8.4. Лабораторна робота №4 «Аналітичні рішення за допомогою дерева рішень».

Завдання до лабораторної роботи.

1. Відповідно до заданого варіанту підготувати навчальну вибірку у вигляді таблиць MS Excel і зберегти їх як персональні файли. Для підготовки даних використовувати тематичні сайти Інтернет, результати проходження практик, довідники і каталоги.

2. Здійснити побудову дерева рішень і виконати його аналіз в пакеті Deductor відповідно до методики, представленої в розділі «Порядок виконання роботи».

3. Отримати правила рішень і провести декілька експериментів відповідно до методики, представленої в розділі «Порядок виконання роботи».

Варіанти завдань.

N варіанту	Процес
1	Діагностика захворювань
2	Діагностика комп'ютерів
3	Діагностика технічних систем
4	Діагностика комп'ютерних мереж
5	Прогнозування курсу валют
6	Прогнозування вартості нерухомості
7	Прогнозування курсу акцій
8	Реакція ринку на підвищення податків
9	Прогнозування рівня інфляції
10	Оцінка фінансового стану фірми
11	Оцінка кредитоспроможності фірми
12	Прогнозування рівня цін на продовольчі товари
13	Прогнозування рівня цін на промислові товари

Теоретична частина.

Дерева рішень (decision trees) є одним з найбільш популярних підходів до вирішення задач інтелектуального аналізу даних. Вони створюють ієрархічну структуру класифікуючих правил типа «якщо … то» (if-then), що має вигляд дерева. Аби прийняти рішення, до якого класу слідує віднести деякий об'єкт або ситуацію, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд «значення параметра A більше B?». Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня; потім знову слідує питання, пов'язане з відповідним вузлом і так далі. Наведений приклад ілюструє роботу так званих бінарних дерев рішень, в кожному вузлі яких, галуження виконується по двох напрямах (тобто на питання, задане у вузлі, є лише два варіанти відповідей, наприклад «Так» чи «Ні»). Проте, в загальному випадку, відповідей, а, отже, гілок, що виходять з вузла, може бути більше. Дерево рішень складається з вузлів – де виконується перевірка умови, і листя – кінцевих вузлів дерева, вказуючих на клас (вузлів рішення) (рис. 8.21).

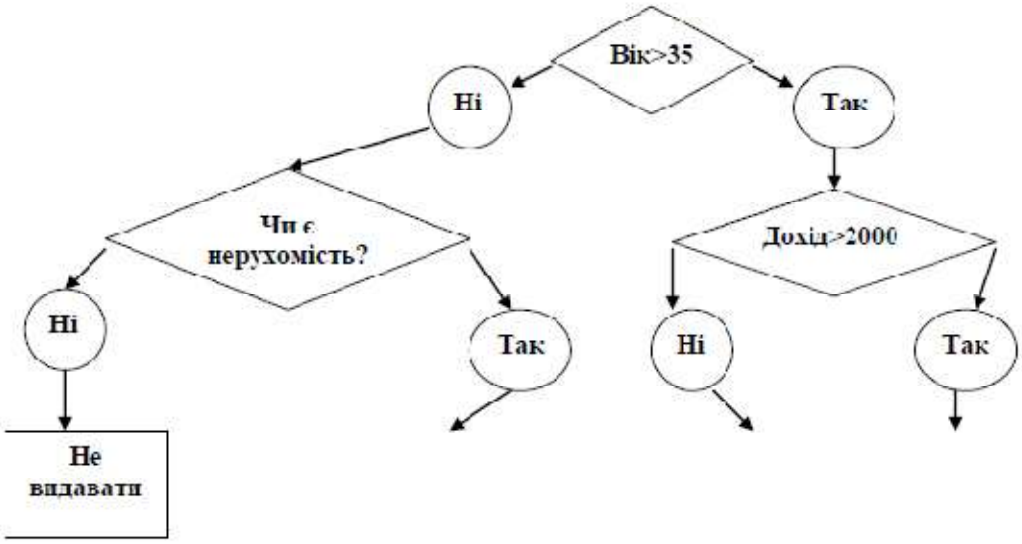

Рис. 8.21. Бінарне дерево рішень.

Сфера застосування дерев рішень в даний час вельми широка, але всі задачі, що вирішуються цим апаратом, можуть бути об'єднані в три класи.

1. Опис даних. Дерева рішень дозволяють зберігати інформацію про дані в компактній формі. Замість громіздких масивів даних можна зберігати дерево рішень, яке містить точний опис об'єктів.

2. Класифікація. Дерева рішень відмінно справляються із задачами класифікації, тобто віднесення об'єктів до одного із заздалегідь відомих класів.

3. Регресія. Якщо цільова змінна має безперервні значення, дерева рішень дозволяють встановити залежність цільової змінної від незалежних (вхідних) змінних. Наприклад, до цього класу відносяться задачі чисельного прогнозування (передбачення значень цільовій змінній).

Підготовка навчальної вибірки.

