7. *Лукичев М.* Особенности развития корпоративных информационных систем в нефтяных компаниях. // Финансовая газета. — 2000. — № 24.

8. *Сопко В. Завгородній В.* Організація бухгалтерського обліку, економічного контролю та аналізу: Підручник. — К.: КНЕУ, 2000. — 260 с.

9. *Терехов А. А., Терехов М. А.* Контроль и аудит: основные методические приемы и технология. — М.: Финансы и статистика. — 1998. — 208 с: ил.

10. *Шуремов Е. Л.* Корпоративные системы: характерные черты и особенности построения // Аудиторские ведомости. — 1998. — № 3. — С. 80—84.

Стаття надійшла до редакції 18.04.06

УДК 330.115:681.513.2

В. Ф. Ситник, д-р екон. наук, проф.
Н. В.Ситник, канд. екон. наук, доц.

ДЕРЕВА РІШЕНЬ В СИСТЕМАХ ДЕЙТАМАЙНІНГУ

В статті розглянуто питання використання алгоритмів дерев рішень у системах дейтамайнінгу для виконання інтелектуального аналізу даних. Розкрита сутність методології дерев рішень і її використання при прийнятті бізнесових рішень. Показані шляхи побудови дерев класифікації. Визначені критерії оцінки якості дерев рішень. Наводяться приклади найрозповсюдженіших пакетів прикладних програм, в яких реалізовані різні алгоритми дерев рішень. Докладніше розглянуто алгоритм побудови дерев рішень ID3 і на прикладі проблеми оцінки кредитного ризику показана можливість практичного його використання.

КЛЮЧОВІ СЛОВА: дейтамайнінг, дерево рішень.

Дейтамайнінг, або по-іншому — інтелектуальний аналіз даних, у даний час широко застосовується при прийнятті рішень у бізнесі [1; 2]. Він включає багато різноманітних аналітичних методологій, зокрема і дерева рішень. Дерева рішень (decision trees) є одним з найпопулярніших підходів до рішення задач дейтамайнінгу. Наприклад, у програмному продукті дейтамайнінгу KnowledgeSTUDIO пропонується п’ять алгоритмів дерев рішень. Система дерева рішень просто ділить таблицю для аналізу даних у менші таблиці за допомогою вибору підмножин, основаних на значеннях для даного атрибута. Дерева рішень відображають надзвичайно багато критичних особливостей дейтамайнінгу, тому во-

ни можуть використовуватися в широкому діапазоні ділових проблем як для дослідження, так і для прогнозування.

В загальному виді *дерево рішень* — це схема (граф), яка відображає структуру задачі багатокрокового процесу прийняття рішень у вибраній сфері аналізу (класифікація набору даних або зразків, пошук оптимального рішення на множині альтернатив, структурізація проблеми, отримання логічного висновку за допомогою аналізу евристики (бази правил)). Свою назву «дерево» цей метод отримав від того, що конфігурація схеми дерева нагадує обрис крони дерева реального ландафту. Правда, на відміну від реальності корневий сегмент дерева рішень знаходиться не внизу (як корінь дерева), а наверху.

Звичайне дерево складається з кореня, гілок, вузлів (місць розгалуження), листя. Точно так дерево рішень складається з вузлів (званих також вершинами), що часто позначаються колами; гілок, що позначаються відрізками, які сполучають вузли. Для зручності дерево рішень зображають звичайно зверху вниз або зліва направо. Найперша (верхня або ліва) вершина називається коренем. Ланцюжок «корінь — гілка — вершина — ... — вершина» закінчується вершиною, яку називають «листом». З кожної внутрішньої вершини (тобто не листа) може виходити дві або більш гілок. Кожному такому вузлу зіставлена деяка характеристика, а гілкам — області значення цієї характеристики, причому ці області дають розбиття множини значень даної характеристики. У випадку, якщо з кожної внутрішньої вершини виходить рівно дві гілки (дерево такого типа називається *дихотомічним* або *бінарним*), кожній гілці можна зіставити істинність або помилковість деякого твердження щодо даної характеристики.

