УДК 004.032.26:004.048

А. Дорошенко
Національний університет "Львівська політехніка",
кафедра автоматизованих систем управління

АНАЛІЗ НЕЙРОМЕРЕЖНИХ МЕТОДІВ DATA MINING ЯК СКЛАДОВОЇ ТЕХНОЛОГІЇ BUSINESS INTELLIGENCE

Проаналізовано особливості бізнес-аналітики та місце у ній методів видобування даних. Визначено перспективи розвитку бізнес-аналітики, проведено порівняльний аналіз різних нейромережних методів видобування даних.

The article analyses the features of the business intelligence and considers Data Mining methods like a part of business intelligence. The future trends of the business intelligence are defined. The different basic methods of classification on the base of neural networks are compared.

Вступ

Сучасні умови ведення бізнесу, що характеризуються зростаючою твердою конкуренцією і нестабільністю економічних умов, висувають підвищені вимоги до оперативності і якості прийнятих рішень на всіх рівнях керування підприємством або організацією. Підтримка прийняття рішень передбачає володіння актуальною всеосяжною інформацією про стан і тенденції розвитку бізнесу. При цьому обсяг інформації, яку необхідно враховувати для формування оптимальних обґрунтованих рішень, неухильно зростає.

Це приводить до ситуації, коли стає неможливо ефективно керувати компанією без використання сучасних засобів інформаційного забезпечення, а саме, методів і засобів бізнес-аналітики (Business Intelligence (BI)) – технологій, що дають можливість організаціям перетворювати накопичені дані на інформацію про бізнес, а потім інформацію – на знання для керування бізнесом.

Мета роботи і постановка задачі

Метою роботи є аналіз існуючої ситуації у сфері business intelligence, визначення перспектив розвитку цього напрямку, а також визначення можливих шляхів підвищення їх ефективності, зокрема розглядається застосування з цією метою різних типів нейромережних методів data mining.

Аналіз останніх досліджень та публікацій

Хоча вперше термін "business intelligence" (BI) з'явився ще наприкінці 1980-х років завдяки аналітикам компанії Gartner та значно еволюціонував з того часу, досі немає єдиного загальноприйнятого визначення цього поняття. Однак сьогодні можні виділити два основні підходи до визначення бізнес-аналітики.

Перший з них розглядає бізнес-аналітику як методи, технології, засоби видобування і представлення знань. Відповідно до початкових визначень, BI – це процес аналізу інформації, вироблення інтуїції і розуміння для поліпшеного і неформального прийняття рішень бізнес-користувачами, а також інструменти для видобування з даних значимої для бізнесу інформації. Треба зазначити, що більшість визначень трактують "business intelligence" як *процес*, технології, методи і засоби видобування і представлення знань. Так, наприклад, Джонатан Ву у своїй статті "Business Intelligence: What is Business Intelligence?" визначає Business Intelligence як "процес

збирання багатоаспектної інформації про досліджуваний предмет. Розроблено програмні додатки, що забезпечують користувачів можливістю проводити такий процес для відповіді на питання бізнесу і для виявлення значимих тенденцій або шаблонів у досліджуваній інформації".

Другий підхід розглядає бізнес-аналітику як знання про бізнес і для бізнесу, тобто не як процес, а як *результат процесу* видобування знань – безпосередньо знання про бізнес для прийняття рішень.

Зокрема, наведемо одне з визначень, характерне для даного підходу: "Business Intelligence - знання, добуті про бізнес з використанням різних апаратно-програмних технологій. Такі технології дають можливість організаціям перетворювати дані на інформацію, а потім інформацію – на знання". Це визначення чітке розмежовує поняття "дані", "інформація" і "знання". Дані розуміються як реальність, що комп'ютер записує, зберігає й обробляє – це "сирі дані". Інформація – це те, що людина в стані зрозуміти про реальність, а знання – це те, що в бізнесі використовується для прийняття рішень. У процесі організації інформації для одержання знання часто застосовують сховища даних, а для представлення цього знання користувачам – інструменти бізнес-інтелекту.

