УДК 519.226, 330.322
DOI: 10.20535/1810-0546.2016.2.63882

О.М. Терентьєв[1], В.Е. Кириченко[1], Н.О. Связінська[1], Т.І. Просянкіна-Жарова[2]
[1]Національний технічний університет України "КПІ", Київ, Україна
[2]Уманська філія Європейського університету, Умань, Україна

ПРОГНОЗУВАННЯ ФІНАНСОВИХ РИЗИКІВ З ВИКОРИСТАННЯМ НАЇВНОГО І ДОПОВНЕНОГО ДЕРЕВОМ КЛАСИФІКАТОРІВ НА ОСНОВІ БАЙЄСІВСЬКИХ МЕРЕЖ

Background. Development and study of characteristics for naïve and tree-augmented classifiers in the form of Bayesian networks in the problem of credit risk estimation.
Objective. To perform estimation of classification quality for the bank credit borrowers using Bayesian classifiers of two types.
Methods. Development of necessary mathematical tools and performing computational experiments aiming towards constructing classifiers in the form of Bayesian networks using actual statistical data characterizing solvency of bank credit borrowers.
Results. The following results were achieved: the methodology of constructing and application of the naïve and tree-augmented Bayesian classifiers for solving the problem of solvency estimation for bank credit borrowers; an analysis of computational algorithmic complexity was performed; two classification models were constructed in the form of Bayesian networks using actual statistical data from banking system; a comparative analysis was performed for the models developed.
Conclusions. It was established that the tree-augmented classifier exhibits higher computational complexity than the naïve Bayesian one, but it showed higher classification results while solving the problem of bank clients classification into two groups: those who return the credits and those who don't.
Keywords: intellectual data analysis; Bayesian networks; credit scoring; financial analysis; macroeconomic indicators.

Вступ

Розв'язання задач класифікації та прогнозування методами інтелектуального аналізу даних є одним із найбільш актуальних напрямів наукових досліджень, оскільки в сучасному світі все частіше виникає необхідність роботи з великими масивами даних [1, 2]. У статті розглядається один із методів розв'язання таких задач – побудова спеціальної мережі Байєса, а саме наївного та доповненого деревом класифікаторів для розв'язання задачі коректної класифікації кредитоспроможності фізичних осіб.

У загальному випадку байєсівська мережа являє собою пару $\langle G, B \rangle$, у якій перша компонента G – це спрямований нециклічний граф, що відповідає випадковим змінним і записується як набір умов незалежності: кожна змінна незалежна від її батьків у G. Друга компонента пари B – це множина параметрів, що визначають мережу [4–6].

Наївний (naïve) та доповнений деревом (tree-augmented, або TAN) байєсівські класифікатори – це ймовірнісні графічні моделі, що використовуються для формалізованого опису великих масивів даних, які містять невизначеності серед своїх взаємозалежних наборів характеристик. Ці моделі широко використовуються для розв'язання задач сегментації зображень, медичної діагностики та інших задач кластеризації і класифікації на основі статистичних даних.

Проблема класифікації полягає у виявленні, до якого класу належить конкретний об'єкт, на основі знань, отриманих після аналізу подібних об'єктів. Кожний елемент описується за допомогою множини змінних, які називають характеристиками або параметрами. Використання наївного байєсівського класифікатора ґрунтується на тому, що всі змінні (характеристики) є незалежними одна від іншої. Це дуже просте уявлення стосовно характеристик системи, але у той же час незалежність, що визначається цією моделлю, не завжди є реалістичною. Модель TAN – це покращений наївний класифікатор Байєса, який враховує ще один рівень взаємодії між параметрами досліджуваної системи, тобто кожна змінна може залежати від деякої іншої змінної. При цьому залежність між характеристиками моделі TAN є реалістичнішою, ніж у наївному класифікаторі [7, 8].

Мережа Байєса – це зручний інструмент для опису досить складних процесів і подій з невизначеностями ймовірнісного характеру. Одним зі світових лідерів прогнозної аналітики є компанія SAS, яка розробила спеціалізований інструмент SAS Enterprise Miner для розв'язання задач інтелектуального аналізу даних. У 2015 р. компанія SAS у межах тринадцятої вер-

сії продукту SAS Enterprise Miner запропонувала новий компонент HP Bayesian Network Classifier, що розроблявся підрозділом R&D у м. Керрі (Північна Кароліна, США) протягом останніх семи років. За допомогою цієї компоненти можна будувати мережі Байєса різних типів, у т.ч. наївну та доповнену деревом.

Постановка задачі

Задачі дослідження є такими: створити опис методики використання наївного та доповненого деревом байєсівських класифікаторів при розв'язанні практичних задач оцінювання кредитоспроможності позичальників кредитів; виконати аналіз методів обчислення алгоритмічної складності наведених алгоритмів; побудувати класифікаційні моделі у формі байєсівських мереж із використанням фактичних статистичних даних щодо визначення кредитоспроможності фізичних осіб; виконати аналіз результатів обчислювальних експериментів.