Якщо йдеться про дослідження проблеми, то гілки дерева відображають різні події, які можуть мати місце, а вершини — стани, в яких виникає необхідність вибору. Для проблем класифікації кожен перехід (гілка) дерева -- питання класифікації, і листя дерева — розділення набору даних з їх класифікацією.

Дерева рішень як метод аналізу і підтримки створення рішень ув різних галузях людської діяльності знайшли широке розповсюдження. Стисло зупинимося на найхарактерніших застосувань алгоритмів дерев рішень.

Дерево рішень як метод сегментації з певним наміром. Для оцінки перспектив розвитку бізнесу дерево рішень може розглядатися як засіб створення сегментації оригінального набору даних, де кожен сегмент був би одним з листків дерева. Сегментація може стосуватися клієнтів, продуктів чи комерційних областей. Вона

виконується з чітким наміром — для прогнозу деякої важливої частини інформації. Записи, які попадають в межі кожного сегменту, мають схожість щодо передбачуваної інформації та вміщують опис характеристик, які і визначають пророчий сегмент.

Застосування дерев рішень до бізнесу. Завдяки своїй деревовидній структурі і здатності легко генерувати правила прийняття рішень дерева рішень — це загально визнаний метод формування легко зрозумілих моделей, що формують наглядну схему бізнес-процесів. На довершення до пророчої моделі завдяки цій ясності моделі дозволяють проводить більш глибоке дослідження альтернатив, застосовуючи різноманітні фінансові показники (наприклад, ROI, NPV чи інші). Крім того, високий рівень автоматизації розрахунків і досить простий переклад моделей дерева рішень на мову SQL для розгортання в реляційних базах даних дозволяють легко інтегрувати процес створення і аналізу дерев рішень з діючими процесами інформаційних технологій в бізнесі.

Використання дерев рішень для дослідження. Технологія дерева рішень може використовуватися для дослідження набору даних і бізнесових проблем. Це часто робиться за допомогою перегляду провісників і значень, які вибрані для кожного розбиття дерева. Часто відслідковування цих провісників забезпечує отримання придатних для використання інтуїтивних міркувань або пропонує питання, на які потрібно відповісти.

Використання дерев рішень для попередньої обробки даних. Технологія дерев рішень може бути використана для попередньої обробки даних з наміром використання отриманих результатів в інших алгоритмах прогнозування (нейромережі, метод найближчих сусідів, стандартні статистичні підпрограми).

Дерева рішення для прогнозування. Хоча деякі форми дерев рішень спочатку розвивалися як дослідницькі інструменти, щоб очистити і наперед обробити дані для більш стандартних статистичних методів подібно логістичній регресії, вони також використовуються все більше і все частіше для прогнозування.

Наразі є багато типів дерев рішень та алгоритмів їх побудови, і для них розроблене відповідне програмне забезпечення. Кожен з цих типів має свої специфічні правила і вимоги стосовно формування графічної схеми. Проте є деякі загальні характеристики, притаманні більшості алгоритмів побудови дерев рішень.

Основний крок у процесі використання методу дерев рішень — це побудова або нарощування самого дерева. Процес нарощування дерева знаходиться в пошуку кращого можливого питання, що ставиться в кожному пункті дерева, де відбувається відгалу-

жується. Іншими словами, щоб вирішити, до якого класу віднести деякий об'єкт або ситуацію, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня.

Різниця між хорошим питанням і поганим питанням (критерієм розщеплення) має відношення до того, наскільки питання можуть добре організувати дані. Існує велика кількість варіантів алгоритмів, що будують дерева рішень. Деякі алгоритми дерева рішень використовують евристики для того, щоб вибрати питання або навіть вибирають ці питання випадково. Наприклад, алгоритм дерева рішень CART вибирає питання у дуже простий спосіб: він пробує їх всі. Багато з алгоритмів дерев рішень використовує рекурсивне застосування деякої процедури розщеплювання спочатку до всієї множини навчальних прикладів, а потім до їх підмножин, що одержуються в результаті розщеплювання початкового набору записів.

