УДК 681.3

В. В. Мазуренко

НЕЧІТКА МОДЕЛЬ ПРОГНОЗУВАННЯ КІЛЬКОСТІ КОМЕНТАРІВ В МЕРЕЖІ ФЕЙСБУК

Вінницький національний технічний університет

Анотація

Запропонована інформаційна система проектування нечіткої моделі по типу бази знань Мамдані на прикладі нечіткої моделі прогнозування кількості коментарів під публікаціями в мережі Фейсбук.

Ключові слова: Мамдані, нечітка логіка, бази знань, коментарі, Фейсбук.

Abstract

An information system for fuzzy model design based on the type of Mamdani knowledge base is proposed with the example of fuzzy model for predicting the number of comments under Facebook posts.

Keywords: Mamdani, fuzzy logic, knowledge base, comments, Facebook.

Вступ

Сьогодні соціальні мережі є провідним середовищем розповсюдження контенту та залучення аудиторії. Саме тому соціальні мережі є предметом аналізу сьогодення з величезною кількістю інструментів та моделей. Однією з важливих задач є задача прогнозування залучення аудиторії та прогнозування тренду [1].

Дана робота розглядає задачу прогнозування кількості коментарів під публікаціями на «Сторінках Фейсбук» [2]. Для чого використовується інформаційна система проектування нечіткої моделі по типу бази знань Мамдані [3].

Метою роботи є перевірка ефективності запропонованої системи та порівняння отриманої нечіткої моделі Мадані [4] з моделями на базі дерев рішення CART [5] та лінійної регресії.

Постановка задачі

Ми зосередилися на методах прогнозування за допомогою баз знань та лінійної регресії, що дають інтерпретабельний результат. Для прогнозування кількості коментарів вирішується задача регресії. Ми уже отримали зріс кількості коментарів у часі [2]. Завдання - це передбачити скільки коментарів до публікації що очікується наступні 24 год. Попередньо підготовленні данні розбиті на навчальну (40,949 екземплярів) і тестовий набір (1000 екземплярів). Навчальний набір використовується для підготовки регресора та оцінки його ефективності, потім регресор оцінюється за допомогою тестових даних.

Використані входи такі:

C1: Загальна кількість коментарів перед обраною часовою міткою.

C2: Кількість коментарів за останні 24 години до часової мітки.

C3: Кількість коментарів за останні 24 години зібрані за 48 годин до часової мітки.

C4: Кількість коментарів за останні 24 години зібрані 72 години до часової мітки.

C5: Різниця між C2 і C3.

Інформаційна система

Вище описані моделі і методи реалізовані у формі інформаційної технології нечіткої ідентифікації. Нечітка ідентифікація здійснюється поетапно, згідно з концепцією "генерація - селекція - редукція - налаштування" [3].

На першому етапі "генерація" відбувається генерація нечітких правил з експериментальних даних методом прямого проходу. Даний метод базується на ідеях генерації бази знань методом Ванга-Менделя [6], з тією лише відмінністю, що консеквентом правила вибирається терм не з максимальною приналежністю для якоїсь однієї стрічки вибірки даних, а терм з максимальною середньою на-

лежності за всіма даними з відповідною зони факторного простору.

Другий етап "селекція" реалізований за допомогою бінарного генетичного алгоритму з кодуванням за Пітсбургською схемою [7].

Третій етап "редукція" реалізований тим же генетичним алгоритмом із застосуванням альтернативного кодування для зменшення кількості антецедентів в кожному з правил.

Останній етап "налаштування" налаштовує внутрішню структуру нечіткої бази знань з використанням градієнтних і квазіньтоновских методів оптимізації. Для збереження інтерпретабельності оптимізація здійснюється з обмеженнями [3].

Результати дослідження

Перед початком експерименту вхідні та вихідні величини було необхідно нормалізувати та збалансувати рис. 1. Для цього вихідні значення Y<5 (публікації з 5 і менше коментарів) були відібрані в окрему групу, туди потрапили 33152 екземпляра. За допомогою Fuzzy c-means clustering екземпляри були об’єднані як знайдені центри 200-т кластерів. Таке стиснення даних не тільки зменшило навчальну вибірку, але збалансувало її. Решта екземплярів була відфільтрована по 0,995 від максимального вихідного значення ($Y<99.5\%*Y_{max}$). Таким чином навчальна вибірка склала 7967 екземплярів.

