4. John, W. S. A Bayesian approach to diagnosis and prognosis using built-in test [Text] / W. S. John, A. K. Mark // IEEE Transactions on instrumentation and measurement. – 2005. – Vol. 54, Issue 3. – P. 1003–1018. doi: 10.1109/tim.2005.847351
5. Jin, L. Accurate testing of analog-to-digital converters using low linearity signals with stimulus error identification and removal [Text] / L. Jin, K. Parthasarathy, T. Kuyel, D. Chen, L. G. Randall // IEEE Transactions on instrumentation and measurement. – 2005. – Vol. 54, Issue 3. – P. 1188–1199. doi: 10.1109/tim.2005.847240
6. Skoczowski, S. A Simple Identification Method for the Order of the Strejc Model and its Application to Autotuning [Text] / S. Skoczowski, A. Osadowski // IFAC Intelligent components and instruments for control applications, 2nd IFAC Symposium. Budapest, Hungary, 1994. – P. 319–325. doi: 10.1016/b978-0-08-042234-3.50054-0
7. Stieber, M. T. Instrumentation architecture and sensor fusion for system control test [Text] / M. T. Stieber, G. Vukovich. // IEEE Transactions on instrumentation and measurement. – 1998. – Vol. 47, Issue 1. – P. 108–113. doi: 10.1109/19.728801
8. Григоренко, І. В. Дослідження впливу нелінійності зміни вхідного сигналу на динамічну похибку вимірювального перетворювача під час проведення тестового контролю [Текст] / І. В. Григоренко // Вестник НТУ «ХПИ». – 2008. – №. 57. – С. 50–57.
9. Григоренко, І. В. Розвиток тестових методів підвищення точності електричних компенсаційних вимірювальних перетворювачів у динамічних режимах [Текст]: дис. ... канд. техн. наук / І. В. Григоренко. – Харків, 2010. – 224 с.
10. Опришкіна, М. І. Тестовий метод підвищення точності електричних давачів з нелінійними функціями перетворення [Текст]: дис. ... канд. техн. наук / М. І. Опришкіна. – Харків, 2013. – 186 с.

Запропоновано оригінальну модифікацію методу k-найближчих сусідів для вирішення задач машинного навчання у кредитному скорингу, а саме розроблено варіанти методу k-plus-найближчих сусідів на множинах дискретних значень вхідних змінних для вирішення задачі ймовірнісної бінарної класифікації відносно бінарної цільової змінної. Наведено частину програмної реалізації запропонованого методу мовою структурованих запитів, використовуючи віконні функції

Ключові слова: метод k-найближчих сусідів, кредитний скоринг, бінарна класифікація, мова структурованих запитів

Предложена оригинальная модификация метода k-ближайших соседей для разрешения задач машинного обучения в кредитном скоринге, а именно разработаны варианты метода k-plus-ближайших соседей на множествах дискретных значений входящих переменных для разрешения задачи вероятностной бинарной классификации относительно бинарной целевой переменной. Приведена часть программной реализации предложенного метода на языке структурированных запросов, используя оконные функции

Ключевые слова: метод k-ближайших соседей, кредитный скоринг, бинарная классификация, язык структурированных запросов

УДК 519.237.8 : 681.518.25
DOI: 10.15587/1729-4061.2015.43730

РОЗРОБКА МЕТОДУ K-PLUS-НАЙБЛИЖЧИХ СУСІДІВ ДЛЯ ЗАДАЧ МАШИННОГО НАВЧАННЯ КРЕДИТНОГО СКОРИНГУ

О. М. Солошенко
Аспірант
Кафедра математичних
методів системного аналізу
Навчально-науковий комплекс
«Інститут прикладного системного аналізу»
Національний технічний університет України
«Київський політехнічний інститут»
пр. Перемоги, 37, м. Київ, Україна, 03056
E-mail: soloshenko_s@ukr.net

1. Вступ

Методи математичного та статистичного моделювання мають надзвичайно широке, важливе, ефективне та успішне застосування в області фінансового ризик-менеджменту [1]. Надзвичайно велика роль в області фінансового ризик-менеджменту відводиться вивченню та моделюванню кредитних ризиків [1]. Управління кредитними ризиками передбачає попередню оцінку кредитоспроможності потенційних клієнтів з метою забезпечення прийнятного рівня ризику у процесі кредитування [1]. Кредитний скоринг – це методологія оцінювання кредитоспроможності потенційних позичальників у ризик-менеджменті [2–5]. Скоринг – це методологія оцінювання кредитоспроможності або майбутньої поведінки на рівні клієнтів або договорів, як потенційних, так і існуючих, тому існує багато категорій скорингу: кредитний (аплі-

каційний) скоринг, поведінковий скоринг, скоринг виявлення та попередження шахрайства, колекторський скоринг, інші численні категорії скорингу [2–4, 6]. Скорингові моделі також називають скоринговими картами (scorecards) [2–7]. Методологія побудови скорингових моделей тісно пов'язана з методами машинного навчання [8], опосередковано та на практиці – з теорією реляційних баз даних [9, 10] (при побудові вибірок, впровадженні моделей, дослідженні та моніторингу їх стабільності та предикативної сили), та напряму – з поняттям інтелектуального аналізу даних (data mining) [2–5, 11]. Одним з найбільш популярних методів, на противагу логістичній регресії [2–5, 12, 13], у машинному навчанні для вирішення задачі класифікації є метод k-найближчих сусідів (k-nearest neighbor method) або метод виводу на основі пам'яті (memory-based reasoning) [2–4], що в термінах машинного навчання ще називається навчанням на основі пам'яті (memory-based learning) [8], який може застосовуватись як до побудови скорингових моделей, так і до проміжного етапу аналізу відхилених заявок (reject inference) з метою включення відхилених заявок в модель аплікаційного скорингу [2, 4].

Класичний метод k-найближчих сусідів визначається у довільному метричному просторі змінних без деталізації та без висвітлення таких можливих питань як: способи нормування змінних, способи вибору метрики серед множини можливих метрик, способи урахування категоріальних змінних, вибір оптимального (а не фіксованого) значення кількості сусідів та критерії такого вибору, способи призначення ваг змінним відносно цільової змінної, оцінка узгодженості цільових класів відносно метричного простору вхідних змінних, оцінка на наявність «викидів», аспекти ймовірнісної класифікації, ситуації з близькими рівновіддаленими множинами елементів, зважування результату відносно відстаней до найближчих сусідів [3, 8]. Тому актуальними та практично цінними з точки зору ризик-менеджменту є питання детального дослідження та модифікацій методу машинного навчання на основі пам'яті за допомогою методу k-найближчих сусідів саме у задачах кредитного скорингу [3], враховуючи існуючі розробки та ключові поняття власне в області скорингу, де ймовірнісна бінарна класифікація посідає ключову роль у методології [2–7].

Ще одним важливим аспектом актуальності дослідження є спосіб збереження великих масивів даних у вигляді таблиць сучасних систем керування базами даних (СКБД), що відповідають реляційній моделі управління даними [9, 10], тому актуальним є питання використання можливостей мови структурованих запитів (Structured Query Language, SQL) [9, 10] для вирішення задач моделювання та аналізу даних без використання сторонніх додаткових програмних засобів.

