УДК 681.3

О. М. Козачко

МОДЕЛЬ ПРОГНОЗУВАННЯ ЗАТРИМКИ ПРИБУТТЯ АВІАРЕЙСІВ

Вінницький національний технічний університет

Анотація

В роботі запропоновано модель прогнозування затримки прибуття авіарейсів за допомогою бінарного класифікатора, який побудований на основі дерева рішень. Для побудови моделі використовувалась така схема: попередній аналіз даних, дослідження впливу атрибуту рейсу на середню тривалість затримки прибуття, побудову класифікатора, навчання та тестування класифікатора.

Ключові слова: бінарний класифікатор, класифікатор, прогнозування затримки прибуття авіарейсів, дерево рішень, авіарейси.

Abstract

In this work the forecasting model of flights arrival delay based on binary classifier, which is constructed on the basis of decision tree is proposed. The following scheme was used to construct the model: preliminary analysis of data, study of the effect of the flight attribute on the average length of arrival delay, classifier design, training and testing of the classifier.

Key words: binary classifier, classifier, prediction of arrival delay of flights, decision tree, flights.

Задачу прогнозування затримки прибуття авіарейсів будемо розглядати у вигляді задачі побудови бінарного класифікатора, який за заданими атрибутам рейсу повинен передбачити наявність або відсутність затримки. Для побудови такого класифікатора будемо використовувати наступну схему: попередній аналіз даних, дослідження впливу атрибутів рейсу на середній час затримки прибуття, побудова моделі класифікатора, навчання і тестування класифікатора. Попередній аналіз та дослідження впливу атрибутів можна познайомитися в роботах [1, 2, 3].

Розглянемо модель прогнозування у вигляді такої функції:

$$y = f(x1, x2, x3, x4, x5, x6, x7, x8, x9) \quad (1)$$

де x1 - день місяця;
x2 - день тижня;
x3 - запланований час відбуття;
x4 - запланований час прибуття;
x5 - компанія-перевізник;
x6 - затримка відбуття;
x7 - аеропорт відправлення;
x8 - аеропорт прибуття;
x9 - дистанція між пунктами відправлення та прибуття.

Вихідна змінна y приймає два значення: 1 - затримка; 0 - відсутність затримки.

Залежність (1) описує вихідна вибірка:

$$(\mathbf{X}^r, y^r), r = 1 \dots m, \quad (2)$$

де m - кількість авіарейсів (розмір вибірки);

$\mathbf{X}^r = \{x_1^r, x_2^r, x_3^r, x_4^r, x_5^r, x_6^r, x_7^r, x_8^r, x_9^r\}$- атрибути r-го авіарейсу, r = 1 ... m;

$y^r = \begin{cases} 1, \text{якщо } ArrDelay > 0 \\ 0, \text{якщо } ArrDelay \le 0 \end{cases}$ ┤ вихідна змінна прогнозу затримки прибуття авіарейсу

Необхідно визначити таку структуру моделі, яка забезпечує мінімальну частоту помилки класифікації [4]:

$$MCR(f) = \frac{\sum_{j=1}^{m} \Delta_j}{m},$$

де $\Delta_j = \begin{cases} 1, \text{якщо } y^j = f(X^j) \\ 0, \text{якщо } y^j \neq f(X^j) \end{cases}$.

У вибірці (2) присутні числові, порядкові і категоріальні атрибути. Категоріальним атрибутами є: компанія-перевізник (x5), аеропорт відправлення (x7); аеропорт прибуття (x8). Порядковими є: день

місяця (x1) і день тижня (x2). Решта атрибути є числовими. Для роботи з категоріальним і порядковими атрибутами необхідно їх попередньо перетворити в числовий формат. Значення кожного атрибута закодуємо числом, який відповідає частоті наявності конкретного значення. Наприклад, значення PS атрибута компанія-перевізник зустрічається у вибірці 3132 раз. Тому це значення замінимо даними числом.

Для побудова моделі використано дерево рішення. Синтез дерева проводився алгоритмом CART, який реалізований в пакеті Statistic Toolbox в програмному середовищі Matlab. В алгоритмі CART розщеплення на гілки відбувається за допомогою принципу мінімізації індексу Джина.

На рис. 1 показано дерево класифікації рейсів, отримане за допомогою алгоритму CART. Безпомилковість класифікації рейсів на тестовій і навчальної вибірки дорівнює 3% та 2% відповідно. З отриманого дерева видно, що атрибути x2 x3 і x7 не використовується через те, що вони не є інформативними.