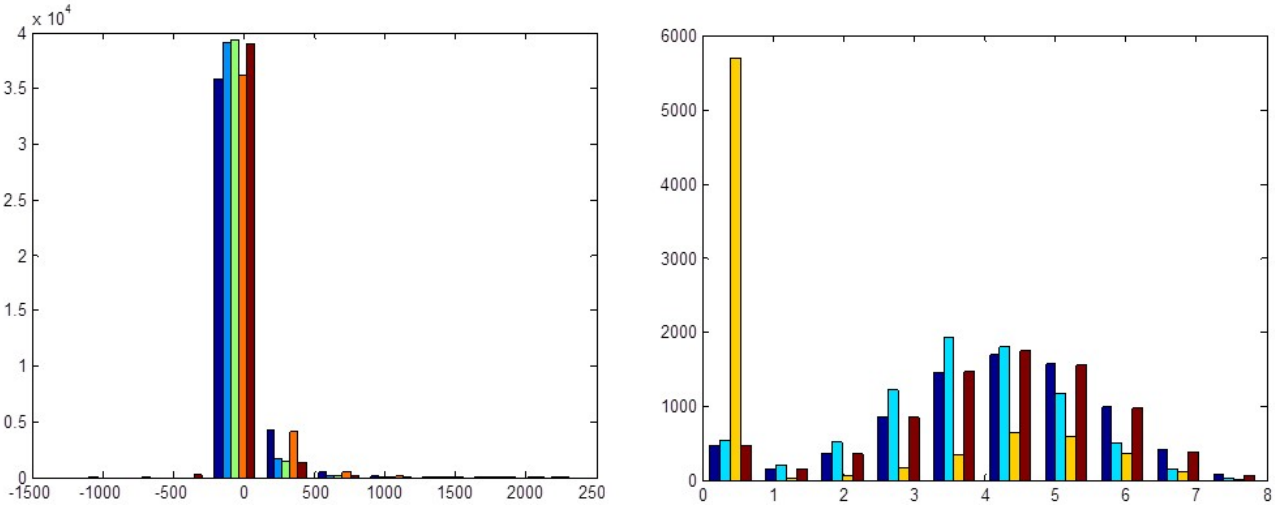

Рис. 1. Розподіл вхідних величин до та після нормалізації

Після чого було проведене навчання для дерева рішень CART (рис.2),

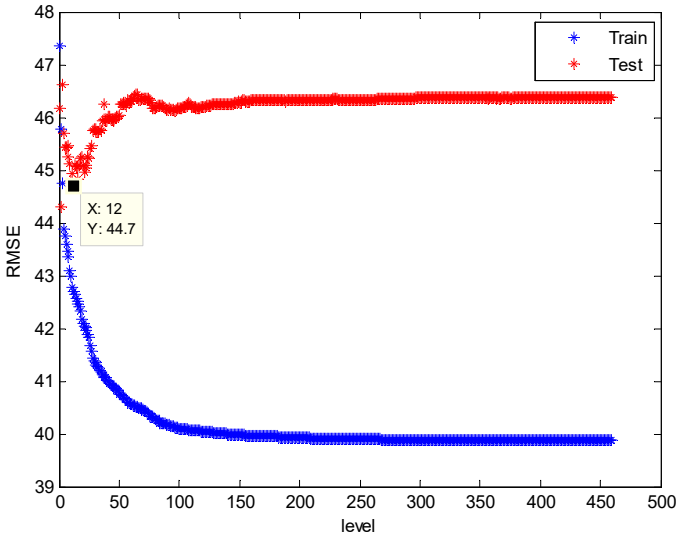

Рис. 2. Результати навчання дерева з різними рівнем підрізання

трьох-етапне навчання нечіткої бази знань Мамдані (рис.3)