2. Аналіз літературних даних та постановка проблеми

Сучасний стрімкий прогрес в області сучасного ризик-менеджменту [1], зокрема в галузі кредитного скорингу [2–7], забезпечується швидким розвитком методів кількісного аналізу [1], розвитком інформаційних технологій [2, 7], розвитком методів інтелектуального аналізу даних (data mining) [2–5, 11], зокрема статистичних та нестатистичних методів побудови скорингових моделей [3].

Основні, але далеко не всі, сучасні методи побудови скорингових моделей можна розділити таким чином [3]:

1) статистичні методи побудови скорингових карт:

1.2) лінійна регресія;

1.3) логістична регресія (нелінійна) [2–5, 12, 13];

1.4) пробіт-регресія (нелінійна);

1.5) дерева рішень (рекурсивний підхід розбиття);

1.6) методи найближчих сусідів:

1.1.5) метод найближчого сусіда;

1.1.6) метод k-найближчих сусідів;

2) нестатистичні методи побудови скорингових карт:

2.1) лінійне програмування;

2.2) цілочисельне програмування;

2.3) нейронні мережі;

2.4) генетичні алгоритми;

2.5) експертні системи;

3) альтернативні змішані методи побудови скорингових карт:

3.1) байєсівські мережі та графічні моделі [11];

3.2) моделі аналізу виживання.

На ринку інформаційних технологій присутні численні рішення у вигляді програмних додатків та статистичних пакетів, що дозволяють здійснювати моделювання зокрема кредитних ризиків: рішення та мова програмування компанії SAS® Institute Inc. [2, 4, 13], проект та вільна мова програмування R [13], статистичний пакет IBM® SPSS® Software з внутрішньою мовою програмування [4, 13] та інші рішення.

Найбільш популярним методом [2, 4, 5] побудови скорингових моделей є логістична регресія [2–5, 12, 13], однак великий інтерес щодо побудови сучасних скорингових моделей для оцінювання кредитоспроможності становить саме метод k-найближчих сусідів (k-nearest neighbor method) [3]. Це пояснюється зокрема концептуальною простотою інтерпретації способу класифікації як машинного навчання на основі пам'яті (memory-based learning) [8] та численними перевагами, що стосуються, наприклад, простоти динамічного он-лайн оновлення моделі через додавання нових елементів вибірки (спостережень) у базу пам'яті та виключення найстаріших елементів (спостережень) з бази пам'яті [3]. Також великий інтерес до методу k-найближчих сусідів обумовлюється питанням вибору оптимальної метрики та відносно низьким ступенем вивченості та експериментального застосування у задачах скорингу [3], особливо при використанні категоріальних або дискретизованих змінних. До основних недоліків даного непараметричного методу належать зокрема лише висока обчислювальна складність при оцінюванні множин елементів, складність вибору метрики, складність регулювання та перекалібрування моделі [3].

Аналіз сучасного джерела [14] свідчить про велику популярність програмних реалізацій класичного методу k-найближчих сусідів, а особливо його модифікації – нечіткого методу k-найближчих сусідів (Fuzzy k-Nearest Neighbor, Fuzzy kNN), де ймовірнісне значення прогнозу присвоюється в залежності від відстаней до найближчих сусідів, тобто за допомогою зважування значень фактичних класів, де вага є

зваженою степеневою функцією від'ємного степеню від відстані згідно з [14] (ідентична формула була описана, наприклад, ще в джерелі [15]). Виникає низка закономірних зауважень та невирішених проблем відносно описаного в [14] нечіткого методу, наприклад: коректне опрацювання нульової відстані, хоча, якщо застосовувати границю для виразу зважування, то можна отримати одиничні значення ваг, але такий підхід дуже чутливий до статистичних «викидів», що знаходяться поблизу вектора, що класифікується, а це становить проблему даного методу; порядок врахування рівновіддалених від вектора груп; важливе питання відносно рівності співвідношення класів на множинах фактів та прогнозів, що забезпечувалося б звичайним, а не зваженим по відстані, усередненням, яке крім того не наскільки чутливе до близьких статистичних «викидів»; питання оптимального вибору k; дослідження інтегральних критеріїв оцінювання якості прогнозів, що притаманні методології кредитного скорингу і т. д. Також у якості постановки задачі може виступати програмна реалізація мовами покоління четвертого покоління (в [14] програмна реалізація представлена мовою третього покоління).

Аналіз сучасного джерела [16] може бути використаний для вибору однієї з можливих метрик для проведення експериментів на числових даних (відстань Евкліда, Мінковського, Махаланобіса), також у даному джерелі описується новітній горизонт застосування будь-яких вдосконалень та форм методу k-найближчих сусідів – аналіз знакових послідовностей та текстових даних на близькість – подібність (міра Хеммінга і т. д.). Робота [16] підтверджує актуальність досліджень та модифікацій методу машинного навчання на основі пам'яті, однак, у даному джерелі також не враховується, наприклад, що не завжди можливо обрати однозначно рівно *k* елементів при існуванні рівновіддалених груп елементів, окрім того, частина зауважень описана при аналізі джерела [14] також має місце.

3. Ціль та задачі дослідження

Проведені дослідження ставили за мету усунути недоліки класичного методу k-найближчих сусідів, включаючи відсутність конкретики та деталізації особливостей застосування машинного навчання на основі пам'яті при використанні категоріальних та дискретизованих змінних в умовах можливості виникнення ситуацій з рівновіддаленими групами найближчих сусідів відносно елемента, що класифікується, а також за мету була поставлена розробка способів зниження обчислювальної складності методу за допомогою подальших вдосконалень пропонованої модифікації методу.

Для досягнення поставленої мети вирішуються такі задачі:

– застосування понять методології кредитного скорингу для утворення метричного простору на основі категоріальних (в т.ч. дискретизованих) змінних з використанням перетворень відносно цільової змінної;

– формалізація вирішення проблеми рівновіддалених груп елементів відносно елемента, що класифікується;

– формулювання та формалізація пропонованого методу k-plus-найближчих сусідів;

– наведення ключових можливостей мови структурованих запитів SQL щодо реалізації пропонованого методу;

– формалізація можливих концептуально значимих вдосконалень методу щодо зменшення обчислювальної складності та наведення відповідних можливостей мови структурованих запитів;

– провести порівняльний аналіз результатів базового методу k-plus-найближчих сусідів для декількох значень вхідного параметру з результатами методу моделювання за допомогою логістичної регресії.