Вибір атрибуту і критерію, по яких проводиться розщеплювання множини, вирішення ситуації, коли не ясно, до якого класу віднести деякий запис, є принциповими моментами, що розрізняють використовувані алгоритми побудови дерев рішень. Для умов, коли набір класифікованих випадків описується неточними даними, перспективним напрямком є використання нечітких критеріїв розщеплювання. Замість певного віднесення кожного запису, що відноситься до вузла, до одного з його нащадків, визначається безперервна (в термінах нечітких множин) міра належності запису до різних підгруп, так що в результаті один і той же запис може належати до різного листя, але з різним ступенем упевненості.

Більшість алгоритмів дерева рішень зупиняють процес нарощування дерева, коли зустрінеться один з трьох критеріїв: сегмент містить тільки один запис; всі записи в сегменті мають ідентичні характеристики; удосконалення не досить істотне, щоб гарантувати створення розбиття

На даний час досить велике число продавців пропонують пакети програмного забезпечення, які основуються на методах дерева рішень як, наприклад, CART. Сюди входять американські корпорації IBM, Pilot Software, Business Objects, Cognos, NeoVista, SAS, Angoss і Integral Solutions (ISL) та інші. Більшість цих систем дозволяє інтерактивне дослідження даних з деревами рішень. Однак, у той час, як основна технологія більшості цих систем подібна, їх виконання відрізняється в реалізації інтерфейсу користувача і легкості використання. Деякі системи мають набагато краще діалогове середовища, ніж інші. Самими поширеними програмними продуктами дейтамайнінгу, що основуються на деревах рішень, є See5/C5.0 (RuleQuest, Австралія), Clementine (Integral Solutions, Великобрита-

нія), SIPINA (University ofLyon, Франція), IDIS (Information Discovery, США). Вартість цих систем варіюється від 1 до 10 тис. дол.

Дерева рішень у дейтамайнінгу застосовуються головно для розв'язування проблем класифікації і регресії [3—5]. Проблеми типу *класифікації* — загалом це ті, де намагаються передбачати значення категорійної залежної змінної (клас, членство в групі тощо) від однієї або більше безперервних та/або категорійних незалежних змінних — предикторів.

В проблемах типу *регресії* ставиться завдання передбачати значення безперервної змінної від однієї або більше безперервних та/або категорійних змінних (предикторів). Розв'язування проблеми регресії вимагає розбиття діапазону значень залежної змінної на декілька інтервалів, в які врешті попадають предмети класифікації. Очевидно, що точність передбачення залежить від числа таких інтервалів: чим більше число інтервалів, тим точніше передбачення, але при цьому зростає обсяг необхідних розрахунків.

Слід зауважити, що більшість алгоритмів побудови і дослідження дерев класифікації успішно справляються з двома означеними вище типами проблем, проте в деяких назвах алгоритів дерев рішень озвучуються ці проблеми. Наприклад, CART (C&RT) є абревіатура від англомовного виразу Classification and Regression Trees, що перекладається як дерева класифікації і регресії.

Дерева класифікації можуть бути, іноді й бувають, дуже складними. Проте використання спеціальних графічних процедур дозволяє спростити інтерпретацію результатів навіть для дуже складних дерев. Можливість графічного представлення результатів і простота інтерпретації багато в чому пояснюють велику популярність дерев класифікації у прикладних областях, проте найважливішими відмітними властивостями дерев класифікації є їх *ієрархічність* і *гнучкість*.

Процедуру формування класифікаційних дерев рішень за статистичними даними прийнято також називати побудовою дерева. Для кожного конкретного завдання статистичного аналізу існує велике число (часто навіть нескінченно багато) різних варіантів дерев рішень. Виникає питання: яке саме дерево краще і як його знайти. Щоб відповісти на першу частину питання, розглянемо різні способи визначення показників, що характеризують якість дерева.

Можна виділити два основні види показників, що характеризують якість дерева: показники точності та показники складності дерева.

Показники точності дерева визначаються за допомогою вибірки і характеризують те, наскільки добре розділені об'єкти

різних класів (у разі задачі розпізнавання), або те, наскільки велика погрішність прогнозування (у разі задачі регресійного аналізу).

