УДК 004.89

Н.С. Лєсна, К.К. Дікарєва

*Харківський національний університет радіоелектроніки, Харків*

# МЕТОДИ І МОДЕЛІ DATA MINING В ПОБУДОВІ ОЦІНОЧНОЇ МОДЕЛІ ПЕРЕВАГ КОРИСТУВАЧА

*Основна мета цієї роботи присвячена вивченню та аналізу існуючих методів data mining. Докладний огляд методу дерев рішень, який дозволяє аналізувати та обробляти дані і будувати прогнози. Значними перевагами методу дерева рішення є те, що він здатний працювати як з категоріальними, так і з інтервальними змінними, не вимагає підготовки даних, дозволяє працювати з великим об'ємом інформації без спеціальних підготовчих процедур.*

***Ключові слова:*** *Data mining, дерева рішень, дані, вузол, зберігання даних, прогнозування, алгоритм, набір даних, вектор, метод, модель.*

## Вступ

В умовах сучасної конкурентної економіки критичне значення для підприємств набуває управління їх найбільш цінними активами – клієнтами та інформацією про них. У зв'язку з вдосконаленням технологій запису і зберігання даних на суспільство обрушилася колосальні потоки інформації в самих різних областях, що вимагає продуктивної переробки [1, 2].

За останні роки, коли, прагнучи до підвищення ефективності і прибутковості бізнесу, при створенні БД всі стали користуватися засобами обробки цифрової інформації, з'явився і побічний продукт цієї активності - гори зібраних данних. У широкому розумінні дані є факти, текст, графіки, картинки, звуки, аналогові або цифрові відео-сегменти. Іншими словами, дані – це необроблений матеріал, що надається постачальниками даних і використовуваний споживачами для формування інформації на основі даних. Корпоративна база даних будь-якого підприємства в сучасному світі містить набір таблиць, що зберігають записи про об'єкти або факти (наприклад, про рахунки, клієнтах, продажах). Як правило, кожен запис в подібній таблиці описує якийсь конкретний факт або об'єкт. Наприклад, запис у таблиці продажів відображає той факт, що якийсь товар проданий такому-клієнтові тоді-то, і, за великим рахунком, нічого, крім цих відомостей, не містить. Однак сукупність великої кількості таких записів, накопичених за кілька років, може стати джерелом додаткової, набагато більш цінної інформації, яку неможливо одержати на основі одного конкретного запису, а саме, відомостей про закономірності, тенденції або взаємозалежностях між даними. Справа в тому, що обсяги даних настільки значні, що людині просто не під силу проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих "сирих даних" укладені знання, які можуть бути використані при

прийнятті рішень. Старі методи, що застосовувалися математиками і статистиками, забирали багато часу, щоб в результаті отримати конструктивну та корисну інформацію. Для того щоб провести автоматичний аналіз даних, використовується Data Mining. Data Mining – це процес виявлення в сирих даних раніше невідомих нетривіальних практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах людської діяльності.

**Мета статті** – проведення аналізу існуючих методів data mining

## Основний розділ

Традиційні методи аналізу даних в основному орієнтовані на перевірку заздалегідь сформульованих гіпотез (verification-driven data mining) і на "грубий" розвідувальний аналіз, що становить основу оперативної аналітичної обробки даних, в той час як одне з основних положень Data Mining – пошук неочевидних закономірностей. Інструменти Data Mining можуть знаходити такі закономірності самостійно і також самостійно будувати гіпотези про взаємозв'язки. Оскільки саме формулювання гіпотези щодо залежностей є найскладнішим завданням, перевага Data Mining в порівнянні з іншими методами аналізу є очевидним. Data Mining допоможе проаналізувати величезні масиви і знайти приховану, але цінну інформацію, яка може допомогти вам краще зрозуміти своїх клієнтів і передбачити їх поведінку [3 – 5].

Озброївшись цією цінною інформацією, можна вибудувати більш близькі відносини з клієнтами, зрозуміти їх, що дозволить:

– краще утримувати клієнтів і уникнути непотрібних дій;

– складати профілі клієнтів і розуміти їх поведінку;

– підтримувати і підвищувати рівень прибутку;

– скоротити витрати клієнтів при покупках;

– виходити на клієнтів з потрібними їм пропозиціями.

Одним із завдань, яке вирішується методами Data Mining є прогнозування (прогноз трендів, динаміка попиту).

Прогнозування є важливим елементом організації управління як окремими господарюючими суб'єктами, так і економіки в цілому.

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

Завдання прогнозування, мабуть, може вважатися однією з найбільш складних задач Data Mining, вона вимагає ретельного дослідження вихідного набору даних і методів, придатних для аналізу.