4. Методи вдосконалення машинного навчання на основі пам'яті та модифікація методу k-найближчих сусідів

4. 1. Методика створення метричного простору для категоріальних та дискретизованих змінних з використанням методології кредитного скорингу

Суть класичного методу k-найближчих сусідів (k-nearest neighbor method) або навчання на основі пам'яті (memory-based learning) [3, 8] полягає у такій формалізації правила класифікації по принципу значення більшості (majority) згідно з формулами (1)–(2):

$$y^* = \begin{cases} 1, \sum_{i=1}^{k} \frac{y_i}{k} > \frac{1}{2}; \\ 0, \sum_{i=1}^{k} \frac{y_i}{k} \le \frac{1}{2}, \end{cases} \quad (1)$$

де y^* прогнозоване значення бінарного класу (цільова змінна), y_i – фактичне бінарне значення класу цільової змінної і-го найближчого сусіда, k – кількість найближчих сусідів (для спрощення можна вважати, що це непарне натуральне число, з метою уникнення ситуації рівного співвідношення), при цьому найближчі сусіди заздалегідь визначаються згідно з метрикою у просторі вхідних серед скінченної множини векторів лише вхідних змінних [3, 8]:

$$\begin{cases} \mathbf{x}_1 = \arg\min_{\mathbf{x} \in X} d(\mathbf{x}, \mathbf{x}^*); \\ \forall i \in \{2,..k\} : \mathbf{x}_i = \arg\min_{\mathbf{x} \in X \setminus \bigcup_{j=1}^{i-1} \mathbf{x}_j} d(\mathbf{x}, \mathbf{x}^*), \end{cases} \quad (2)$$

де $\mathbf{x}$ – довільний вектор простору лише вхідних змінних (вектор спостереження) навчальної вибірки, X – скінченна множина векторів спостережень навчальної вибірки для вхідних змінних, $\mathbf{x}^*$ – вхідний вектор значень вхідних змінних, що класифікується (до якого застосовується прогноз), d – метрика визначена на просторі векторів вхідних змінних.

Таким чином суть алгоритму полягає у присвоєнні елементу, що класифікується, значення локальної статистичної моди у якості прогнозованого цільового класу. На рис. 1 зображено приклад застосування методу трьох найближчих сусідів у двовимірному просторі [8], коли вектору прогнозовано одиничний клас згідно з методом.

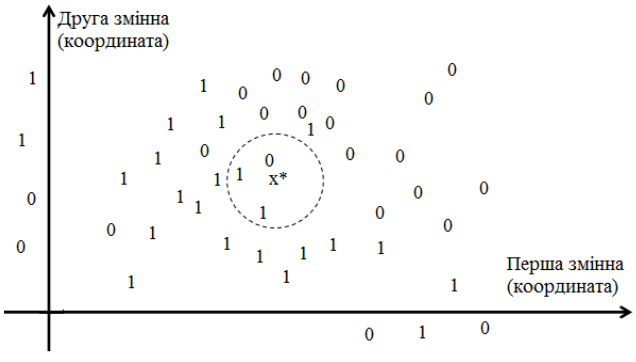

Рис. 1. Приклад застосування методу k-найближчих сусідів у двовимірному просторі при k=3 (результат прогнозу: $y^*=1$)

Даний метод відносно просто застосовний, наприклад, для побудови скорингових моделей у просторі вхідних неперервних змінних, якщо можливе деяке нормування неперервних змінних, що забезпечує незалежність від масштабу вхідних змінних, де бінарна цільова змінна означає індикатор кредитоспроможності, однак багато питань щодо налаштування метрики та вибору числа найближчих сусідів залишаються відкритими. Зокрема, також залишається невирішеним питання врахування можливих спостережень, що рівновіддалені від $\mathbf{x}^*$ на відстані k-го найближчого сусіда, а також залишається відкритим просте питання ймовірнісної класифікації. Однак найголовнішими невирішеними питаннями з точки зору скорингового моделювання, що виступають у якості постановки даного дослідження, окрім способів ймовірнісного висновку, також лишаються способи врахування категоріальних змінних, що власне й часто призводитимуть до ситуацій з рівновіддаленими точками, та власне критерії оптимальності кількості найближчих сусідів відносно обраної метрики на множині спостережень навчальної вибірки.

Класична методологія кредитного скорингу передбачає два основні методи числового оперування категоріальними (в т. ч. дискретизованими) змінними, що значно відрізняються, головне, через наявність необхідності залучення цільової бінарної змінної або необхідності заміни оригінальної множини змінних [2].

Перший метод, менш популярний та ефективний, полягає у розбитті кожної категоріальної змінної на множину бінарних змінних, що відповідають окремим значенням окремої категоріальної змінної. Даний метод у статистиці ще називають методом створення фіктивних змінних (dummy variables) [2]. Недоліком даного методу є заміна кожної категоріальної змінної на множину бінарних змінних, що відповідають можливим значенням категоріальної змінної, що призводить до збільшення сукупної кількості змінних в процесі моделювання. Недоліком з точки зору методу навчання на основі пам'яті є можливість рівноцінного входження всіх суто бінарних координат в формулу метрики без врахування наявності факту взаємозв'язку значення бінарної координати з цільовою змінною.

Другий методом інтерпретації категоріальних як числових значень є метод перетворення кожного категоріального значення окремої змінної в вагу значення змінної (Weight Of Evidence, WOE), тобто вагу атрибуту змінної [2, 5, 6], що входить як до формули розрахунку відстані Кульбака-Лейблера [5], так і до формули розрахунку індексу значення інформації (Information Value, IV) [2, 5, 6]. Для кожного значення категоріальної змінної вага значення змінної обчислюється як натуральний логарифм від відношення долі одиничних («хороших») значень цільової змінної, що відповідають значенню категоріальної змінної, відносно всіх одиничних («хороших») значень цільової змінної, до долі нульових («негативних») значень цільової змінної, що відповідають значенню категоріальної змінної, відносно всіх нульових («негативних») значень цільової змінної [2, 5, 6]:

$$WOE_{ij} = \ln\left(\frac{g_{ij}}{b_{ij}}\right), \quad (3)$$

де i – номер змінної, j – номер можливого значення (категорії) конкретної змінної, g_{ij} – доля одиничних («good») значень цільової змінної, що відповідають j-й категорії i-ї змінної, відносно всіх одиничних («good») значень цільової змінної, – доля нульових («bad») значень цільової змінної, що відповідають j-й категорії i-ї змінної, відносно всіх нульових («bad») значень цільової змінної, тобто:

$$g_{ij} = \frac{G_{ij}}{\sum_{k=1}^{m_i} G_{ik}}, \quad (4)$$

$$b_{ij} = \frac{B_{ij}}{\sum_{k=1}^{m_i} B_{ik}}, \quad (5)$$

тут m_i – кількість категорій i-ї змінної, G_{ij} – кількість одиничних («good») значень цільової змінної, що відповідають j-й категорії i-ї змінної, B_{ij} – кількість нульових («bad») значень цільової змінної, що відповідають j-й категорії i-ї змінної.

Очевидно, що виконується така тотожність:

$$\forall i \in \{1,..n\}: \sum_{j=1}^{m_i} g_{ij} = \sum_{j=1}^{m_i} b_{ij} = 1, \quad (6)$$

де n – кількість змінних, що окремо аналізуються.