Щороку кількість даних у світі подвоюється, але від цього мало користі, хоча їх можна перетворити на корисні інформацію і знання – інформація сама по собі не дуже підходить для прийняття рішень через її величезний обсяг. Засоби бізнес-інтелекту і сховищ даних покликані знаходити у величезній кількості даних і інформації те істотне, що реально додається до наших корисних знань. Вони не намагаються цілком замінити людину, а використовують для формування гіпотез інтуїцію, засновану на її підсвідомості й особистому досвіді.

Отже, бізнес-інтелект (business intelligence) у широкому змісті слова означає:

– процес перетворення даних на інформацію і знання про бізнес для підтримки прийняття поліпшених і неформальних рішень;

– інформаційні технології (методи і засоби) збирання даних, консолідації інформації і забезпечення доступу користувачів до знань;

– знання про бізнес, добуті в результаті заглибленого аналізу детальних даних і консолідованої інформації.

Характерні особливості Business intelligence

В основу технології BI покладено організацію доступу кінцевих користувачів і аналіз структурованих кількісних даних і інформації про бізнес. BI породжує ітераційний процес бізнес-користувача, що передбачає доступ до даних та їх аналіз, формування висновків, знаходження взаємозв'язків, щоб ефективно керувати підприємством. BI має широкий спектр користувачів на підприємстві, включаючи керівників і аналітиків.

Деякі схильні доволі широко трактувати BI, включаючи в це поняття і технологію керування знаннями (Knowledge Management), яка, однак, більш пов'язана з аналізом неструктурованої або слабоструктурованої інформації (наприклад, HTML), що не є предметом аналізу BI-інструментів [6]. Технологія керування знаннями забезпечує категоризацію, розвідку і семантичну обробку текстів, розширений пошук інформації тощо. Технологія BI має відношення до аналізу фактографічної структурованої (бази даних, плоскі файли й інші ODBC або OLE DB-джерела даних) і квазіструктурованої інформації (наприклад, XML). Щільні стики і перетини можливі при підготовці довідкової інформації для аналізу за допомогою розвідки (text mining) і очищення тексту, а також при розширенні пошуку інформації на аналітичні БД. Корпорації IBM і Microsoft реалізують стратегії інтеграції програмних засобів бізнес-інтелекту й інструментів керування знаннями, маючи на меті створення нового покоління ПЗ, що оброблятиме як структуровані, так і неструктуровані дані [7].

Технологія Business intelligence визначає методи і засоби доступу й оперативного аналізу інформації в термінах предметної області. BI-засоби переважно працюють в інфраструктурі сховищ даних, хоча це не є обов’язковим. Однак, враховуючи те, що концепція, методи і засоби сховищ даних визначають підходи і забезпечують інтеграцію, очищення, ретроспективне збереження

інформації, призначеної для аналізу у випадку взаємодії сховищ даних та методів BI, проблема очищення й узгодження даних покладається на них, причому здійснювати ці операції доведеться в режимі online. Крім того, є ефект впливу на продуктивність і надійність оперативної системи обробки трансакцій [8].

Класифікація продуктів Business intelligence

Сьогодні BI-продукти поділяються на BI-інструменти та BI-ужитки. Перші, своєю чергою, поділяються на: генератори запитів і звітів; розвинуті BI-інструменти, – насамперед інструменти оперативної аналітичної обробки (online analytical processing, OLAP). Головна частина BI-інструментів поділяється на корпоративні BI-набори (enterprise BI suites, EBIS) та BI-платформи. Засоби генерації запитів і звітів переважно поглинаються і заміщаються корпоративними BI-наборами. Багатовимірні OLAP-механізми або сервери, а також реляційні OLAP-механізми є BI-інструментами й інфраструктурою для BI-платформ. Більшість BI-інструментів застосовується кінцевими користувачами для доступу, аналізу і генерації звітів за даними, що найчастіше розташовуються в сховищі, вітринах даних або оперативних складах даних. Розробники ужитків використовують BI-платформи для створення і впровадження BI-ужитків, що не розглядаються як BI-інструменти. Прикладом BI-ужитків є інформаційна система керівника EIS [2,5].