Показники *складності дерева* характеризують його форму безвідносно до вибірки. До цих показників відносяться число листя дерева, число його внутрішніх вершин, максимальна довжина шляху з кореня в кінцеву вершину.

Показники складності і точності взаємозв'язані: чим складніше дерево, тим воно, як правило, точніше (якщо розглянути дерево, в якому кожному листу відповідає один об'єкт, то точність буде максимальною). Менш складне дерево, за інших рівних умов, переважніше. При виборі якнайкращого дерева рішень повинен досягатися певний компроміс між показниками точності і складності.

Існуючі методи побудови дерев рішень (всього їх кілька десятків), у принципі, можуть бути розділені на дві основні групи. До першої групи відносяться методи побудови строго-оптимальної по заданому критерію якості дерева, до другої групи — методи побудови приблизно-оптимального дерева.

Завдання пошуку оптимального варіанту дерева можна віднести до завдання дискретного програмування або вибору з кінцевого (але дуже великого) числа варіантів. У дискретному програмуванні розглядаються три основні види методів: повний перебір, метод динамічного програмування і метод гілок та меж. Проте ці методи в прикладенні до дерев рішень, як правило, є дуже трудомісткими, особливо при великому числі спостережень і характеристик. Тому доцільно обмежитися наближеними методами, до яких належать метод послідовного галуження, метод усікання і рекурсивний метод.

Метод послідовного галуження. Даний метод є процедурою поетапного галуження, при якому на кожному кроці вибирається кращий варіант розділення.

У методі усікання повчальна вибірка ділиться на дві частини. Перша частина використовується для побудови дерева методом послідовного галуження, причому параметри правила зупинки задаються такими, щоб забезпечити максимально можливу точність одержаного рішення, при цьому число листя дерева буде дуже великим. Друга частина вибірки служить для усікання (спрощення) одержаного дерева.

Рекурсивний метод. Для побудови дерева рішень у разі складної залежності між характеристиками, застосовують методи, складніші, ніж метод послідовного галуження. Загальна схема методу аналогічна тій, яка використовувалася в методі послідовного галуження, з тією різницею, що замість операції розділення вико-

ристовується складніша операція розростання. При порівнянні різних варіантів галуження деякої вершини необхідний критерій, який дозволяв би порівнювати ці варіанти і вибирати якнайкращий з них. Таким критерієм може бути частота помилок або відносна дисперсія. Якщо число помилок для цих варіантів співпадає, то щоб враховувати подібні випадки, для визначення якості розділення можна використовувати критерій ентропії.

Для прикладу застосування дерев рішень для класифікації розглянемо один з найвідоміших алгоритмів — індуктивний алгоритм побудови дерева рішень ID3. Згідно алгоритму ID3 на основі множини навчальних прикладів будуються дерева рішень, які забезпечують класифікацію об'єкта за його властивостями (атрибутами). Кожний внутрішній вузол дерева рішень відповідає за одну із властивостей об'єкта-кандидата і використовує його значення для вибору наступної гілки дерева. В процесі проходу по дереву перевіряються різні властивості. Ця процедура продовжується до тих пір, поки не буде досягнутий один з листків дерева, що означає клас, до якого відноситься даний об'єкт. В алгоритмі ID3 для упорядкування тестів і побудови майже оптимального дерева рішень використовується функція вибору тестів, побудована на основі теорії інформації

До вибіркових даних, використовуваних ID3, висуваються певні вимоги:

- Опис значення атрибуту*: такі ж самі атрибути повинні описувати кожний приклад і мати фіксоване число значень.*
- Наперед визначені класи*: атрибути прикладу мають бути вже визначені, а не навчені ID3.*
- Дискретні класи*: класи мають бути чітко окреслені. Безперервні класи, що розбиті в нечіткі (невиразні) категорії, як, наприклад, щодо металів «твердий, вельми твердий, гнучкий, м'який, вельми м'який», є підозрілими.*
- Достатність прикладів*: оскільки використовується індуктивне узагальнення, має бути досить випробувальних випадків, щоб відрізняти значимі взірці від випадкових подій.*

ID3 був інтегрований у низку комерційних пакетів типу «правила—індукції». Деякі реальні застосування включають медичний діагноз, оцінку ризику неплатежу по кредитній заяві на позику, виявлення причин аварійних режимів устаткування, класифікацію web-пошуку та інше. В подальшому розглянемо приклад ризику неплатежів.