Тоді формула обчислення індексу значення інформації [2, 5, 6] (або інформаційної статистики [3]) для кожної змінної має такий вигляд:

$$IV_i = \sum_{j=1}^{m_i} (g_{ij} - b_{ij}) \ln\left(\frac{g_{ij}}{b_{ij}}\right) = \sum_{j=1}^{m_i} (g_{ij} - b_{ij}) WOE_{ij}, \quad (7)$$

де IV_i – індекс значення інформації i-ї змінної.

Як наслідок, кожне спостереження навчальної вибірки (без відображення в списку координат власне цільової змінної) можна описати вектором конкретних ваг категорій n змінних:

$$\mathbf{x}_p^T = \left(WOE_{1j_1(p)}\ WOE_{2j_2(p)} \ldots WOE_{ij_i(p)} \ldots WOE_{nj_n(p)}\right), \quad (8)$$

де $j_i(p)$ – фактичний номер категорії для i-ї змінної вектору $\mathbf{x}_p$.

У дослідженні у якості метрики запропоновано використовувати класичну відстань Евкліда (Euclidean distance) [2, 3, 8] саме у просторі векторів ваг категорій змінних розмірності *n*:

$$d\left(\mathbf{x}_p,\mathbf{x}_r\right)=$$

$$\|\mathbf{x}_p-\mathbf{x}_r\|=\sqrt{\sum_{i=1}^{n}(\mathrm{WOE}_{ij_i(p)}-\mathrm{WOE}_{ij_i(r)})^2}\,, \qquad (9)$$

де d – метрика.

Застосування даної метрики опосередковано (через попереднє обчислення WOE_{ij}) залежить від бінарної цільової змінної, тому має бути ефективним. Тим паче, запропонована метрика узгоджується з поняттям індексу значення інформації, оскільки високе значення індексу значення інформації IV_i зазвичай свідчить про високу дисперсію ваг категорій i-ї змінної $\mathrm{Var}\left(\left\{\mathrm{WOE}_{ij_i(p)}\right\}_{p=1}^{N}\right)$ на навчальній вибірці розміру N та відповідно про достатньо високу різницю («розмах») між максимальним та мінімальним значенням $\max\limits_{p\in\{1..N\}}\mathrm{WOE}_{ij_i(p)}-\min\limits_{p\in\{1..N\}}\mathrm{WOE}_{ij_i(p)}$, що означає, що саме змінні з високим значенням індексу інформації грають ключову роль у варіації запропонованої метрики. Також легко показати, що при низькій варіації ваг категорії змінної її значення близьке до нуля, оскільки виконується наступна рівність [17], що пов'язує вагу категорії змінної з долею нульового («bad») класу по окремій категорії та взагалі на всій вибірці (або середньозваженій по кількості долі):

$$\mathrm{WOE}_{ij}=\ln\left(\frac{\frac{B}{B+G}}{1-\frac{B}{B+G}}\right)-\ln\left(\frac{\frac{B_{ij}}{B_{ij}+G_{ij}}}{1-\frac{B_{ij}}{B_{ij}+G_{ij}}}\right)=$$

$$=\ln\left(\frac{p(B)}{1-p(B)}\right)-\ln\left(\frac{p_{ij}(B)}{1-p_{ij}(B)}\right), \qquad (10)$$

де B – загальна кількість спостережень з нульовим («bad») класом (знаменник відношення (5)), G – загальна кількість спостережень з одиничним («good») класом (знаменник відношення (4)), p(B) – доля нульового класу на всій вибірці, $p_{ij}(B)$ – доля нульового класу по j-й категорії i-ї змінної.

Таким чином, навіть при включенні в простір змінних з низьким індексом значенням інформації вплив таких змінних буде низьким, тобто важливою властивістю запропонованої метрики є її чутливість до предикативної сили змінних.

Також суть даної метрики можна описати спробою перемістити більшість одиничних значень у частину простору, де всі координати позитивні, а більшість нульових значень – де всі координати негативні.

4. 2. Методика вирішення проблеми рівновіддалених груп елементів відносно елемента, що класифікується

Класичний метод k-найближчих сусідів [3, 8] не дає рекомендацій відносно вирішення ситуацій, коли, наприклад, починаючи з якогось найближчого сусіда йде велика група великої кількості рівновіддалених елементів, що не можуть поміститися разом з попередніми найближчими сусідами в число k.

У рамках пропонованого методу пропонується використовувати такий алгоритм, що дозволяє розглядати щонайменше (а не точно рівно) k сусідів:

1) відсортувати множину, що представляє собою склад пам'яті, по зростанню відстані від елемента, що класифікується, при цьому внутрішнє сортування елементів рівновіддалених груп можна проводити випадковим чином;

2) обрати перші окремі k елементів згідно з сортуванням по першому пункту;

3) доповнити k елементів всіма елементами, що перебувають на такій же відстані від елемента, що класифікується, як останній обраний (k-й) найближчий сусід, якщо такі елементи існують.

4. 3. Формулювання та формалізація базового методу k-plus-найближчих сусідів та його вдосконалення з використанням пропонованого критерію оптимальності

Формалізація пропонованого базового методу k-plus-найближчих сусідів та його застосування:

1) розрахувати значення ваг категорій для всіх змінних, що використовуються при проектуванні метричного простору навчальної вибірки, згідно з формулою (3) або (10);

2) визначити метрику на просторі ваг категорій змінних згідно з формулою (9);

3) задати значення k;

4) для кожної окремої категорії кожного окремого вхідного елемента з множини елементів, що класифікуються, присвоїти значення ваги цієї окремої категорії відповідне такому ж значенню категорії в навчальній вибірці (тобто категорії елемента, що класифікується, перетворюються в числа – ваги категорій змінної – згідно з відповідністю «категорія-WOE» в навчальній вибірці);

5) для кожного елемента з множини елементів, що класифікуються, отримати щонайменше k сусідів з навчальної вибірки згідно з алгоритмом методики описаної в підрозділі 4.2 та метрикою, що визначена в п. 2;

6) для кожного елемента з множини елементів, що класифікуються, розрахувати долю одиничних («good») елементів в числі обраних щонайменше *k* сусідів згідно з п. 5, що й буде дорівнювати прогностичній ймовірності належності до одиничного («good») класу. Таким чином визначається спосіб виводу ймовірнісного висновку. Формула (1) заміняється пропонованою формулою (11):

$$y^*=\sum_{i=1}^{k^+(\mathbf{x}^*)}\frac{y_i}{k^+(\mathbf{x}^*)}, \qquad (11)$$

де k ( )≥k – фактична кількість найближчих сусідів (не менше k) для вектору $\mathbf{x}^*$.

Також класичний метод k-plus-найближчих сусідів (навіть з детермінованим висновком) не встановлює критеріїв вибору k, виходячи з навчальної вибірки [3, 8]. У вдосконаленні методу k-plus-найближчих сусідів у якості критерію оптимальності вибору k пропо-

нується використання індексу Джині [2–4, 7] , але за допомогою перехресної валідації (cross-validation) [3] на навчальній вибірці за допомогою методу виключення одного елементу – методу «leave-one-out» [3].

