УДК 519.816

КЛАСТЕРИЗАЦІЯ БАЗИ КЛІЄНТІВ НА ОСНОВІ МЕТОДУ К-СЕРЕДНІХ

Струбицька І.П. [1)], Мельник І.Є. [2)]
Тернопільський національний економічний університет
[1)] к.т.н., доцент, [2)] магістрант

І. Постановка проблеми

У сучасних умовах ефективне управління є цінним ресурсом організації поряд з матеріальними, фінансовими та людськими ресурсами. Підвищення ефективності управлінської діяльності стає одним із напрямків удосконалення роботи підприємства в цілому.

У кожній організаційній структурі є низка проблем. Салон краси не є виключенням. У даний час в містах дуже популярні мережі салонів краси. Їх об'єднує одна проблема – відсутність будь-якої автоматизації процесів, в тому числі організації роботи із клієнтами.

Клієнти бувають постійними, що відвідують салон регулярно та новими – вперше в ньому, які мають високий чи середній фінансовий статус, а також малозабезпечені. Для кожного із них потрібно створити власний підхід, заохотити до отримання послуг, щоб отримати максимальний прибуток.

Ще однією проблемою вважається число об'єктів (клієнтів), яке є великим і затрудняє їх вивчення та прогнозування цих об'єктів. Наприклад, прогнозування категорії клієнтів (за віком та місцем проживання), які використовують найдорожчі процедури, тим самим приносять найбільший прибуток салону. Також до проблеми на системному рівні можна віднести відсутність апріорних відомостей класів, яким належать об'єкти досліджуваного набору даних.

Усунути подібні проблеми може одна із задач інтелектуального аналізу даних (Data Mining) – кластеризація.

ІІ. Мета роботи

Мета роботи полягає у побудові профілів клієнтів за допомогою алгоритму кластеризації k-середніх шляхом врахування схожої поведінки об'єктів у плані частоти відвідування послуг та проведення оцінки сегментів, інформація яких в подальшому використовуватиметься для створення індивідуального підходу до клієнта.

ІІІ. Застосування алгоритму k-середніх для кластеризації бази клієнтів

Кластеризація клієнтської бази – це розподіл відвідувачів по групах, які відповідають стійким ознакам, так званих «ознак сегментації» (рис. 1). Вибір ознак сегментації залежить від мети кластеризації. Як правило, в якості цих ознак використовують:

- географічні характеристики (регіональний поділ);
- характеристики споживчої поведінки (інтенсивність відвідування послуг, отримані суми за обслуговування);
- демографічні ознаки (стать, вік).

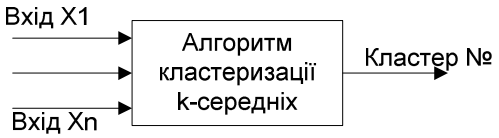

Рисунок 1 – Задача кластеризації клієнтів

Множина об'єктів $x = (x_1, ..., x_n)$ – це набір вхідних даних. Кожний і-й об'єкт з X визначається як $x = (x_1, i, ...x_i, d)$. Часто X представляють в формі матриці характеристик розмірності $n \times d$. Кластер – це підмножина «близьких один до одного» об'єктів X. Кластери бувають непересічні (non-overlapping) та пересічні (overlapping). Відстань $d(x_i, x_j)$ між об'єктами x_i та x_j – це результат застосування вибраної метрики в просторі характеристик [1].

Для підвищення ефективності побудови профілів клієнтів шляхом їхньої споживчої поведінки в плані частоти відвідування послуг при сегментації клієнтської бази, як правило, використовують характеристику споживчої поведінки: частоту та величину середнього чеку. Також характеристику споживчої поведінки застосовують і для оцінки сегментів, які дають найбільше та найменше доходів. Іноді в якості додаткових параметрів враховують загальний час взаємодії клієнта із салоном краси чи давність замовлення та відвідування послуги.

Кожна група може бути розбита на декілька підгруп для сегментації клієнтської бази оцінки вікової категорії клієнтів. Для досягнення цього використовують характеристику, що представляє собою демографічну ознаку (дату народження відвідувача).

Алгоритм кластеризації k-means включає наступні кроки [2]:

1. Задається число кластерів *k*, яке повинно бути сформоване з об'єктів вихідної вибірки.
2. Випадковим чином вибираються *k*-ті записи, які будуть початковими центрами (точками), з яких потім отримаємо кластер. Кожен такий запис являє собою так званий «ембріон» кластера, який складається тільки з одного елемента.
3. Для кожного запису вихідної вибірки визначається найближчий до неї центр кластера.
4. Обчислюється нове положення центрів – центроїдів. Це робиться шляхом визначення середнього значення кожної ознаки всіх записів кластера (1):

$$(x, y) = \left(\frac{x_1 + x\ + x_3 + ... + x_i}{N} \right), \left(\frac{y_1 + y_2 + y_3 + ... + y_i}{N} \right), \quad (1)$$

де $(x_1 + y_1), (x_2 + y_2), (x_3 + y_3), ..., (x_i + y_i)$ – набір ознак; N – кількість наборів ознак.