Згідно алгоритму ID3 дерево рішень будується зверху вниз. Зауважемо, що кожна властивість (атрибут) дозволяє розбити на-

бір навчальних прикладів на непересічні підмножини, до кожної з яких відносяться всі приклади з однаковим значенням цієї властивості. По алгоритму ID3 кожен вузол дерева представляє деяку властивість, на підставі якої виконується розділення набору прикладів. Таким чином, алгоритм рекурсивно будує піддерево для кожного розділу. Ця процедура триває до тих пір, поки всі елементи розділу не будуть віднесені до одного і того ж класу. Цей клас стає кінцевим вузлом дерева. Оскільки для побудови простого дерева рішень важливу роль грає порядок тестування, в алгоритмі ID3 реалізований спеціальний критерій вибору тесту для кореневого вузла кожного піддерева.

Маючи набір навчальних прикладів і кілька дерев рішень, що дозволяють коректно класифікувати ці приклади, слід вибрати дерево, яке з найбільшою вірогідністю дозволить коректно класифікувати невідомі екземпляри. По алгоритму ID3 таким деревом вважається просте дерево рішень, що покриває всі навчальні приклади.

Вибір атрибутів (властивостей) в ID3 здійснюється на основі теорії інформації. Кожну властивість можна розглядати з погляду його внеску в процес класифікації. Алгоритм ID3 при виборі кореня поточного піддерева оцінює питому вагу інформації, що додається кожною властивістю. Потім він вибирає властивість, що має найбільшу інформативність.

Як вирішує ID3, який атрибут є кращим? Використовується статистична властивість, названа *інформаційним виграшем (gain)* (приріст інформації). Виміри (оцінки) gain показують, як добре надані атрибути навчальних прикладів формують кінцеві класи. Вибирається атрибут з найвищою інформацією. Для того, щоб визначити вигоду, використовується концепція *ентропія* з теорії інформації. В теорії інформації поняття ентропії увів американський математик-інженер К. Шеннон, який розглядав ентропію як міру невизначеності випадкової величини.

Ентропія вимірює кількість інформації в атрибуті. Якщо задана сукупність S з певним числом виходів (результатів), то ентропія обчислюється так:

$$E = -\sum_{i \in I} p_i \log_2 p_i , \qquad (1)$$

де p_i — пропорційна належність S до класу I. Зауважимо, що в даному разі S не атрибут, а є повним вибірковим набором.

Для прикладу використання алгоритму ID3 розглянемо проблему оцінки кредитного ризику (залежна категорійна змінна) на основі категорійних змінних — провісників: *кредитної історії*,

поточного боргу, наявності *поручительства* (застави) та неперервного провісника — *доходу*, який має три градації [6]. У табл.1 представлені приклади з відомим кредитним ризиком. Один з варіантів дерева рішень, яке розглядає повний набір провісників, показаний на рис. 1. Він містить приведені в табл. 1 дані і дозволяє коректно класифікувати всі об'єкти в таблиці. Кожен внутрішній вузол дерева рішень представляє деяку властивість, наприклад, борг або дохід. Кожному можливому значенню цієї властивості відповідає гілка дерева. Вузли-листя відображають результати класифікації, зокрема, низький або середній ризик. За допомогою цього дерева можна класифікувати клієнта, тип якого невідомий: для кожного внутрішнього вузла перевіряється значення відповідної властивості для даного клієнта і здійснюється перехід по відповідній гілці. Процес завершується досягши кінцевого вузла, що визначає клас об'єкта.