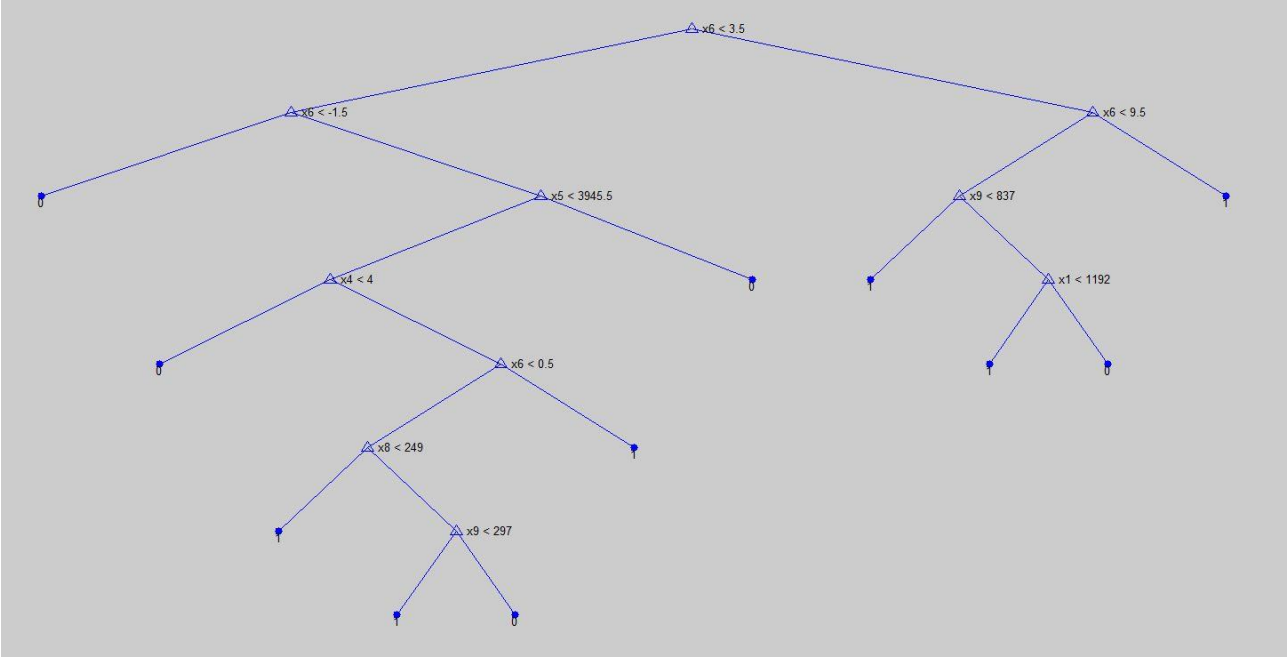

Рис. 1. Дерево рішень, що класифікує затримки прибуття авіарейсів

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Методологія та організація наукових досліджень : навчальний посібник / Б. І. Мокін, О. Б. Мокін. – 2-е вид., змін. та доп. – Вінниця : ВНТУ, 2015. – 317 с.

2. Мітюшкін Ю.І., Мокін Б.І., Ротштейн О.П. Soft Computing: ідентифікація закономірностей нечіткими базами знань. Монографія. - Вінниця: УНІВЕРСУМ-Вінниця, 2002. - 145 с.

3. Технології обробки та моделювання екологічної та економічної інформації [Електронний навчальний посібник] / В. Б. Мокін, А. В. Поплавський, М. П. Боцула, А. Р. Ящолт. — Вінниця : ВНТУ, 2015. — 120 с.

4. Штовба С.Д., Галущак А.В. Ідентифікація багатофакторних залежностей за допомогою баз знань. Лабораторний практикум : електронний навчальний посібник. – Вінниця: Вінницький національний технічний університет, 2016. – 96 с

Козачко Олексій Миколайович — к.т.н., доцент, доцент кафедри системного аналізу комп'ютерного моніторингу та інженерної графіки, Вінницький національний технічний університет, Вінниця, lekoz80@gmail.com.

***Kozachko Oleksiy* —** Cand. Sc. (Eng), Assistant Professor of Department of system analysis, computer monitoring and engineering graphics, Vinnytsia National Technical University, Vinnytsia, lekoz80@gmail.com.