В ужитки для бізнес-аналізу часто вбудовані BI-інструменти (OLAP, генератори запитів і звітів, засоби моделювання, статистичного аналізу, візуалізації і видобування даних (data mining)). BI-ужитки зазвичай орієнтовані на конкретну функцію організації або задачу, зокрема: аналіз і прогноз продажів, прогнозування, аналіз ризиків, аналіз тенденцій тощо. Вони можуть застосовуватися і ширше як ужитки керування ефективністю підприємства (enterprise performance management) або система збалансованих показників (balanced scorecard).

Тенденції розвитку Business intelligence

Серед BI-інструментів найінтенсивніше розвиваються EBIS, що є наслідком конкуренції, яка посилилася у сучасній економіці. Використання інструментів для генерації запитів і звітів, аналізу даних знижується, організації оновлюють їх і заміняють корпоративними BI-наборами. Основні інструменти (незаплановані запити, звітність і основний OLAP-аналіз) усе ще залишаються найпоширенішими, задовольняючи більшість потреб. Однак все частіше застосовуються прогресивніші BI-інструменти, подібні до технології data mining. Однак частка автономних інструментів data mining зменшується, ця технологія поглинається та інтегрується в інші BI-інструменти, наприклад, у розширення СУБД.

Очікується, що протягом 5 років такі можливості, як XML для аналізу (XML/A), BI Web-сервіси, спільна робота, безпровідні та мобільні комунікації об'єднаються у вигляді мереж бізнес-аналізу (BI networks), які будуть доповнені засобами моніторингу бізнес-діяльності (Business activity monitoring, BAM). Нова технологія BAM є, власне кажучи, операційним BI, що інтегрує ужитки реального часу з можливостями бізнесу-аналізу. Використовуючи транзакційні дані, витягнуті із систем обробки транзакцій у реальному часі, BI-інструменти аналізують ці дані і попереджають про критичні події, видаючи інформацію операційним користувачам, що приймають безпосередні рішення [10].

Перспективи розвитку Business intelligence в умовах економічного спаду

Сьогодні у всіх секторах інформаційних технологій спостерігається значне зменшення інвестицій. З початку економічної кризи, що розпочалась у вересні минулого року, галузеві експерти відзначають скорочення IT-бюджетів, особливо наголошуючи на зниження витрат у межах великих проектів, зокрема у галузі інфраструктур. Однак, незважаючи на це, компанії прагнуть упровадити нові інструменти інтеграції даних, які формально належать до компонентів інфраструктури, керування якістю даних і MDM.

Це можна пояснити тим, що ключовими компонентами, які стимулюють подальший розвиток і активне використання інтеграції даних, є: керування якістю даних, керування нормативно-довідковою інформацією, складна обробка подій та інші важливі для бізнесу компоненти. Клієнти впроваджують такі проекти, тому що сподіваються, що окупність (пов'язана з покращанням якості і підвищенням актуальності даних в операційних системах) буде достатньо високою. Незважаючи на скорочення бюджетів, витрати на інтеграцію даних не зменшуються [11].

Ще один напрямок IT, який не лише не постраждав від економічної кризи, але й залишається на підйомі – це розроблення аналітичних ужитків та аналітичних баз. Цей сегмент ринку особливо розвинувся протягом останніх трьох років, тому що компанії усе більше і більше змушені контролювати стрімкі зміни в бізнесі, а також шукати способи ефективного реагування на них.

Найкраще ці задачі вирішуються на аналітичних базах, де зібрані терабайти вихідних даних. Бізнес-аналітики виконують нерегламентовані запити (або досліджують дані за допомогою прогнозної аналітики), щоб визначити, як ці сутності змінилися і як на ці зміни треба реагувати. Саме в ті періоди, коли зміни відбуваються швидко, а старі стереотипи руйнуються, найкориснішою є технологія BI. BI та BPM-системи допоможуть дати відповіді на такі питання, як: які клієнти припинили купувати продукцію і чому; які замовники відмовилися від співробітництва; що можна зробити, щоб їх залучити; як вплине на дохід деяке зниження цін; які клієнти можуть виявитися найприбутковішими в умовах кризи; як можна залучити найбільшу кількість клієнтів тощо.

Застосування нейромережних методів видобування даних в системах BI

Видобування даних є процесом виявлення кореляції, тенденцій, шаблонів, зв'язків і категорій. Воно виконується шляхом ретельного дослідження даних з використанням технологій розпізнавання шаблонів, а також статистичних і математичних методів.

