УДК 681.301

К.т.н., доцент Маслянко П.П., магістрант Бабченко О.О.

**Національний технічний університет України
«Київський політехнічний інститут»**

ДОСЛІДЖЕННЯ ТА РОЗРОБКА ПОКРАЩЕНОГО АЛГОРИТМУ ПОБУДОВИ ДЕРЕВА РІШЕНЬ ДЛЯ СИСТЕМИ СКОРИНГУ

Abstract

Pavlo P. Maslyanko, assoc. prof., PhD; Oleksiy Babchenko, student
Research and development of improved partitioning criteria for scoring systems

This article describes an algorithm and its improved modification for designing scoring system. Improved partitioning criteria corrects deficiencies. The ways for further research are proposed as well.

Вступ

Розробка покращеного критерію, який би дозволив побудувати більш адекватну систему скорингу, є дуже необхідним для власників кредитних спілок України [1]. Впровадження такої системи дозволяє:

- збільшити дохід кредитної спілки;
- зменшити ризик при видачі кредиту.

Постановка задачі

Мета роботи – розробити покращений критерій розбиття для системи скорингу кредитної спілки.

Об’єктом дослідження є система скорингу кредитної спілки.

Предметом дослідження є процес побудови дерева для системи скорингу.

Опис алгоритму побудови дерева

Нехай задано множину прикладів T, де кожний елемент цієї множини описується m атрибутами. Кількість прикладів в множині T будемо називати потужністю |T|. Нехай мітка класу приймає наступні значення C1, C2 … Ck.

Наша задача буде полягати в побудові ієрархічної класифікаційної моделі у вигляді дерева із множини прикладів T. Процес побудови дерева

буде відбуватись зверху вниз. Спочатку корінь дерева, за ним нащадки кореня и т.д.

Нехай ми маємо перевірку X, яка приймає n значень A1, A2 ... An. Тоді розбиття T по перевірці X дасть нам підмножини T1, T2 ... Tn, при X рівному відповідно A1, A2… An. Єдина доступна нам інформація – це те, яким чином розподілені класи в множині T та її підмножинах, отриманих при розбитті по X. Саме це ми і використаємо при означенні критерію.

Нехай $freq(C_j, S)$ – кількість прикладів із деякої множини S, які відносяться до класу Cj. Тоді ймовірність P того, що довільно обраний приклад із множини S буде належати до класу Cj

$$P = \frac{freq(C_j, S)}{|S|} \quad (1)$$

Згідно теорії інформації, кількість інформації яка міститься в повідомленні, залежить від її ймовірності [2].

$$\log_2\left(\frac{1}{P}\right) \quad (2)$$

Оскільки ми використовуємо логарифм з двійковою основою, то вираз (2) дає кількісну оцінку в бітах.

Вираз

$$Info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} * \log_2\left(\frac{freq(C_j, T)}{|T|}\right) \quad (3)$$

дає оцінку середньої кількості інформації, необхідної для визначення класу прикладу з множини T. Вираз (3) називається ентропією множини T.

Цю ж оцінку, але після розбиття T по X, дає наступний вираз:

$$\text{Info}_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * Info(T_i) \quad (4)$$

Тоді критерієм вибору атрибута буде наступна формула:

$$\text{Gain}(X) = \text{Info}(T) - \text{Info}_X(T) \quad (5)$$

Критерій (5) рахується для всіх атрибутів. Обирається атрибут, який максимізує даний вираз. Цей атрибут буде перевіркою в поточному вузлі дерева, а потім по цьому атрибуту відбувається подальша побудова дерева.

Такі ж судження можна застосувати до отриманих підмножин T1, T2 … Tn і продовжити рекурсивно процес побудови дерева. Застосувавши методику відсікання (pruning) можна усунути розлогі дерева, які з’являться в процесі побудови.

Покращений критерій розбиття

Розглянемо гіпотетичну задачу медичної діагностики, де один із атрибутів ідентифікує особу пацієнта. Оскільки кожне значення цього

атрибута унікальне, то при розбитті множини прикладів по цьому атрибуту отримаємо підмножини, які містять тільки по одному прикладу. Так як всі ці множини "одноприкладні", то і приклад відповідно відноситься до одного класу , тоді

$$Info_X(T) = 0 \tag{6}$$

Отже критерій (5) приймає своє максимальне значення, і саме цей атрибут буде обраний алгоритмом. На практиці не так часту зустрічаються подібні задачі, але необхідно передбачити і такі випадки.

