Марчук Д.К., студент, група ПІ-47м,
Марчук Г.В., старший викладач,
Житомирський державний технологічний університет

ДЕРЕВА РІШЕНЬ У DATA MINING

Ухвалення рішення - це процес раціонального або ірраціонального вибору альтернатив, що має на меті досягнення усвідомлюваного результату. Один з методів автоматичного аналізу даних є дерева рішень. Перші ідеї створення дерев рішень відносяться до робіт Ховленда (Hoveland) і Ханта (Hunt) кінця 50-х років XX століття. Однак, основною роботою, що дала імпульс для розвитку цього напрямку, стала книга Ханта (Hunt, E.B.), Меріна (Marin J.) і Стоуна (Stone, P.J) «Experiments in Induction», що побачила світ у 1966 р.

Дерева рішень, що використовуються в Data Mining, бувають двох основних типів:

– Аналіз дерева класифікації, коли результат, що передбачається є класом, до якого належать дані;

– Регресійний аналіз дерева, коли результат, що передбачається, можна розглядати як дійсне число (наприклад, ціна на будинок, або тривалість перебування пацієнта у лікарні).

Data mining (видобування даних, інтелектуальний аналіз даних, глибинний аналіз даних) - збірна назва, що використовується для позначення сукупності методів виявлення в даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень в різних сферах людської діяльності. Термін введений Григорієм Пятецким-Шапіро у 1989 році.

На сьогоднішній день існує велика кількість алгоритмів, що реалізують дерева рішень: CART, C4.5, CHAID, CN2, NewId, ITrule та інші.

Дерево прийняття рішень використовується в області статистики та аналізу даних для моделей, що прогнозуються. На ребрах дерева рішення записані атрибути, від яких залежить цільова функція, в вершинах записані значення цільової функції, а в інших вузлах - атрибути, за якими розрізняються випадки.

Мета полягає в тому, щоб створити модель, яка передбачає значення цільової змінної на основі декількох змінних на вході:

$$(x,Y)=(x_1, x_2, x_3, \dots , x_k, Y)$$

Залежна змінна Y є цільовою змінною, яку необхідно проаналізувати, класифікувати і узагальнити. Вектор x складається з вхідних змінних $x_1, x_2, x_3, \dots, x_k$, які використовуються для виконання завдання.

Процес гри «Хрестики-нулики» теж можна представити у вигляді дерева рішень, вузли якого є станом ігрового поля після ходу одного з гравців:

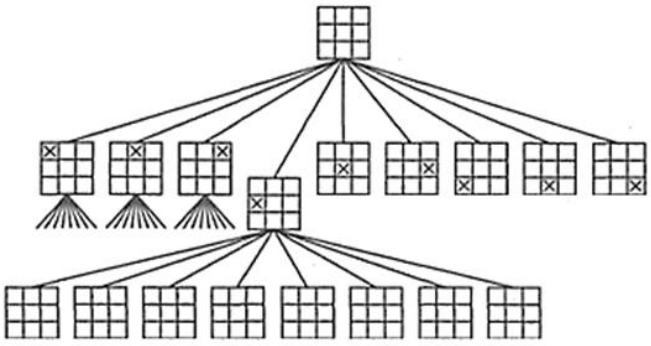

Рис. 1. Дерево рішень гри «Хрестики-нулики»

До речі гра має 362 880 сценаріїв розвитку.

C4.5 - алгоритм для побудови дерев рішень, розроблений Джоном Квінлану. В даному алгоритмі було додано відсікання гілок, можливість роботи з числовими атрибутами, а також можливість побудови дерева з неповною навчальною вибіркою, в якій відсутні значення деяких атрибутів.

Алгоритм CART (Classification and Regression Tree), як видно з назви, вирішує завдання класифікації і регресії побудови дерева рішень. Він розроблений в 1974-1984 роках чотирма професорами статистики: Лео Брейманом (Берклі), Джеромом Фрідманом (Jerome H. Friedman, Стенфорд), Чарлзом Стоуном (Charles Stone, Берклі) і Річардом Олшеном (Richard A. Olshen, Стенфорд).

Розглядаючи дані алгоритми, можна прийти до такого висновку. Перевагами дерев рішень є: простота в розумінні та інтерпретації; не вимагають підготовки даних; дозволяють оцінити модель за допомогою статистичних тестів; дозволяють створювати класифікаційні моделі в тих областях, де аналітику досить складно формалізувати знання; алгоритм конструювання дерева рішень не вимагає від користувача вибору вхідних атрибутів; швидко навчаються.

Недоліками дерев рішень є те, що можуть з’явитися занадто складні конструкції, які при цьому недостатньо повно представляють дані.