх основі доходять висновку про вплив викидів на кінцевий результат аналізу. Чутливість до викидів різна залежно від вибраних методів Data Mining. Цей факт необхідно враховувати при виборі методу аналізу даних. Процедури очищення від шумів можна знайти в багатьох сучасних інструментах Data Mining.

Висновки

Як уже зазначалося, наявність брудних даних у загальному їх наборі негативно позначається на результатах Data Mining. Такі результати можуть виявитися ненадійними і непотрібними. Утім наявність таких даних не завжди призводить до необхідності їх очищення або запобігання їх появі. Завжди має бути розумний вибір між наявністю брудних даних і витратами, необхідними для їх очищення.

Сьогодні проблема очищення даних вельми актуальна і інтерес до неї дедалі підвищується. Очищенням даних опікується чимало дослідницьких груп. Це й не дивно, адже від якості підготовлених даних безпосередньо залежить результат Data Mining. Ціна помилок може бути дуже висока, насамперед у фінансовому плані, оскільки впровадження засобів Data Mining на виробництві коштує недешево. І все ж успішний результат Data Mining може забезпечити великій компанії прибуток, що на порядок вищий за витрати.

Список використаної літератури

1. ***Григорьев, Л. И.*** *Научно-методические и технологические основы информационной системы управления качеством учебного процесса / Л. И. Григорьев.— М.: РГУ нефти и газа им. И. М. Губкина, 2008.— 132 с.*

2. ***Баргесян, А. А.*** *Методы и модели анализа данных: OLAP и Data Mining / [А. А. Баргесян, М. С. Куприянов, В. В. Степаненко, И. И. Холод].— М.: БХЗ-Петербург, 2009.— 336 с.*

3. ***Чубукова, И. А.*** *Data Mining: учеб. пособие.— [2-е изд., перераб.].— М.: Інтернет-университет информационных технологий, 2008.— С. 239–252.*

4. ***Aggarwal, C. C.*** *Data Clustering: Algorithms and Applications / C. C. Aggarwal, C. K. Reddy.— CRC Press, 2014.*

Рецензент: доктор техн. наук, професор **Л. Н. Беркман**, Державний університет телекомунікацій, Київ.

Е. С. Тихонов

DATA MINING И ПРОБЛЕМА ИСПОЛЬЗОВАНИЯ «ГРЯЗНЫХ ДАННЫХ»

В современной жизни интеллектуальный анализ данных получил широкое признание как мощный и универсальный инструмент анализа данных: не только в информационных технологиях, но и во многих других отраслях, прежде всего в клинической медицине, социологии, физике. Вычислительный процесс анализа больших объемов данных имеет целью извлечение полезной информации. В этой статье будут рассмотрены методы борьбы с грязными данными, которые значительно замедляют поиск ценных данных.

Ключевые слова: получение данных (Data Mining); аналитическая обработка в реальном времени (OLAP); качество данных; пропущенные значения; грязные данные.

Y. Tykhonov

DATA MINING AND USE PROBLEM «DIRTY DATA»

In modern life, Data Mining has been widely recognized as a powerful yet versatile analysis tools in various fields, not only in information technology but also clinical medicine, sociology, physics. Data calculated defined as a computational process of analyzing large amounts of data to extract useful information. This article will discuss methods of dealing with dirty data, because they significantly slow the search for valuable data.

Keywords: Data Mining; analytical processing in real time (OLAP); data quality; missing values; dirty data.

УДК 621.325.5:621.382.049.77

M. KOSOVETS,
L. TOVSTENKO,
Quantor scientific and production enterprise
Institute of cybernetics of V. Glushkov NAS of Ukraine

THE CONCEPT OF CREATION OF THE MODERN CLOUD COMPUTING ON THE BASIS THE DISTRIBUTED MULTIPROCESSOR OF REAL TIME

The solution of the organization of cloud computing on the basis of architecture of the multiprocessor of real time distributed in space is proposed, using perspective modules of processing and telecommunications. The attempt of convergence of system of telecommunications 5G, the Internet of things (IoT) and the multiprocessor distributed in space is made. This combining allows realizing all types of cloud computing, exchange and information representation at the level of perception by the person.

Keywords: cloud computing; multiprocessor of real time; convergence.

INTRODUCTION

Development of computer technique and telecommunications are intensified researches with development of computation on remote computers. Their use led to a concept of the virtual computing environment, and soon generated a successful brand — *cloud computing*. Now the mankind endure a boom on their creation and implementation.

Urgent there is a research of the technique of creation of cloud computing based on the information processing multiprocessor distributed in space in real time with the wireless front-side bus driver of exchange and open, flexible architecture allowing to increase a computing resource for today.

Each type of cloudy services and method of deployment provides the level of monitoring, flexibility and controllability. The «infrastructure as service» model includes in itself basic elements for creation of cloudy IT structure. In this model we get access to network resources, the virtual calculators and databases. The user has all advantages which provide clouds virtually. In case of desire the user can independently create clouds and provide paid access to their resources. It will provide the maximum protection against information leakage, low cost of support of service and complete freedom of creativity. SPC «Quantor» organized cloud computing in Italy with control from Ukraine. It is some kind of outsourcing of cloud.

Creation of cloud requires knowledge of parameters of information flows, redistribution of tasks between specialized coprocessors accelerators and the central multiprocessor system, technical characteristics of the processing modules, features of algorithmic support and program service.

The architecture of the calculator must maintain high parallelism and a tunable configuration that allows to solve problems of mathematical physics, digital filtering, image processing, electrodynamics and others (tasks of processing of multivariate signals). The special part is assigned to the communication environment, allowing to establish flexibly connection between processors.