5. Старий центр кластера зміщується в його центроїд.

Кроки 3 та 4 повторяються доти, поки виконання алгоритму не буде перервано або поки не буде виконана угода відповідно з деякими критерієм схожості.

Вихід з алгоритму робиться тоді, коли межі кластерів і положення центроїдів перестануть змінюватись, тобто на кожній ітерації в кожному кластері залишається один і той же набір записів. Алгоритм k-means зазвичай знаходить набір стабільних кластерів за кілька десятків ітерацій.

Критерій схожості розраховується як сума квадратів помилок між центроїдом кластера та всіма вхідними його записами (2):

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2, \quad (2)$$

де $p \in C_i$ – довільна точка даних, яка належить кластеру C_i; m_i – центроїд даного кластеру.

Вхідними даними для кластеризації задано інформацію про клієнтів: прізвище, ім'я, по-батькові, величина середнього чеку, кількість відвідувань салону. Параметрами для проведення сегментації клієнтської бази салону краси є такі характеристики: величина середнього чеку та інтенсивність відвідування клієнтом салону.

Результатом роботи алгоритму є утворення чотирьох груп відповідно до заданих параметрів кластеризації:

- випадкові відвідувачі – клієнти, які замовили за весь час одну послугу, сума середнього чеку – мінімальна по клієнтській базі (кластер №1);
- відвідувачі – клієнти, що отримали декілька послуг протягом певного періоду часу (кластер №2);
- постійні відвідувачі, які періодично користуються послугами салону краси (кластер №3);
- прихильники – активні клієнти, величина середнього чеку є вищою середнього рівня (кластер №4).

Приклад сформованого кластеру №3 із даними про відвідувачів наведено у таблиці 1.

До кластеру ввійшли записи про клієнтів, у яких кількість відвідувань знаходиться у межах від 7 до 11 разів на рік з величиною середнього чеку від 250 до 350 грн.

Таблиця 1

Результат кластеризації

№п/п	Прізвище	Ім'я	По-батькові	Середній чек (грн)	Кількість відвідувань
1	Москаль	Наталя	Михайлівна	257	7
3	Кутна	Христина	Петрівна	251	8
4	Кушнір	Анастасія	Іванівна	325	11
7	Мелих	Ярина	Андріївна	282	9
10	Попович	Дарина	Сергіївна	255	8

Висновок

За допомогою алгоритму k-середніх проведено сегментацію клієнтської бази салону краси, у результаті чого, з використанням параметрів споживчої поведінки (частоти відвідування салону та величини середнього чеку), виділились чотири кластери.

Таку кількість кластерів обрано тому, що двох або трьох кластерів недостатньо і кластеризація буде нечіткою та призведе до втрати інформації кожного із об'єктів. Більше десяти кластерів – забагато і важко тримати в короткій пам'яті стільки підмножин.

Список використаних джерел

1. Чубукова И.А. Data Mining: Учебное пособие / И.А. Чубукова. М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.
2. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям: Учебное пособие, 2-е издания, испр. / Н.Б. Паклин, В.И. Орешков – СПб.: Изд-во Питер, 2013. – 704 с.

УДК 004.896

WEB-ОРІЄНТОВАНА СИСТЕМА КОНТРОЛЮ ЗНАНЬ УЧНІВ З ВИКОРИСТАННЯМ СЕМАНТИЧНОЇ МОДЕЛІ

Струбицька І.П.[1)], Тимець В.І.[2)]

Тернопільський національний економічний університет

[1)] к.т.н., доцент; [2)] магістрант

І. Постановка проблеми

Контроль та оцінка знань, умінь і навичок учнів є невід'ємним структурним компонентом навчального процесу. Процес навчання – це система із внутрішніми взаємозв'язками між їх компонентами. В умовах сучасного стрімкого розвитку інтелектуальних технологій, для загальноосвітніх закладів, найкраще підходить, методологія контролю успішності учнів на основі тестових завдань.

Проблема оцінки якості навчання за допомогою тестів завжди розглядалась як важлива і одночасно "небезпечна". "Небезпечність" педагогічного тестування полягає в тому, що будь яка необґрунтованість або поспішність у висновках може призвести до випадкових висновків, із необґрунтованим рекомендаціям, сумнівним педагогічним наслідкам.

Отже, актуальними залишаються задачі вдосконалення технічного й інформаційного забезпечення, яке використовуватиметься у навчальних закладах, та відповідно підвищить якість контролю та перевірки знань учнів.

ІІ. Мета роботи

Метою роботи є підвищення швидкодії, програмної системи контролю знань учнів на основі семантичної моделі, яка побудована на принципах генерації тестових завдань за допомогою понятійно-тезисного підходу. Ця модель базується на додання у алгоритм автоматичного генерування набір "поняття-теза" із семантичного фрагменту.