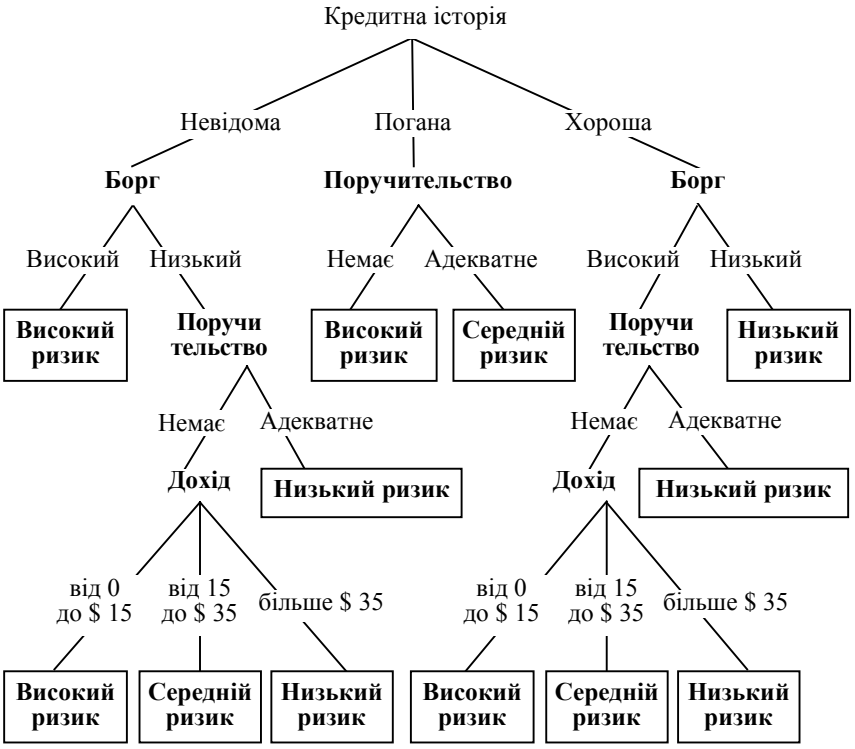

Рис. 1. Варіант дерева рішень для оцінки кредитного ризику

Помітимо, що при класифікації кожного конкретного екземпляра за допомогою цього дерева враховуються не всі властивості, представлені в табл. 1. Наприклад, якщо людина має хорошу кредитну історію і низький борг, то згідно дереву без урахування доходу і поручительства з ним зв'язується низький ризик. Це дерево дозволяє коректно класифікувати всі приклади.

В цілому, розмір дерева, необхідного для класифікації конкретного набору прикладів, варіюється залежно від властивостей, що перевіряються. На рис. 2 показано набагато простіше дерево, яке з використання алгоритму ID3, але також дозволяє коректно класифікувати приклади з табл. 1.