Під час видобування даних багаторазово виконуються різні операції і перетворення над сирими даними (добір ознак, стратифікація, кластеризація, візуалізація і регресія), що призначені для знаходження представлень, які є інтуїтивно зрозумілими для людей, які, своєю чергою, краще розуміють бізнес-процеси, покладені в основу їхньої діяльності та для знаходження моделей, що можуть прогнозувати результат або значення визначених ситуацій, використовуючи історичні або суб'єктивні дані.

На відміну від використання OLAP, видобування даних значно менше скеровується користувачем, замість цього покладається на спеціалізовані алгоритми, що встановлюють співвідношення інформації і допомагають розпізнати важливі (і раніше невідомі) тенденції, вільні від упередженості і припущень користувача.

Серед методів видобування даних широко відомими є пакети статистичного аналізу й аналіз часових рядів і оцінки ризиків; засоби моделювання; пакети нейронних мереж; засоби нечіткої логіки, експертні системи тощо. Важливим моментом є використання засобів для графічного оформлення результатів: засобу аналітичної картографії і топологічних карт; засоби візуалізації багатомірних даних тощо.

Згідно з описаними тенденціями розвитку BI можна зробити висновок, що сьогодні особливо актуальним є розроблення нових методів видобування даних, які повинні мати такі характеристики, як висока точність результату та висока швидкодія, можливість використання в системах реального часу.

Враховуючи саме ці параметри, порівняємо результати розв’язання типової задачі видобування даних різними нейромережевими методами, а саме: багатошаровим перцептроном, РБФ-мережею, імовірнісною нейтронною мережею, нейронною мережею з лінійною архітектурою. Архітектури нейронних мереж, що використовуються, наведено на рис.1.

Як одну з типових та важливих задач видобування даних розглядатимемо задачу класифікації клієнтів Інтернет-магазину на надійних клієнтів (з високою ймовірністю оплати замовлення) та

ненадійних покупців з метою оцінки ризиків втрати коштів. Розв'язання такої задачі можливе завдяки створенню в усіх online-магазинах інформаційних сховищ даних, в яких зберігається інформація про всі замовлення, зроблені клієнтами, та їх особові дані. За допомогою інтелектуального аналізу даних, які можуть мати дуже великий обсяг та бути різнорідними (якісними, кількісними, текстовими), ми можемо видобути приховані закономірності, залежності, отримати конкретні і зрозумілі результати (у нашому випадку – класифікацію всіх покупців на два попередньо відомі класи).