Проблема розв'язується введенням деякої нормалізації. Нехай суть інформації повідомлення,яке відноситься до прикладу, вказує не на клас, а на вихід. Тоді, по аналогії з визначенням Info(T), ми маємо

$$split\ \text{info}(X) = -\sum_{i=1}^{n} \frac{T_i}{T} \log_2\left(\frac{T_i}{T}\right) \tag{7}$$

Вираз (7) оцінює потенційну інформацію, отриману при розбитті

$$gain\ ratio(X) = \frac{gain(X)}{split\ \text{info}(X)} \tag{8}$$

Нехай вираз (8) є критерієм обрання атрибута.
Очевидно, що атрибут, який ідентифікує пацієнта, не буде високо оцінений критерієм (8). Нехай ми маємо k класів, тоді чисельник виразу (8) максимально буде дорівнювати $\log_2(k)$ і нехай n – кількість прикладів навчальної виборки і одночасно кількість значень атрибута, тоді знаменник максимально дорівнює $\log_2(n)$. Якщо припустити, що кількість прикладів більша кількості класів, то знаменник зростає швидче, ніж чисельник, і, відповідно, вираз буде мати найбільше значення. Таким чином, ми можемо замінити критерій (5) на новий критерій (8), і знову же обрати той атрибут, який має максимальне значення по критерію.

Розглянутий нами алгоритм, передбачає, що для атрибута, обираємого в якості перевірки, існують всі значення. Тобто для будь-якого прикладу з навчальної виборки існує значення по цьому атрибуту.

Доки ми працюємо з синтетичними даними, то ми можемо "згенерувати" необхідні дані. Але буває що реальні дані далекі від ідеальних, і що часто зустрічаються пропущені, суперечливі та аномальні дані. Позначимо через U кількість невизначених значень атрибута A. Змінимо формули (3) і (4) таким чином, щоб враховувати тільки ті приклади, у яких існують значення по атрибуту A.

$$Info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T| - U} * \log_2\left(\frac{freq(C_j, T)}{|T| - U}\right) \tag{9}$$

$$\text{Info}_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T| - U} * Info(T_i) \tag{10}$$

В цьому випадку при розрахунку freq(Cj, T) враховуються тільки приклади з існуючими значеннями атрибута A.

Тоді критерій (4) можна переписати

$$\text{Gain}(X) = \frac{|T| - U}{|T|}\left(\text{Info}(T) - \text{Info}_X(T)\right) \quad (11)$$

Подібним чином змінюється критерій (8). Якщо перевірка має n вихідних значень, то критерій (8) розраховується як у випадку, коли вихідну множину розділено на n+1 підмножин.

Нехай тепер перевірка X з вихідними значеннями O1, O2 … On обрана на основі модифікованого критерію (10).

Необхідно вирішити, що робити з пропущеними даними. Якщо приклад із множини T з відомим Oi асоційований з підмножиною Ti, вірогідність того, що приклад з множини Ti рівна 1. Нехай тоді кожний приклад з підмножини Ti має вагу, яка вказує ймовірність того, що приклад належить Ti. Якщо приклад має значення по атрибуту A, тоді вага дорівнює 1, в іншому випадку приклад асоціюється з всіма множинами T1, T2 … Tn, із відповідними значеннями ваги $\frac{|T_1|}{|T|-U}, \frac{|T_2|}{|T-U|}, \ldots, \frac{|T_n|}{|T|-U}$.

Легко пересвідчитись, що

$$\sum_{i=1}^{n} \frac{|T_i|}{|T|-U} = 1 \quad (12)$$

Цей підхід можна сформулювати таким чином: передбачається, що пропущені значення по атрибуту розподілені пропорційно частоті появи існуючих значень.

Висновки

Розроблений алгоритм впроваджений в програмному забезпеченні “CuSol”, яке працює в кредитних спілках «Аккорд» та «Флагман».

В цілому, подібні алгоритми лише починають використовувати в Україні, тому доцільно накопичувати напрацювання у даній сфері, адже такі рішення користуватимуться великим попитом і дозволять істотно підвищити ефективність кредитного скорингу.

Література

1. І наукова конференція «Прикладна математика та комп’ютинг ПМК-2009”, Київ, 15-17 квітня 2009 р.: зб.тез/ред кол.: С.В. Сирота (гол.ред.) та ін. К.: НТУУ «КПІ», 2009. – с.25-29.
2. *К. Шеннон* / Работы по теории информации и кибернетике. М. Иностранная литература - 2000. – с.10-24.
3. *J. Ross Quinlan* «C4.5: Programs for Machine learning» – Morgan Kaufmann Publishers, 1993. – 115-117 p.