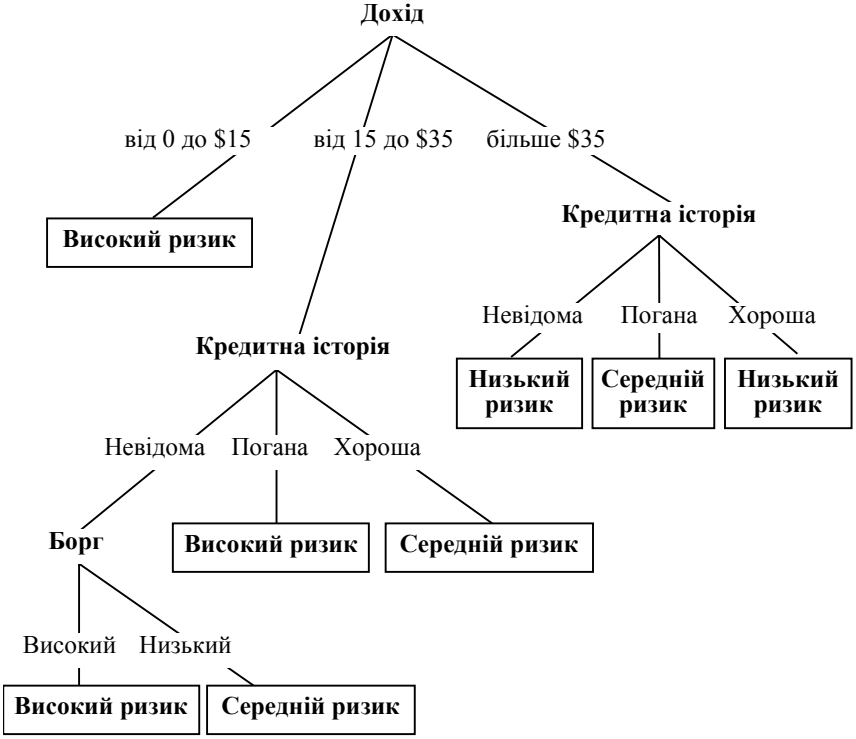

Рис. 2. Спрощене дерево рішень для оцінки кредитного ризику

Розглянемо процес побудови цього дерева на основі даних з табл. 1. Маючи повну таблицю прикладів, алгоритм ID3 вибирає як кореневу властивість значення доходу на основі функції вибору, описаної раніше. При цьому множина прикладів ділиться на

три частини, як показано на рис. 3. Елементи кожної частини представлені порядковими номерами прикладів в табл. 1.

Таблиця 1

ДАНІ ПРО КРЕДИТНУ ІСТОРІЮ

№	Ризик	Кредитна історія	Борг	Поручитель-ство	Дохід
1	Високий	Погана	Високий	Немає	від 0 до
2	Високий	Невідома	Високий	Немає	від 15 до
3	Середній	Невідома	Низький	Немає	від 15 до
4	Високий	Невідома	Низький	Немає	від 0 до
5	Низький	Невідома	Низький	Немає	більше
6	Низький	Невідома	Високий	Адекватне	більше
7	Високий	Погана	Низький	Немає	від 0 до
8	Середній	Погана	Низький	Адекватне	більше
9	Низький	Хороша	Низький	Немає	більше
10	Низький	Хороша	Високий	Адекватне	більше
11	Високий	Хороша	Високий	Немає	від 0 до
12	Середній	Хороша	Високий	Немає	від 15 до
13	Низький	Хороша	Високий	Немає	більше
14	Високий	Погана	Високий	Немає	від 15 до

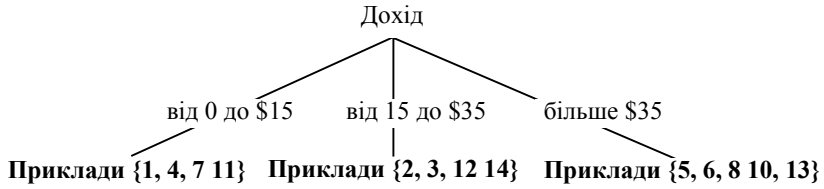

Рис. 3. Перший фрагмент дерева рішень

До розділу {1, 4, 7, 11} відносяться клієнти з високим ризиком; алгоритм ID3 створить відповідний кінцевий вузол. Потім як кореневий вузол піддерева розділу {2, 3, 12, 14} вибирається властивість «кредитна історія». На рис. 4 елементи цього розділу в свою чергу розбиваються на три групи: {2, 3}, {14} і {12}. Аналогічно будуються всі фрагменти дерева, що представлене на рис. 2.

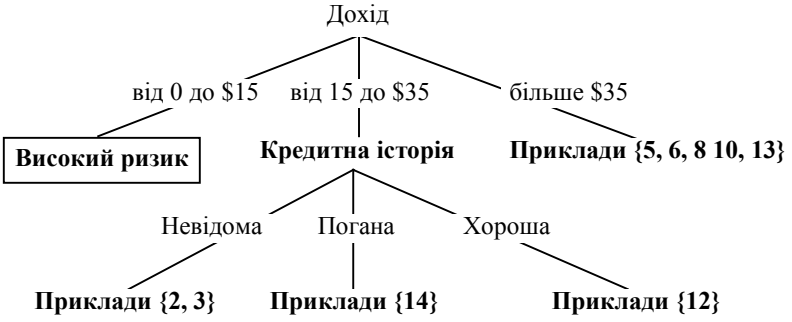

Рис. 4. Другий фрагмент дерева рішень

Дерево рішень можна розглядати з погляду інформації про приклади. Інформативність дерева обчислюється на основі вірогідності різних типів класифікації. Наприклад, якщо припустити, що всі приклади в табл. 1 з'являються з однаковою вірогідністю, то p(високий ризик) = $^{6}/_{11}$, p (середній ризик) = $^{3}/_{14}$ і p (низький) = $^{5}/_{14}$.