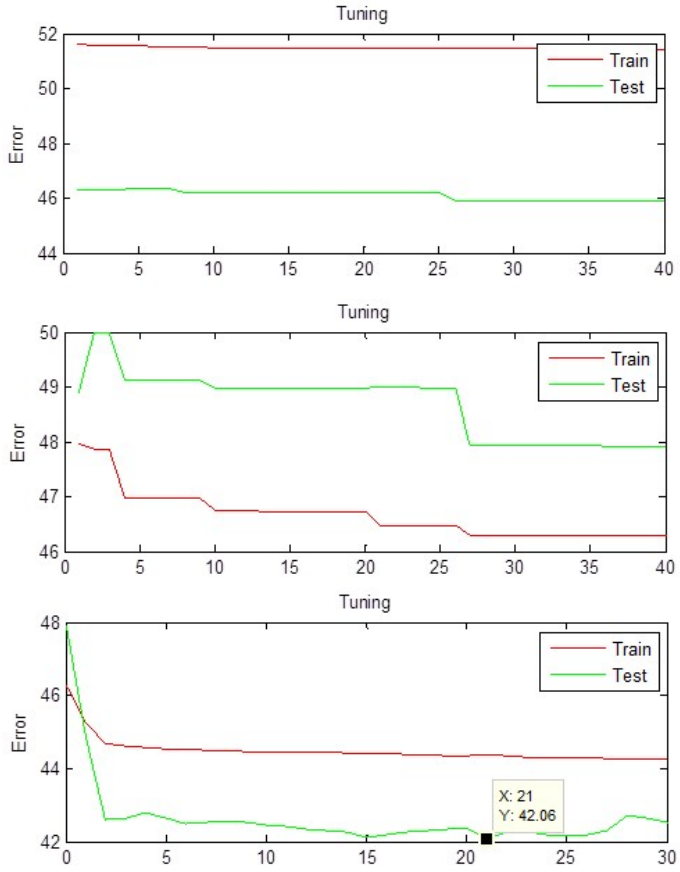

Рис. 3 Результати 3х етапів навчання нечіткої моделі мамдані

та побудована проста регресійна модель

Регресійна модель :

26.9132108550669
13.0865833162764
-2.29361141754382
-32.7859203758629
0.00793769287394669

RMSE = 49.77

Регресійна модель використовується виключно як базис для порівняння.

З рис. 2 випливає, що дерево рішень має 12 правил та середньоквадратичне відхилення 44.7, в свою чергу нечітка модель (рис. 3) значно покращує даний результат 42,6 маючи всього 25 неповних правил. Отже нечітка модель покращила результат на 4,7% та значно відрізняється від базису на 14.3%

Висновки

Встановлено, що запропонований інформаційна система проектування нечіткої моделі по типу бази знань Мамдані спроможна створити нечітку модель для передбачення кількості коментарів під публікаціями на «Сторінках Фейсбук». Продемонстровано покращенні результати отриманої нечіткої моделі Мамдані в порівнянні з конкуруючими моделями дерев рішень CART.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Buza, Krisztian. Feedback prediction for blogs. Data analysis, machine learning and knowledge discovery, Springer, Cham, 2014, P. 145-152.

2. Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. Comment volume prediction using neural networks and decision trees. IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015, 2015.

3. Штовба, С. Д., В. В. Мазуренко, Р. О. Тылец. Информационная технология нечеткой идентификации для синтеза точных, компактных и интерпретабельных баз знаний. Computer Sciences and Telecommunications 1, 2016, С. 8-22.

4. Штовба С.Д. Проектирование нечетких систем средствами MATLAB. Москва: Горячая линия – Телеком, 2007.

5. Breiman, L., J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Boca Raton, FL: CRC Press, 1984.

6. Wang, L. X., & Mendel, J. M. Generating fuzzy rules by learning from examples. Systems, Man and Cybernetics, IEEE Transactions on, 1992, Vol. 22, No. 6, P. 1414-1427.

7. Cordon O., Gomide E., Herrera E., Homannc E. Magdalena L. Ten years of genetic fuzzy systems: current framework and new trends. Fuzzy Sets and Systems, 2004, Vol. 141, P. 5–31.

Мазуренко Віктор Володимирович — IT фахівець, приватний підприємець, Вінниця, e-mail: viktor.mazurenko@gmail.com

Науковий керівник: ***Штовба Сергій Дмитрович*** — д-р техн. наук, професор кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця

Mazurenko Viktor V. — IT specialist, Entrepreneur, Vinnytsia, email : viktor.mazurenko@gmail.com

Supervisor: ***Shtovba Serhiy S.*** — Dr. Sc. (Eng.), Professor, Professor with the Computer Control Systems Department, Vinnytsia National Technical University, Vinnytsia