Для побудови дерева рішень готується навчальна вибірка так, як це описано для нейромережі. Різниця полягає в тому, що вихідне поле для дерева рішень може бути лише одне - *дискретне*

Для полів, що подаються на входи і вихід дерева рішень, також задається нормалізація. Можна задати або лінійну нормалізацію, або нормалізацію унікальними значеннями. Налаштування навчальної вибірки виконується аналогічно, як і для нейромережі. Параметри навчання дерева рішень наступні:

- мінімальна кількість прикладів, при якій буде створений новий вузол. Задається мінімальна кількість прикладів, яка можлива у вузлі. Якщо прикладів, які потрапляють в даний вузол, буде менше заданого - вузол вважається листом (тобто подальше галуження припиняється). Чим більше цей параметр, тим менш гіллястим виходить дерево;
- будувати дерево з достовірнішими правилами в збиток складності. Включає спеціальний алгоритм, який, ускладнюючи структуру дерева, збільшує достовірність результатів класифікації. При цьому дерево виходить, як правило, більш гіллястим;
- рівень довіри, що використовується при відсіканні вузлів дерева. Значення цього параметра задається у відсотках і повинне лежати в межах від 0 до 100. Чим більше рівень довіри, тим більше гіллястим виходить дерево, і, відповідно, чим менше рівень довіри, тим більше вузлів буде відсічено при його побудові.

Якість побудованого дерева після навчання можна оцінити по декількох параметрах. По-перше, це число розпізнаних прикладів в навчальному і тестовому наборах даних. Чим вище це число, тим якісніше побудоване дерево. По-друге, це кількість вузлів в дереві. При дуже великому їх числі дерево стає важким для сприйняття. Це також означає дуже слабку залежність вихідного поля від вхідних полів.

Кожне правило характеризується *підтримкою* і *достовірністю*. Підтримка – загальна кількість прикладів класифікованих даним вузлом дерева. Достовірність – кількість правильно класифікованих даним вузлом прикладів.

Порядок виконання роботи.

1. Продовжимо приклад, розглянутий в попередній лабораторній роботі, присвяченій нейронним мережам. Всіх кредиторів можна розділити на два класи – кредитоспроможних і некредитоспроможних. Вочевидь, існують деякі правила віднесення кредиторів до того або іншого класу. Але при чималому числі їх характеристик майже неможливо побудувати ці правила. Це дозволяють зробити дерева рішень.

Підготуємо і завантажимо навчальну вибірку згідно варіанту за допомогою «*Мастера импорта*» і викличемо «*Мастер обработки*». У вікні, що з'явилося, виберемо режим – «*Дерево решений*». Кроки 1 – 4 «*Мастера обработки*» виконуються аналогічно попередній лабораторній роботі.

2. При побудові правил задамо мінімальну кількість прикладів, при якій буде створений новий вузол, рівним 1. Будуватимемо дерево з достовірнішими правилами в збиток складності (у вікні «*Параметры отсечения*» задамо 100) – крок 5 «*Мастера обработки*».

3. На кроці 6 запускаємо процес побудови дерева рішень і на кроці 7 вибираємо способи відображення даних – «*Дерево решений*» і «*Правила*».

4. Отримане дерево рішень містить 3 вузли і 7 правил (рис. 8.22).

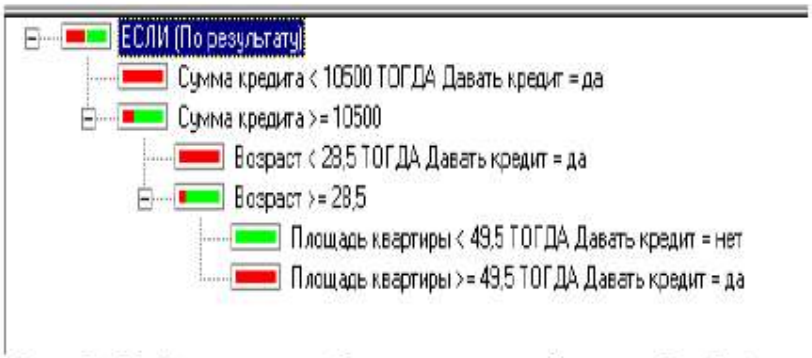

Рис. 8.22. Результат побудови дерева рішень в Deductor.

Таке дерево містить в собі правила, слідуючи яким можна віднести кредитора в одну з груп ризику і зробити висновок про видачу кредиту. Правила читаються

з вузлів, розташованих правіше. Побудовані правила можна також проглянути у вигляді списку правил (рис. 8.23).

Итого правил: 4

N	Условие	Следствие (Давать кредит)	Поддержка		Достоверность	
			%	Кол-во	%	Кол-во
1	Сумма кредита < 10500	да	30,00	3	100,00	3
2	Сумма кредита >= 10500 И Возраст < 28,5	да	10,00	1	100,00	1
3	Сумма кредита >= 10500 И Возраст >= 28,5 И Площадь квартиры < 49,5	нет	50,00	5	100,00	5
4	Сумма кредита >= 10500 И Возраст >= 28,5 И Площадь квартиры >= 49,5	да	10,00	1	100,00	1

Рис. 8.23. Список правил.