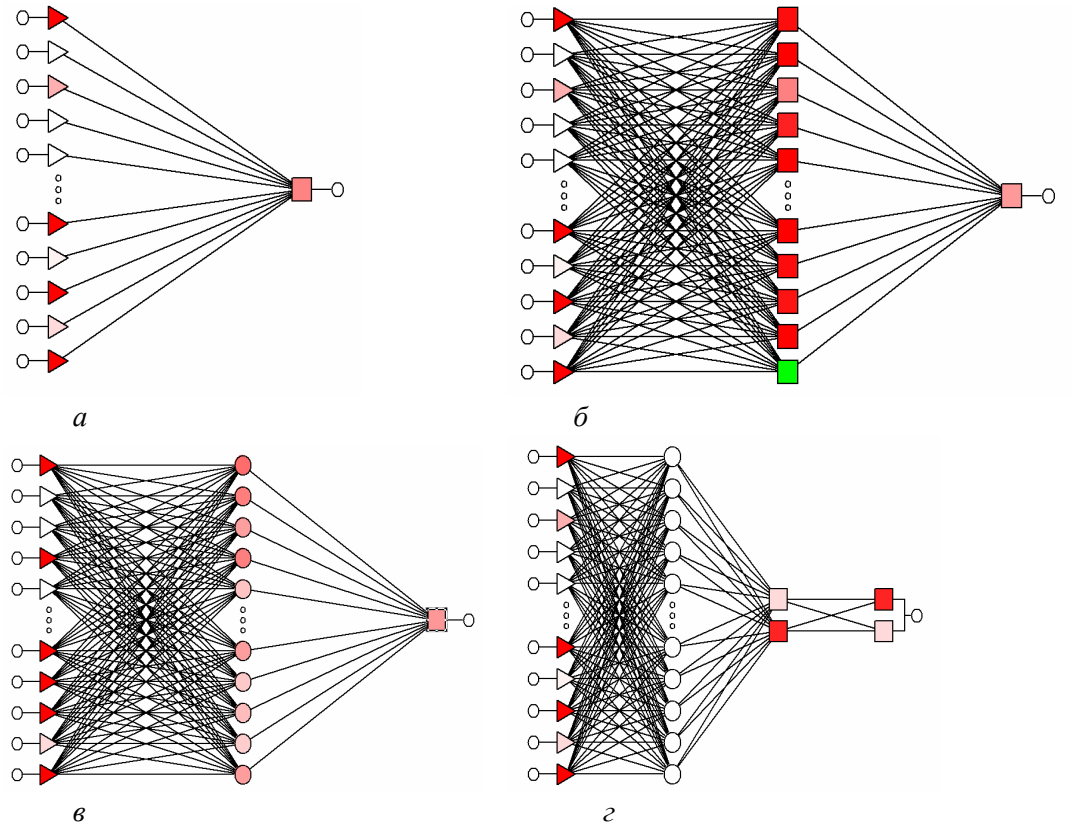

Рис.1. Архітектури нейронних мереж: а – лінійна архітектура 62:62-1:1; б – багатошаровий перцептрон 62:62-21-1:1; в – РБФ-мережа 38:38-51-1:1, точність; г – ймовірнісна нейронна мережа 62:62-1500-2-2:1

Тренувальна вибірка складається з даних про 30000 замовлень, в яких на основі спостережень за 4 тижні визначено їхню належність до одного з двох класів. Необхідно розробити таку систему, яка б дала змогу передбачати факт втрати оплати для подальших замовлень, які надходитимуть, та відносити їх до одного з двох класів відповідно до наведеної матриці вартостей. Розроблену систему тестуємо на виборці з 20000 замовлень, належність яких до одного з двох класів є невідомою, однак може бути перевіреною.

Кожне замовлення описують 44 ознаками, серед яких: ідентифікаційний номер замовлення, дата народження клієнта, наявність вказаних даних про телефон, e-mail, адресу для листування клієнта („так" чи „ні"), метод оплати замовлення, тип кредитної картки, вартість замовлення, час замовлення, кількість та артикули (коди) зроблених замовлень, інформація про те, чи було протягом наступних трьох днів зроблено інші замовлення з цієї IP-адреси, від цього самого клієнта та ін. Всі ознаки є різнорідними: числовими, текстовими, континуальними та векторними. Тому для коректної роботи з такими даними їх попередньо обробляють, у результаті чого всі дані перекодовують у числовий формат, кількість вхідних ознак при цьому збільшується до 62.

Результати класифікації на етапі тестування для всіх типів нейронних мереж, що порівнювались, наведено в табл. 1.

Таблиця 1

Результати класифікації

	Багатошаровий перцептрон		РБФ-мережа		Ймовірнісна нейронна мережа		Лінійна архітектура	
	Клас 1 (Nein)	**Клас 2 (Ja)**	**Клас 1 (Nein)**	**Клас 2 (Ja)**	**Клас 1 (Nein)**	**Клас 2 (Ja)**	**Клас 1 (Nein)**	**Клас 2 (Ja)**
Разом	18430	11570	18430	11570	18430	11570	18430	11570
Правильно	13680	8500	12830	7880	12600	7800	12820	7910
Неправильно	4750	3070	5600	3690	5830	3770	5610	3660
% правильно класифікованих	74,227	73,466	69,615	68,107	68,367	67,416	70,560	69,366
% неправильно класифікованих	25,773	26,534	30,385	31,893	31,633	32,584	29,440	30,634

Порівняємо отримані результати із результатами розв'язання цієї задачі за допомогою машини геометричних перетворень (МГП) з архітектурою зображеною на рис. 2 [14].