Метою прогнозування є передбачення майбутніх подій. Прогнозування спрямоване на визначення тенденцій динаміки конкретного об'єкта або події на основі ретроспективних даних, тобто аналізу його стану в минулому і сьогоденні.

Застосування data mining сприяє стрімкому розвитку електронного бізнесу. Здатність визначати інтереси і переваги кожного відвідувача, спостерігаючи за його поведінкою, є серйозною і критичною перевагою конкурентної боротьби на ринку електронної комерції. За допомогою технології Data Mining аналізується така інформація:

– які сторінки переглядав користувач;

– яка послідовність перегляду сторінок.

Також аналізується, які групи користувачів можна виділити серед загального їх числа на основі історії перегляду Web-вузла.

Дані повинні мати певний достатній обсяг, щоб в них можна було виявити закономірності, але при цьому не занадто великий, щоб аналіз можна було провести порівняно швидко. Зазвичай вихідні дані беруться зі сховищ даних. При використанні Data Mining виникає два типи задач. Перша стосується збору даних, друга – використання методів персоніфікації. В результаті збору деякого об'єму персоніфікованих ретроспективних даних про конкретного клієнта, система накопичує певні знання про нього і може рекомендувати йому, наприклад, певні набори товарів або послуг.

Задачі DataMining можна підрозділити на такі категорії: попередня обробка даних; виявлення шаблонів і відкриття знань з використанням асоціативних правил, часових послідовностей, класифікації та кластеризації; аналіз отриманого знання.

Дерева рішень – один з методів автоматичного аналізу даних. Метод дерев рішень (decision trees) є одним з найбільш популярних методів вирішення завдань класифікації та прогнозування. Іноді цей метод Data Mining також називають деревами вирішальних правил, деревами класифікації та регресії.

Дерева рішень – це спосіб представлення правил в ієрархічній, послідовної структурі, де кожному об'єкту відповідає єдиний вузол, що дає рішення.

Під правилом розуміється логічна конструкція, представлена у вигляді «якщо ... то ...».

Завданням дерева рішень – відповісти на запитання. Для його вирішення потрібно відповісти на ряд питань, які знаходяться у вузлах цього дерева, починаючи з його кореня.

В інтелектуальному аналізі даних, дерева рішень можуть бути використані в якості математичних і обчислювальних методів, щоб допомогти описати, класифікувати й узагальнити набір даних, які можуть бути записані таким чином:

$$(x, Y) = (x_1, x_2, x_3 ... x_k, Y) .$$

Залежна змінна $Y$ є цільової змінної, яку необхідно проаналізувати, класифікувати й узагальнити. Вектор x складається з вхідних змінних $x_1, x_2, x_3$ і т.д., які використовуються для виконання цього завдання.

Дерева рішень, використовувані в Data Mining, бувають двох основних типів:

– аналіз дерева класифікації, коли пророкував результати є класом, до якого належать дані;

– регресійний аналіз дерева, коли передбачений результат можна розглядати як дійсне число (наприклад, ціна на будинок, або тривалість перебування пацієнта в лікарні).

Серед інших методів Data Mining, метод дерева прийняття рішень має кілька переваг. Класифікаційна модель, представлена у вигляді дерева рішень, є інтуїтивною і спрощує розуміння розв'язуваної задачі. Результат роботи алгоритмів конструювання дерев рішень, на відміну, наприклад, від нейронних мереж, що представляють собою "чорні ящики", легко інтерпретується користувачем. Це властивість дерев рішень не тільки важливо при віднесенні до певного класу нового об'єкта, але й корисно при інтерпретації моделі класифікації в цілому. Дерево рішень дозволяє зрозуміти і пояснити, чому конкретний об'єкт належить до того чи іншого класу.

Метод добре працює навіть в тому випадку, якщо були порушені початкові припущення, включені в модель. Дозволяє працювати з великим об'ємом інформації без спеціальних підготовчих процедур. Даний метод не вимагає спеціального обладнання для роботи з великими базами даних.

Дерева рішень дають можливість витягати правила з бази даних на природній мові. Приклад правила: Якщо Вік> 35 і Дохід> 200, то видати кредит.

Дерева рішень дозволяють створювати класифікаційні моделі в тих областях, де аналітику досить складно формалізувати знання.

Алгоритм конструювання дерева рішень не вимагає від користувача вибору вхідних атрибутів (незалежних змінних). На вхід алгоритму можна подавати всі існуючі атрибути, алгоритм сам вибере найбільш значущі серед них, і тільки вони будуть використані для побудови дерева. У порівнянні, наприклад, з нейронними мережами, це значно полегшує користувачеві роботу, оскільки в нейронних мережах вибір кількості вхідних атрибутів істотно впливає на час навчання.