Згідно з [7] в інтегральній формі формулу обчислення індексу Джині можна записати таким чином відносно кумулятивних функцій розподілу прогнозів на двох підмножинах, що відповідають двом класам:

$$GINI = \frac{\int\limits_{y^* \in Y^*} F_B(y^*)dF_G(y^*) - \frac{1}{2}}{\frac{1}{2}}, \qquad (12)$$

де Y^* відображає множину значень ймовірнісних прогнозів на множині, що досліджується на якість прогнозу в порівнянні з фактичними класами, F_G – емпірична кумулятивна функція розподілу прогнозів на множині фактичного одиничного («good») класу, F_B – емпірична кумулятивна функція розподілу прогнозів на множині фактичного нульового («bad») класу.

Пропонований метод передбачає, що Y^* відповідає множині навчальної вибірки, але формується за допомогою перехресної валідації (cross-validation) з використанням методу «leave-one-out» [3], а оптимальне значення k на навчальній вибірці відповідає максимальному значенню індексу Джині на навчальній вибірці при застосуванні перехресної валідації, що й пропонується за критерій оптимальності.

Таким чином, вдосконалений метод k-plus-найближчих сусідів передбачає наступні кроки:

1) створити порожній масив для значень k та *Gini* (масив «кількість-Джині»);

2) перший цикл (зовнішній): для кожного значення k від 1 до зменшеного на два значення розміру навчальної вибірки N–2 (всюди включно) провести наступні дії:

2.1) створити новий порожній масив для значень ймовірнісного прогнозу y^* та бінарного факту y (масив значень «прогноз-факт»);

2.2) другий цикл (внутрішній): для кожного елемента вибірки від 1 до N (всюди включно):

2.2.1) згідно з методом «leave-one-out», сформувати підмножину навчальної вибірки розміром N–1 без урахування поточного елементу, що розглядається;

2.2.2) обчислити значення ймовірнісного прогнозу y^* для поточного елементу на основі підмножини навчальної вибірки без його врахування, використовуючи базовий метод k-plus-найближчих сусідів;

2.2.3) додати значення ймовірнісного прогнозу y^* та бінарного факту y у відповідний масив значень «прогноз-факт»;

2.3) підрахувати значення емпіричних кумулятивних функцій розподілу, використовуючи масив «прогноз-факт»;

2.4) розрахувати значення критерію оптимальності – індексу Джині згідно з формулою (12), використовуючи класичний метод трапецій [7];

2.5) додати значення k та *Gini* у відповідний масив «кількість-Джині»;

2.6) видалити масив значень «прогноз-факт»;

3) знайти максимальне значення *Gini* у відповідному масиві «кількість-Джині», що відповідатиме оптимальному значенню k: k_o.

Очевидно, що вироджений випадок k=N–1 виключається, бо, після застосування базового методу (п. 2.2.2) для всіх елементів вибірки на внутрішньому циклі, індекс Джині дорівнюватиме мінус одиниці (видимість штучно отриманої анти-класифікації), оскільки для кожного елемента прогностичне значення y^* буде лінійно залежним від бінарного факту при застосуванні методу «leave-one-out»:

$$y^* = \frac{G - y}{B + G - 1}. \qquad (13)$$

Також на практиці достатньо та рекомендовано брати праву границю для зовнішнього циклу по k меншою наприклад в декілька разів за N–2, тим більше застосування дуже великого значення k не є ефективним, швидким та логічним.

4. 4. Методика проведення експерименту з використанням ключових можливостей мови структурованих запитів SQL

Мова структурованих запитів SQL відноситься до мов четвертого покоління та дозволяє ефективно оперувати з множинами та великими масивами інформації [7, 9, 10, 13]. Застосування віконних функцій [10] дає змогу лаконічно та швидко розрахувати для кожного рядка навчальної вибірки вагу категорії відповідної рядку для будь-якої змінної згідно з формулою (3), наприклад, таким чином (де поле GOOD – бінарний індикатор цільової змінної): **SELECT LOG(**1.0***(SUM(**GOOD**)** OVER**(PARTITION BY** *<поле-змінна>***))/(SUM(**GOOD**)** OVER**())/ (SUM(**1-GOOD**)** OVER**(PARTITION BY** *<поле-змінна>***))*(SUM(**1-GOOD**)** OVER**())) AS** WOE **FROM** *<повна таблиця навчальної вибірки>*.

У даному підрозділі наведено приклад застосування базового методу k-plus-найближчих сусідів на тестову вибірку, що не входить до складу навчальної, але має бінарний фактичний результат цільової змінної для подальшого оцінювання якості прогнозів.

Приклад лаконічного (в порівнянні з мовами третього покоління) коду генерації ваг категорій п'яти змінних для навчальної вибірки, що відповідає проведеному експерименту на даних споживчого кредитування:

```
IF OBJECT_ID('dbo.TEMP_DEV_WOE_MEMORY') IS NOT NULL DROP TABLE dbo.TEMP_DEV_WOE_MEMORY;
SELECT t.*, ROWNUM = IDENTITY(INT, 1, 1) INTO dbo.TEMP_DEV_WOE_MEMORY
FROM
(
 SELECT VAR2_GENDER_AGE, VAR2_EDU_CURREXP,
  VAR3_EDU_MARR_CHILD, VAR3_BRANCH_POSITION_TOTALEXP,
  VAR2_BRANCH_SECTOR,

  LOG(1.0*(SUM(GOOD) OVER(PARTITION BY VAR2_GENDER_AGE))/
    (SUM(GOOD) OVER())/(SUM(1-GOOD) OVER(PARTITION BY
      VAR2_GENDER_AGE))*(SUM(1-GOOD) OVER())) AS WOE1,
```

```
    LOG(1.0*(SUM(GOOD) OVER(PARTITION BY VAR2_EDU_CURREXP))/
        (SUM(GOOD) OVER())/(SUM(1-GOOD) OVER(PARTITION BY
        VAR2_EDU_CURREXP))*(SUM(1-GOOD) OVER())) AS WOE2,

    LOG(1.0*(SUM(GOOD) OVER(PARTITION BY VAR3_EDU_MARR_CHILD))
        /(SUM(GOOD) OVER())/(SUM(1-GOOD) OVER(PARTITION BY
        VAR3_EDU_MARR_CHILD))*(SUM(1-GOOD) OVER())) AS WOE3,

    LOG(1.0*(SUM(GOOD) OVER(PARTITION BY VAR3_BRANCH_POSITION_TOTALEXP))/
        (SUM(GOOD) OVER())/(SUM(1-GOOD) OVER(PARTITION BY
        VAR3_BRANCH_POSITION_TOTALEXP))*(SUM(1-GOOD) OVER()))
    AS WOE4,

    LOG(1.0*(SUM(GOOD) OVER(PARTITION BY VAR2_BRANCH_SECTOR))/
        (SUM(GOOD) OVER())/(SUM(1-GOOD) OVER(PARTITION BY
        VAR2_BRANCH_SECTOR))*(SUM(1-GOOD) OVER())) AS WOE5,
    GOOD
    FROM dbo.T_V1_DEV_GRAY
    ) t;
```