Отже, інформативність розподілу T.1, описаного в табл.1, а значить і будь-якого дерева, що покриває ці приклади, згідно формули (1) складає: $E[T.1] = -\ 6/14 \times \log_2(6/14) - 3/14 \times \log_2(3/14) - 5/14 \times \log_2(5/14) = -\ 6/14 \times (-\ 1{,}222) - 3/14 \times (-\ 2{,}222) - 5/14 \times (-\ 1{,}485) = 1{,}531$ біт.

Виграш від використання властивості P обчислюється згідно формули (1) як різниця загальної інформативності дерева і об'єму інформації, необхідного для завершення побудови дерева. Повертаючись до прикладів з табл. 1, при виборі кореня дерева властивість «дохід», приклади будуть розділені на три групи: C_1 = {1,4,7,11}, C_2 = {2, 3, 12, 14} і C_3 = {5, 6, 8, 9, 10, 13}. Інформація, необхідна для завершення побудови дерева, складає: $E[\text{дохід}] = 4/14 \times E[C_1] + 4/14 \times E[C_2] + 6/14 \times E[C_3] = 4/14 \times 0{,}0 + 4/14 \times 1{,}0 + 6/14 \times 0{,}650 = 0{,}564$ біт.

Інформаційний виграш (gain) от такого розбиття даних табл. 1 складає: $\text{gain}(\text{дохід}) = E[D_{7.2}] - E[\text{дохід}] = 1{,}531 - 0{,}564 = 0{,}967$ біт. Аналогично можно показати, що gain (кредитна історія) = 0,266; gain(борг) = 0,581; gain(застава) = 0,576.

Оскільки дохід забезпечує найбільший інформаційний виграш (0,967), то саме ця властивість вибирається як корінь дерева рішень алгоритму ID3. Такий аналіз рекурсивно виконується для кожного піддерева до повної побудови всього дерева.

Не дивлячись на те, що алгоритм ID3 будує просте дерево рішень, проте зовсім не очевидно, що за допомогою цих дерев

можна ефективно класифікувати невідомі приклади. Тому алгоритм ID3 був протестований на контрольних прикладах і реальних застосуваннях. Тести підтвердили його хорошу працездатність. Існують варіанти алгоритму ID3, що дозволяють вирішувати задачі в умовах зашумленних даних і дуже великих навчальних множин.

Проте при використанні алгориму ID3 виникає множина проблем, які часто виникають при роботі з великими масивами даних. Вирішення цих проблем привело до створення нового покоління алгоритмів навчання, заснованих на побудові дерева рішень, зокрема алгоритма C4.5.

Література

1. *Ситник В. Ф.* Засоби дейтамайнінгу для аналізу бізнесових рішень // Науково-практичний журнал «Науково-технічна інформація». — № 3. — 2002. — С. 60—64.
2. *Ситник В. Ф., Ситник Н. В.* Проблеми впровадження дейтамайнінгу в бізнесі/ Вчені записки: Наук. зб. — Вип. 6 . — К.: КНЕУ. 2004. — С. 58—64.
3. *Breiman L., Friedman J., Olsen R.* and *Stone C.* Classification and Regression Trees. Monterey, CA: Wadsworth, 1984
4. Classification Trees. — http://www.statsoft.com/textbook/stclatre.html
5. Classification and Regression Trees (C&RT). — http://www.statsoft.com/textbook/stcart.html
6. *Джорж Ф Лютер*. Искусственный интеллект: стратегии и методы решения сложных проблем, 4-е издание. Пер. с англ. — М.: Издательский дом «Вильямс», 2003. — 864 с.

Стаття надійшла до редакції 18.04.06