Точність моделей, створених за допомогою дерев рішень, порівнянна з іншими методами побудови класифікаційних моделей.

Розроблено ряд масштабованих алгоритмів, які можуть бути використані для побудови дерев рішення на надвеликих базах даних; масштабованість тут означає, що із зростанням кількості прикладів або записів бази даних час, що витрачається на навчання, тобто побудова дерев рішень, зростає лінійно.

Приклади таких алгоритмів: SLIQ, SPRINT.

На побудову класифікаційних моделей за допомогою алгоритмів конструювання дерев рішень потрібно значно менше часу, ніж, наприклад, на навчання нейронних мереж.

Більшість алгоритмів конструювання дерев рішень мають можливість спеціальної обробки пропущених значень.

Багато класичних статистичних методів, за допомогою яких вирішуються завдання класифікації, можуть працювати тільки з числовими даними, в той час як дерева рішень працюють і з числовими, і з категоріальними типами даних.

Багато статистичних методів є параметричними, і користувач повинен заздалегідь володіти певною інформацією, наприклад, знати вид моделі, мати гіпотезу про вид залежності між змінними, припускати, який вид розподілу мають дані. Дерева рішень, на відміну від таких методів, будують непараметричні моделі. Таким чином, дерева рішень здатні вирішувати такі завдання Data Mining, в яких відсутня апріорна інформація про вид залежності між досліджуваними даними.

## Висновки

Розглянуто Data Minig, необхідність його використання в сьогоденні для виявлення нових, раніше невідомих знань, які в подальшому можна буде використовувати на практиці, та один з методів data minig – дерева рішень, який використовується для аналізу даних для прогнозних моделей.

Data Mining (видобуток даних, інтелектуальний аналіз даних, глибинний аналіз даних) використовується для позначення сукупності методів виявлення в даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності. Значними перевагами методу дерева рішення є те, що він простий в розумінні та інтерпретації, не вимагає підготовки даних, здатний працювати як з категоріальними, так і з інтервальними змінними, є надійним методом та дозволяє працювати з великим об'ємом інформації без спеціальних підготовчих процедур, не вимагає спеціального обладнання для роботи з великими базами даних.

## Список літератури

*1. Nisbet R., Elder J., Miner G. Handbook of Statistical Analysis and Data Mining Applications 2009. – 736 p.*

*2. Michael R. Berthold. Intelligent Data Analysis. Publisher: Springer; 2nd edition, 2007. – 525 p.*

*3. Michael J. A. Berry, Gordon S. Linoff. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Publisher: *Wiley Computer Publishing; 2nd edition, 2004. – 672 p.*

*4. David J. Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining (Adaptive Computation and Machine Learning). Publisher: A Bradford Book, 2001. – 584 p.*

*5. Дюк В. Data Mining: учебный курс / В .Дюк, А. Самойленко. – СПб.: Изд. Питер, 2001. – 368 с.*

*Надійшла до редколегії 26.02.2013*

**Рецензент:** канд фіз.-мат. наук, ст. наук. співр. О.О. Можаєв, Національний технічний університет «ХПІ», Харків.

## МЕТОДЫ И МОДЕЛИ DATA MINING В ПОСТРОЕНИИ ОЦЕНОЧНОЙ МОДЕЛИ ПРЕДПОЧТЕНИЙ ПОЛЬЗОВАТЕЛЯ

Н.С. Лесная, К.К. Дикарева

*Основная цель этой работы посвящена изучению и анализу существующих методов data mining. Подробный обзор метода деревьев решений, который позволяет анализировать и обрабатывать данные и строить прогнозы. Значительными преимуществами метода дерева решения является то, что он способен работать как с категориальными, так и с интервальными переменными, не требует подготовки данных, позволяет работать с большим объемом информации без специальных подготовительных процедур.*

***Ключевые слова:*** *Data mining, деревья решений, данные, узел, хранения данные, прогнозирования, алгоритм, набор данных, вектор, метод, модель.*

## METHODS AND MODELS OF DATA MINING IN BUILDING ESTIMATION MODELS OF USER PREFERENCES

N.S. Lesna, K.K. Dikarieva

*The main objective of this work is devoted to the study and analysis of existing methods of data mining. A detailed review of the method of decision tree that allows you to analyze and process the data and make predictions. The significant advantage of the method of decision tree is that it can work both with categorical, or interval variables, doesn't require the preparation of data, you can work with a large amount of information without any special preparation procedures..*

***Keywords:*** *Data mining, trees of decisions, information, knot, storages are information, prognostications, algorithm, set of data, vector, method, model.*