ПОРІВНЯННЯ МЕТОДІВ ПОБУДОВИ ДЕРЕВ РІШЕНЬ В ОБРОБЦІ ВЕЛИКИХ МАСИВІВ ДАНИХ

Ярощук О. В., Якушина А. О., Шпінарева І. М.
Одеський національний університет імені І.І. Мечникова

Анотація. Розглядається застосування дерев рішень в задачах прогнозування при обробці великих масивів даних. Метою роботи є дослідження алгоритмів дерев рішень в задачі прогнозування. Серед алгоритмів побудови дерев рішень були обрані ID3, C4.5 та CART. Прикладом використання дерев рішень, який розглядається в роботі, є діагностика наявності серцево-судинних захворювань. Також з метою подолання недоліків дерев рішень був використаний ліс рішень.

Ключові слова: інтелектуальний аналіз даних, дерево рішень, ліс рішень, класифікація, медицина.

Інтелектуальний аналіз даних використовується для вилучення корисної інформації з великих наборів даних і для її відображення в легко інтерпретованих візуалізаціях. Дерева рішень є одним з найбільш ефективних методів аналізу даних. Вони широко використовуються в різних сферах життя, тому що вони прості у використанні, вільні від двозначності і надійні навіть при наявності пропущених значень. Як дискретні, так і неперервні змінні можуть використовуватися як цільові або як незалежні змінні. Останнім часом методологія дерева рішень стала також популярною в медичних дослідженнях.

В даний час розроблено значну кількість алгоритмів навчання дерева рішень, але найбільшого поширення і популярність отримали ID3, C4.5 та CART. В таблиці 1 розглянуті основні ознаки цих алгоритмів [1-2].

Таблиця 1 – Порівняння алгоритмів побудови дерев рішень

Алгоритми	ID3	C4.5	CART
Критерій розбиття	Критерій приросту інформації	Критерій приросту інформації	Індекс Джині
Відсікання	-	Попереднє	Попереднє
Задачі	Класифікація/ Прогнозування	Класифікація / Прогнозування/ Регресія	Класифікація / Прогнозування/ Регресія
Вид розбиття	Множинне	Множинне	Бінарне

Критерій зупинки	Об'єкти у вузлі одного класу / Немає зменшення ентропії критерію	Обмеження на число об'єктів в листі	Обмеження на число об'єктів в листі

Для дослідження методів дерев рішень використовувалась база даних Heart Disease UCI [3]. Ця база даних містить 76 атрибутів, але всі опубліковані експерименти відносяться до використання підмножини з 14 з них. Експерименти з базою даних Клівленда сконцентрувались на простій спробі відрізнити присутність від відсутності серцево-судинних захворювань.

Як критерій якості моделей була обрана F-міра, яка являє собою гармонійне середнє між точністю і повнотою. Після застосування методів отримані такі результати: ID3 – 75,5%; C4.5 – 76,2%; CART – 80,3%.

Для покращення результату застосовано ліс рішень з метою вирішення можливої проблеми перенавчання дерев рішень [5]. Ліс рішень передбачає собою кілька дерев, результат класифікації яких визначається шляхом голосування. Основою ліса рішень став алгоритм CART. В результаті отримали ліс рішень, F-міра якого складала 89,6%.

Одним з основних переваг дерев рішень є їх інтерпретованість. Для прикладу розглянемо перші три рівні побудованого лісу рішень (рис.1). У кореневому вузлі розташувалась ознака болі в грудній клітці типу стенокардії, присутність якого в більшій частині передбачає наявність захворювань серцево-судинних систем. Далі йдуть такі ознаки, як артеріальний тиск у стані спокою (в мм рт.ст. при надходженні до лікарні), нахил піку сегмента ST, частота серцебиття, рівень холестерину в мг / дл.

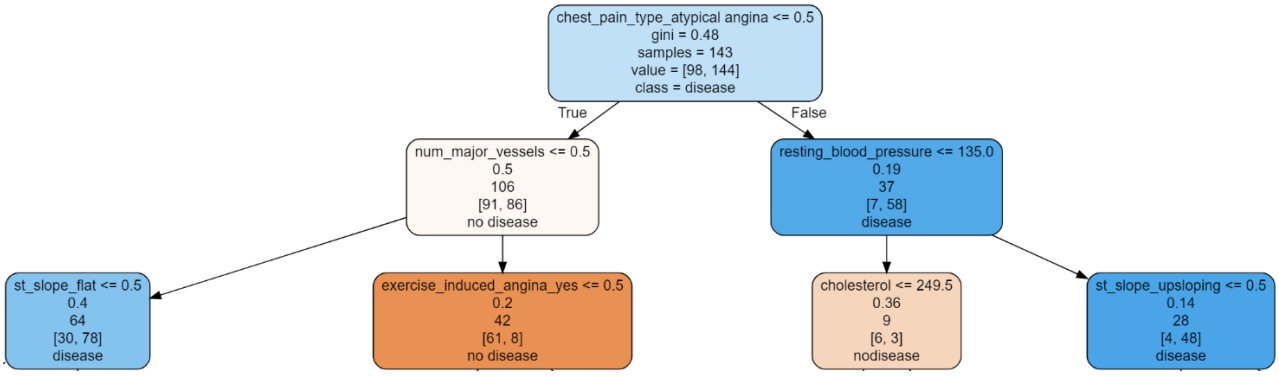

Рис. 1 – Перші три рівня побудованого ліса рішень.

Таким чином, дерева рішень дозволяють не тільки вирішити поставлене завдання регресії або, в нашому випадку, прогнозування, але також дають можливість проаналізувати отриманий результат і виявити ознаки, які найбільше вплинули на отриманий результат. Проте, серед недоліків алгоритмів побудови дерев рішень варто відзначити можливість швидкого перенавчання, але

проблема вирішується шляхом відсіканням гілок, а також використанням лісу рішень.

Література

1. Colin A. Building decision trees with the ID3 algorithm, 1996.Dr.Dobb's Journal of Software Tools for Professional Programmer,21(6),107-109.
2. Quinlan J. R. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers Inc., 1993.
3. Heart Disease Data Set [Електронний ресурс] – Режим доступу: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.