Приклад коду доповнення тестової вибірки значеннями WOE з навчальної вибірки (тобто доповненнями числовими перетвореннями категоріальних змінних), використовуючи конструкцію JOIN [9] для з'єднання таблиць:

```
    IF OBJECT_ID('dbo.TEMP_VAL_WOE') IS NOT NULL DROP TABLE dbo.TEMP_VAL_WOE;
    SELECT t.*, convert(float, NULL) AS P_GOOD_FORECAST, ROWNUM = IDENTITY(INT, 1, 1) INTO dbo.TEMP_VAL_WOE
    FROM
    (
    SELECT val.VAR2_GENDER_AGE, val.VAR2_EDU_CURREXP, val.VAR3_EDU_MARR_CHILD, val.VAR3_BRANCH_POSITION_TOTALEXP,
    val.VAR2_BRANCH_SECTOR,
    t1.WOE1, t2.WOE2, t3.WOE3, t4.WOE4, t5.WOE5,
    val.GOOD
    FROM dbo.T_V2_VAL_GRAY val

    LEFT JOIN
    (SELECT DISTINCT VAR2_GENDER_AGE, WOE1
    FROM dbo.TEMP_DEV_WOE_MEMORY) t1
    ON ISNULL(t1.VAR2_GENDER_AGE, 'NULL')=
    ISNULL(val.VAR2_GENDER_AGE, 'NULL')

    LEFT JOIN
    (SELECT DISTINCT VAR2_EDU_CURREXP, WOE2
    FROM dbo.TEMP_DEV_WOE_MEMORY) t2
    ON ISNULL(t2.VAR2_EDU_CURREXP, 'NULL')=
    ISNULL(val.VAR2_EDU_CURREXP, 'NULL')

    LEFT JOIN
    (SELECT DISTINCT VAR3_EDU_MARR_CHILD, WOE3
    FROM dbo.TEMP_DEV_WOE_MEMORY) t3
    ON ISNULL(t3.VAR3_EDU_MARR_CHILD, 'NULL')=
    ISNULL(val.VAR3_EDU_MARR_CHILD, 'NULL')

    LEFT JOIN
    (SELECT DISTINCT VAR3_BRANCH_POSITION_TOTALEXP, WOE4
    FROM dbo.TEMP_DEV_WOE_MEMORY) t4
    ON ISNULL(t4.VAR3_BRANCH_POSITION_TOTALEXP, 'NULL')=
    ISNULL(val.VAR3_BRANCH_POSITION_TOTALEXP, 'NULL')

    LEFT JOIN
    (SELECT DISTINCT VAR2_BRANCH_SECTOR, WOE5
    FROM dbo.TEMP_DEV_WOE_MEMORY) t5
    ON ISNULL(t5.VAR2_BRANCH_SECTOR, 'NULL')=
    ISNULL(val.VAR2_BRANCH_SECTOR, 'NULL')
    ) t;

    CREATE INDEX TEMP_VAL_WOE_INDEX5 ON dbo.TEMP_VAL_WOE(WOE1, WOE2, WOE3, WOE4, WOE5);
```

Очевидно, для найпростішого прискорення подальших обчислень застосовується індексування доповненої тестової вибірки (про більш ефективні пропоновані методи прискорення методу йдеться у наступному підрозділі).

Надалі для ймовірнісної класифікації застосовується ключова можливість мови структурованих запитів SQL, що ідеально застосовна отримання $k^+(\mathbf{x}^*)$ найближчих сусідів, а саме конструкція «TOP N WITH TIES» [9, 10]. Наведемо приклад коду для ймовірнісної класифікації при k=10:

```
    DECLARE @k int;
    SET @k = 10;
    UPDATE dbo.TEMP_VAL_WOE
    SET P_GOOD_FORECAST =
    (SELECT avg(convert(float, t2.GOOD))
    FROM (SELECT TOP(@k) WITH TIES t1.GOOD
    FROM dbo.TEMP_DEV_WOE_MEMORY t1
    WHERE t1.ROWNUM != dbo.TEMP_VAL_WOE.ROWNUM
    ORDER BY SQRT(POWER(t1.WOE1-dbo.TEMP_VAL_WOE.WOE1,2)+
        POWER(t1.WOE2-dbo.TEMP_VAL_WOE.WOE2,2)+
        POWER(t1.WOE3-dbo.TEMP_VAL_WOE.WOE3,2)+
        POWER(t1.WOE4-dbo.TEMP_VAL_WOE.WOE4,2)+
        POWER(t1.WOE5-dbo.TEMP_VAL_WOE.WOE5,2))
    ) t2
    );
```

4. 5. Формалізація вдосконалень методу щодо зменшення обчислювальної складності при проведенні експерименту та наведення відповідних можливостей мови структурованих запитів

Програмну реалізацію даного методу можна значно вдосконалити методом агрегування за допомогою агрегатних функцій визначених на групах [9, 10], прогнозуючи за допомогою ймовірності спочатку цільову змінну для векторів $\mathbf{x}^*$, що відразу по нульовій відстані (точній рівності) потрапляють у групу рівних по координатах векторів **x**, яка становить щонайменше k векторів, а далі класифікуючи за допомогою ймовірності методом описаним в підрозділі 4. 4 всі інші вектори, що ще не класифікувалися:

```
DECLARE @k int;
SET @k = 10;

IF OBJECT_ID('dbo.TEMP_QUICK_MEMORY') IS NOT NULL
DROP TABLE dbo.TEMP_QUICK_MEMORY;
/*для випадків нульової відстані до груп чисельністю не менше k:*/
SELECT WOE1, WOE2, WOE3, WOE4, WOE5,
 count(*) AS cnt, sum(GOOD) AS goods,
 1.0*sum(GOOD)/count(*) AS P_good,
        ROW_NUMBER() OVER(ORDER BY count(*) DESC) RN
INTO dbo.TEMP_QUICK_MEMORY
FROM dbo.TEMP_DEV_WOE_MEMORY
GROUP BY WOE1, WOE2, WOE3, WOE4, WOE5
HAVING count(*) >= @k
ORDER BY count(*) DESC;

CREATE UNIQUE INDEX TEMP_QUICK_MEMORY_PK ON dbo.TEMP_QUICK_MEMORY(WOE1, WOE2, WOE3, WOE4, WOE5);

/*на всякий випадок, додаткова очистка від попередніх значень прогнозів:*/
UPDATE dbo.TEMP_VAL_WOE SET P_GOOD_FORECAST = NULL;

UPDATE VAL
SET VAL.P_GOOD_FORECAST = QM.P_good
FROM dbo.TEMP_VAL_WOE VAL
      INNER JOIN
 dbo.TEMP_QUICK_MEMORY QM
 ON QM.WOE1 = VAL.WOE1 AND QM.WOE2 = VAL.WOE2
 AND QM.WOE3 = VAL.WOE3 AND QM.WOE4 = VAL.WOE4
AND QM.WOE5 = VAL.WOE5;

UPDATE dbo.TEMP_VAL_WOE
SET P_GOOD_FORECAST =
 (SELECT avg(convert(float, t2.GOOD))
 FROM (SELECT TOP(@k) WITH TIES t1.GOOD
 FROM dbo.TEMP_DEV_WOE_MEMORY t1
 WHERE t1.ROWNUM != dbo.TEMP_VAL_WOE.ROWNUM
 ORDER BY SQRT(POWER(t1.WOE1-dbo.TEMP_VAL_WOE.WOE1,2)+
        POWER(t1.WOE2-dbo.TEMP_VAL_WOE.WOE2,2)+
        POWER(t1.WOE3-dbo.TEMP_VAL_WOE.WOE3,2)+
        POWER(t1.WOE4-dbo.TEMP_VAL_WOE.WOE4,2)+
        POWER(t1.WOE5-dbo.TEMP_VAL_WOE.WOE5,2))
 ) t2
 )
WHERE P_GOOD_FORECAST IS NULL;
```