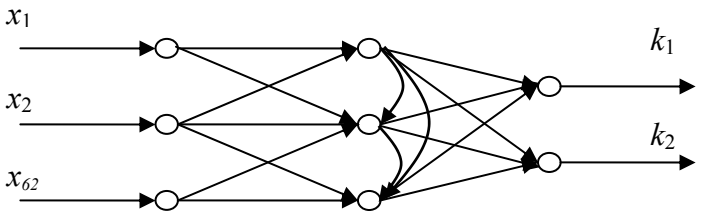

Рис. 2. Архітектура машини геометричних перетворень

Результати класифікації за допомогою машини геометричних перетворень, отримані на етапі тестування, наведено в табл. 2.

Аналіз отриманих результатів показав, що за практично однаково високої швидкодії всіх типів порівнюваних нейронних мереж (що дає змогу інтегрувати їх в системи реального часу), точність класифікації для цього типу задач є значно вищою у випадку використання машини геометричних перетворень. Також необхідно зазначити ще й таку перевагу МГП, як повторюваність результатів завдяки відсутності будь-якої імовірнісної складової у алгоритмі її навчання та функціонування.

Таблиця 2

Результати класифікації за допомогою МГП

	Машина геометричних перетворень	
	Клас 1 (Nein)	**Клас 2 (Ja)**
Разом	18430	11570
Правильно	16402	9950
Неправильно	2028	1620
% правильно класифікованих	88,996	85,538
% неправильно класифікованих	11,004	14,462

Висновки

Аналізуючи результати експериментів, можна зробити висновок, що, враховуючи тенденції розвитку BI серед методів видобування даних, які характеризуються високою точністю результату

та високою швидкодією, що дає змогу використовувати їх в системах реального часу, можна зазначити, що нейроподібні структури на основі машини геометричних перетворень для задач видобутку даних мають значну перевагу порівняно із такими типами нейронних мереж, як багатошаровий перцептрон, РБФ-мережа, імовірнісна нейронна мережа, нейронна мережа з лінійною архітектурою. Однак необхідно зазначити, що кожна конкретна задача видобування даних має свої особливості та вимагає ґрунтовного аналізу та, як правило, певної модифікації розроблених методів та алгоритмів.

1. Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. Базы данных. Интеллектуальная обработка информации. // М.: Нолидж, 2001. 2. Kimbal R. The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Willey&Sons, 1996. 3. Thomsen E. OLAP Solutions: Building Multidimensional Information Systems. Wiley Computer Publishing, 1997. 4. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, реализация. Том.1: Пер. с англ. – М.: Вильямс, 2001. 5. Архипенков С., Голубев Д., Максименко О. Хранилища данных. От концепции до внедрения/ Под общ. Ред. С.Я. Архипенкові – М.: ДИАЛОГ-МИФИ, 2002. 6. Liautaud B., Hammond M. e-Business Intelligence: Turning Information into Knoledge into Profit. McGraw-Hill, 2001. 7. Business Intelligence: Data Mining and Optimization for Decision Making, Carlo Vercellis, Wiley, 2008. 8. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Galit Shmueli, Nitin R. Patel, Peter C. Bruce, Wiley, 2008. 9. Интеграция данных. «Полный вперед», несмотря на кризис (Data Integration: Full Steam Ahead Despite Weak Economy), Стивен Суойер (Stephen Swoyer), апрель 2009; 10. Экономика не вмешалась в планы интеграции данных (Economy Hasn't Dampened Data-Integration Plans), Лорани Лоусон (Loraine Lawson), березень 2009 г. 11. TDWI's Best of Business Intelligence, Vol. 6), The Economy and BI. Two perspectives, Philip Russom, Wayne Eckerson, 2008 г. 12. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с. 13. Дюк В., Самойленко А. Data Mining: учебный курс. – СПб: Питер, 2001. – 386 с.14. Ткаченко Р. О. Нейромережні компоненти систем технічного зору // Інформаційні технології і системи. – 2005. – Т. 8 - №1. – С. 86–89. 15. Хайкин С. Нейронные сети: полный курс//Пер с англ. – 2-е изд. – М.: Вильямс, 2006. – 1104 с. 16. Donghui Li; Azimi-Sadjadi, M.R.; Robinson, M. Comparison of different classification algorithms for underwater target discrimination // IEEE Transactions on neural networks. – 2004.– Vol.15, № 1.– P.189–194.