4. 6. Оцінка якості прогнозів на тестовій вибірці при проведенні експерименту засобами мови структурованих запитів

Оскільки критерієм оптимальності вдосконаленого методу k-plus-найближчих сусідів обрано індекс Джині, як найбільш популярний показник якості прогнозів у кредитному скорингу [2–4, 7], то, використовуючи віконні функції [10], наведемо його програмну реалізацію, ідея та реалізація якої для висвітлена в [7], адаптувавши її до особливостей системи керування базами даних (СКБД) MS SQL Server (нижче наведено код сумісний як мінімум починаючи з версії 2005):

```
WITH smpl(BAD, GOOD, score) AS
(
 /*start sample*/
 SELECT 1 - GOOD AS BAD, GOOD, P_GOOD_FORECAST AS SCORE
 FROM dbo.TEMP_VAL_WOE
 /*end sample*/
),
distr AS
(
 SELECT score,
  1.0*(sum(GOOD))/(sum(sum(GOOD)) over()) AS GOOD,
 1.0*(sum(BAD))/(sum(sum(BAD)) over()) AS BAD
 FROM smpl
 GROUP BY score
),
cum AS
(
 SELECT D_BASE.SCORE,
  sum(D_LESS.GOOD) AS GOOD, sum(D_LESS.BAD) AS BAD,
  ROW_NUMBER() OVER(ORDER BY D_BASE.SCORE) AS RN
 FROM distr d_base LEFT OUTER JOIN distr d_less
 ON D_LESS.SCORE<=D_BASE.SCORE
 GROUP BY D_BASE.SCORE
),
cum_with_lag AS
(
 SELECT cum.*, ISNULL(cum_prev.GOOD, 0) AS GOOD_PREV,
 ISNULL(cum_prev.BAD, 0) AS BAD_PREV
 FROM cum LEFT JOIN cum AS cum_prev ON cum_prev.RN = cum.RN - 1
)
SELECT 'GINI' AS "Indicator",
```

convert(varchar(6**),** **convert(numeric(**5, 2**),** **ROUND((**1.0*****
sum((GOOD - GOOD_PREV**)*(**BAD + BAD_PREV**) / 2)-** 0.5**)/**0.5, 4**)***100**)) +** '%' **AS** "Value" **FROM** cum_with_lag**;**

Дана програмна реалізація легко застосовна для оцінювання якості прогнозів довільних бінарних ймовірнісних класифікаторів.

4. 7. Додаткові засоби моніторингу процесу виконання експерименту за допомогою SQL

Оскільки процес виконання експерименту на етапі останнього «UPDATE» має високу обчислювальну складність, то практично цінним є можливість он-лайн контролю ходу виконання експерименту. Для цього можна застосувати рівень ізоляції транзакції «READ UNCOMMITTED» [9], щоб порахувати наприклад миттєву кількість векторів з тестової вибірки, яким ще не присвоєно прогноз:

SET TRANSACTION ISOLATION LEVEL READ UNCOMMITTED;
SELECT COUNT(*)
FROM dbo.TEMP_VAL_WOE
WHERE P_GOOD_FORECAST **IS NULL;**

5. Результати проведення експерименту на базі даних споживчого кредитування

Як згадується в підрозділі 4.4, моделювання за допомогою базового методу k-plus-найближчих сусідів здійснюється на даних споживчого кредитування з використанням п'яти комбінованих змінних, що включають в себе атомарні змінні (наприклад, змінна VAR2_GENDER_AGE включає дані про стать та вік клієнта).

Результати якості порівняння прогнозів (в т. ч. з логістичною регресією) наведені в табл. 1 (скорочено запропонований метод будемо називати k-plus-NN, тобто походить від «k-plus-nearest neighbor»).

Таблиця 1

Порівняння якості прогнозів базового методу k-plus-найближчих сусідів з логістичною регресією

Метод моделювання	Індекс Джині	Кількість параметрів, що оптимізуються	Кількість безумовно заданих параметрів
Логістична регресія	40,32 %	**6**	**0**
k-plus-NN (k=10)	30,45 %	**0**	**1**
k-plus-NN (k=50)	36,58 %	**0**	**1**

Кількість параметрів логістичної регресії включає зміщення (intercept), а застосована тут формула для порівняння результатів має вигляд [2, 12]:

$$y_p^* = \frac{1}{1+e^{-f(\mathbf{x}_p)}}, \qquad (14)$$

де

$$f(\mathbf{x}_p) = c_0 + \sum_{i=1}^{m} c_i WOE_{ij_i(p)}, \qquad (15)$$

де m – кількість змінних.

6. Аналіз результатів проведеного експерименту побудови моделей навчання на основі пам'яті та за допомогою логістичної регресії (для порівняння) на даних споживчого кредитування

На основі аналізу табл. 1 можна зробити висновок, що навіть базовий метод k-plus-найближчих сусідів, на фоні повної відсутності оптимізації будь-яких параметрів (тут значення k) та лише завдяки обраній метриці на множині ваг категорій змінних та заданим значенням k, дає результати не набагато гірші та цілком порівнянні з результатами логістичної регресії, тому, логічно, вдосконалений метод k-plus-найближчих сусідів, де значення k оптимізується при застосуванні складної обчислювальної процедури, даватиме набагато кращі результати.

7. Висновки

1. Досліджено та запропоновано простір числових перетворень категоріальних (в т. ч. дискретизованих) змінних з використанням перетворень відносно цільової змінної, використовуючи ваги категорій змінних згідно з класичною методологією кредитного скорингу, застосовано класичну метрику та детально досліджено її властивості у рамках саме термінології скорингового моделювання. Таким чином тісніше пов'язано методологію побудови скорингових карт з теорією машинного навчання на основі пам'яті. Надалі експериментально доведено доцільність використання запропонованої метрики, як однієї з можливих ефективних метрик.

2. Формалізовано вирішення проблеми рівновіддалених груп елементів відносно елемента, що класифікується, що є однією з невисвітлених (або неналежно розв'язних) проблем у класичному методі k-найближчих сусідів. Продемонстровано нагальну необхідність вирішення цього питання саме у випадку числових перетворень категоріальних змінних, коли область визначення всіх змінних скінченна. Доведено ефективність та детермінованість запропонованого методу вирішення проблеми рівновіддалених груп у рамках загального методу, а також простоту та спеціальні готові засоби реалізації саме за допомогою мови структурованих запитів діалекту MS SQL (T-SQL) – транзакційної SQL.

3. Наведено чітке формулювання формалізації пропонованого базового методу k-plus-найближчих сусідів та вдосконаленого методу на основі базового з використанням критерію оптимальності класичного для кредитного скорингу – індексу Джині. Ключовими особливостями методу та його вдосконалення є: ймовірнісне значення прогнозів, коректне та детерміноване врахування рівновіддалених груп, використання метрики на просторі класичних показників у рамках методології кредитного скорингу, визначення критерію оптимальності моделі та вибору значення вхідного параметру на основі перехресної валідації, проста інтерпретація методу, властивостей простору та опосередкований взаємозв'язок з показниками аналізу характеристик. У результаті метод передбачає всі можливі ситуації та більш детально пояснює особливості та застереження відносно використання мето-

ду виключення одного елементу – методу «leave-one-out» – у якості побічного результату.

4. Запропоновано повну програмну реалізацію базового методу k-plus-найближчих сусідів та розрахунку критерію оптимальності з належним та влучним використанням ключових можливостей мови структурованих запитів SQL діалекту MS SQL (T-SQL). Використана мова СКБД четвертого покоління має високу властивість читабельності мови високого рівня, забезпечує оперування множинами та можливість обробки даних безпосередньо у середовищі їх збереження, що є суттєвою перевагою. Продемонстровано використання саме транзакційних особливостей діалекту T-SQL на прикладі контролю процесу виконання методу.

5. Представлено оригінальний підхід до прискорення процесу виконання методу за допомогою зменшення його обчислювальної складності через розбиття на два етапи процесу прогнозування засобами агрегування мови SQL. Важливість даного методу прискорення прогнозування надзвичайно важлива при малих значеннях вхідного параметру загального пропонованого методу.

6. Проведено порівняльний аналіз результатів базового методу k-plus-найближчих сусідів для декількох значень вхідного параметру з результатами методу моделювання за допомогою логістичної регресії. У якості висновків, відзначено якість моделей, основаних на базовому пропонованому методі, порівнянною з результатами логістичної регресії на прикладі даних масового споживчого кредитування. Основними перевагами перед логістичною регресією є простота реалізації (а для базового методу взагалі відсутність параметрів, які оптимізуються), також більш явне самостійне врахування метрикою предикативної сили вхідних змінних з нівелюванням впливу змінних зі слабким взаємозв'язком з цільовою змінною.

Література

1. Барбаумов, В. Е.Энциклопедия финансового риск-менеджмента [Текст] / В. Е. Барбаумов, М. А. Рогов, Д. Ф. Щукин и др.; под ред. А. А. Лобанова, А. В. Чугунова. – М.: Альпина Паблишер, 2003. – 786 с.
2. Siddiqi, N. Credit risk scorecards: developing and implementing intelligent credit scoring [Text] / N. Siddiqi. – Hoboken: John Wiley & Sons, Inc., 2006. – 196 p.
3. Thomas, L. C. Credit Scoring and its Applications [Text]: monograph / L. C. Thomas, D. B. Edelman, J. N. Crook. – Philadelphia: SIAM, 2002. – 248 p.
4. Ванг, Вэй Руководство по кредитному скорингу [Текст] / Вэй Ванг, А. Д. Влатса, К. Д. Гленнон и др.; пер. с англ. И. М. Тикота; науч. ред. Д. И. Вороненко; под. ред. Э. Мэйз. – Минск: Гревцов Паблишер, 2008. – 464 с.
5. Солошенко, О. М. Вдосконалення методу ітеративної класифікації з включення відхилених заявок у кредитному скорингу [Текст] / О. М. Солошенко // Наук. вісті НТУУ «КПІ». – 2014. – № 5. – С. 63–69.
6. Солошенко, О. М. Дослідження відстані Кульбака-Лейблера у задачах моделювання у кредитному скорингу [Текст]: сб. науч. трудов междунар. конф. / О. М. Солошенко // Развитие информационно-ресурсного обеспечения образования и науки в горно-металлургической отрасли и на транспорте. – Днепропетровск: НГУ, 2014. – С. 328–333.
7. Солошенко, О. М. Спосіб розрахунку показника Джині, статистики Колмогорова-Смирнова та відстані Махаланобіса у кредитному скорингу засобами мови SQL [Текст] / О. М. Солошенко // Наук. вісті НТУУ «КПІ». – 2015. – № 1. – С. 29–35.
8. Haykin, S. Neural networks: a comprehensive foundation. 2nd edition [Text] / S. Haykin. – Delhi: Pearson Education, Inc., 2005. – 823 p.
9. Ben-Gan, I. Microsoft® SQL Server® 2012 T-SQL fundamentals [Text] / I/ Ben-Gan. – Sebastopol: O'Reilly Media, Inc., 2012. – 412 p.
10. Ben-Gan, I. Microsoft® SQL Server® 2012 high-performance T-SQL using window functions [Text] / I. Ben-Gan. – Sebastopol: O'Reilly Media, Inc., 2012. – 221 p.
11. Терентьєв, О. М. Моделі і методи побудови та аналізу байєсівських мереж для інтелектуального аналізу даних [Текст]: дис. ... канд. техн. наук / О. М. Терентьєв. – К., 2009. – 258 с.
12. Allison, P. D. Logistic regression using the SAS® system: theory and application [Text] / P. D. Allison. – Cary: SAS Institute Inc., 1999. – 287 p.
13. Шипунов, А. Б. Наглядная статистика. Используем R! [Текст] / А. Б. Шипунов, Е. М. Балдин, П .А. Волкова и др. — М.: ДМК Пресс, 2014. — 298 с.
14. Егорова, И. Н. Программная реализация методов классификации [Текст] / И. Н. Егорова, С. В. Егоров // Східно-Європейський журнал передових технологій. – 2010. – Т. 1, № 5 (43). – С. 52–54. – Режим доступу: http://journals.uran.ua/eejet/article/view/2579/2384
15. Keller, J. M. A fuzzy k-nearest neighbor algorithm [Text] / J. M. Keller, M. R. Gray, J. A. Jr. Givens // IEEE transactions on systems, man and cybernetics. – 1985. – Vol. SMC-15, Issue 4. – P. 580–585. doi: 10.1109/tsmc.1985.6313426
16. Берзлев, О. Ю. Метод прогнозування знаків приростів часових рядів [Текст] / О. Ю. Берзлев // Східно-Європейський журнал передових технологій. – 2013. – Т. 2, № 4 (62). – С. 8–11. – Режим доступу: http://journals.uran.ua/eejet/article/view/12362/10250
17. Солошенко, О. М. Адаптація формул підрахунку ваг категорій змінної та значення інформації змінної при відомому розподілі категорій та відомих умовних ймовірностях негативних значень цільової змінної [Текст] / О. М. Солошенко // Проблеми науки. – 2014. – № 10 (166). – С. 45–47.