Наївний байєсівський класифікатор

Наївний байєсівський класифікатор – особлива форма ймовірнісної моделі у вигляді байєсівської мережі, яка характеризується тим, що має сильні припущення стосовно незалежності змінних. Структуру наївного байєсівського класифікатора подано на рис. 1.

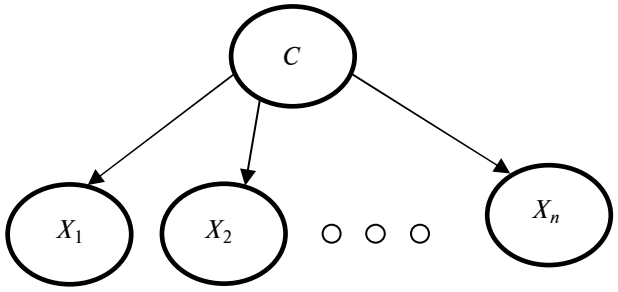

Рис. 1. Приклад структури наївного байєсівського класифікатора

Ця модель широко використовується для розв'язання задач класифікації. На рис. 1 C являє собою цільову змінну, що може набувати значення $\{C_1,...,C_k\}$, де C_i – стан цільової змінної, а $X_1,...,X_n$ – множина незалежних вхідних змінних процесу, які можуть впливати на цільову. Основним припущенням моделі є те, що всі вхідні змінні незалежні між собою, але це рідко відповідає дійсності [8].

Для прикладу розглянемо модель, яка використовується для прийняття рішення стосовно того, чи варто грати в теніс у той день, коли відомі деякі погодні умови [9]. Вхідними змінними цієї моделі є: рівень видимості, вологість, температура, вітер і місце розташування. Цільова змінна вказує, підходить погода для гри в теніс чи ні, тобто вона може набувати двох значень {так, ні}. У цьому прикладі, якщо погода сонячна або вологість висока, то більш імовірно, що і температура висока. Усі вхідні змінні системи пов'язані між собою. Такий підхід є швидким для використання та розуміння і надає при цьому кращі результати порівняно з іншими підходами. З рис. 1 можна побачити, що спільний розподіл випадкових величин зводиться до рівняння

$$P(X_1, X_2, ..., X_n) = P(C) \cdot \prod_{i=1}^{n} P(X_i | C).$$

У цій моделі умовні ймовірності всіх змінних, з урахуванням цільових, обчислюються у процесі навчання моделі. На етапі тестування моделі розраховуються апостеріорні ймовірності для кожного стану цільової змінної. Рішення вибирається (приймається) за максимальним значенням ймовірності. Згідно з теоремою Байєса маємо

$$P(C | X_1, X_2, ..., X_n) = \frac{P(X_1, X_2, ..., X_n | C) \cdot P(C)}{P(X_1, X_2, ..., X_n)}.$$

Знаменник $P(X_1, X_2, ..., X_n)$ у цьому рівнянні є однаковим для всіх станів цільової змінної для всіх станів навчальної вибірки. Оскільки знаменник не відіграє ролі при максимізації $P(C | X_1, X_2, ..., X_n)$ відносно цільової змінної, то його можна ігнорувати, а для класифікації стану розраховувати тільки чисельник.

Доповнений деревом байєсівський класифікатор

У реальному світі змінні будь-якої досліджуваної системи корелюють між собою, а об'єкти з усіма незалежними змінними трапляються дуже рідко. Якщо модель враховує кореляції між змінними, то точність класифікації, як правило, поліпшується. Одним із варіантів вирішення цієї проблеми є доповнена деревом байєсівська модель. Вона підтримує структуру

наївного байєсівського класифікатора і доповнює її додаванням ребер між вершинами, які являють собою змінні досліджуваного процесу, з метою відображення кореляції між змінними.

Одночасно такий підхід призводить до підвищення обчислювальної складності алгоритмів. Разом із тим використання наївного байєсівського класифікатора вимагає зберігання тільки умовних ймовірностей належності до станів цільової змінної, а доповнена модель – перебору всіх можливих топологій мережі Байєса.

Для того щоб зменшити обчислювальну складність, а також врахувати кореляції між змінними, необхідно накласти обмеження на рівні взаємодії між змінними. Однією з таких моделей є модель класифікатора, розширеного деревом (TAN). Ця модель вводить обмеження на кількість батьківських змінних: їх може бути не більше двох. У моделі TAN всі вхідні змінні пов'язані з цільовою змінною за допомогою спрямованих ребер. Отже, при визначенні умовної ймовірності $P(C|X_1, X_2, ..., X_n)$ до уваги беруться всі незалежні параметри. При цьому кожна змінна у графі може мати двох батьків, а саме: цільову змінну та іншу вхідну змінну, що не є нащадком.

Обчислювальна складність такої моделі значно зменшується, адже кожна змінна може мати не більше двох батьків. Таким чином, TAN має ненабагато більшу обчислювальну складність, ніж наївний байєсівський класифікатор, але при цьому показує вищу точність класифікації. На рис. 2 подано приклад TAN-класифікатора.

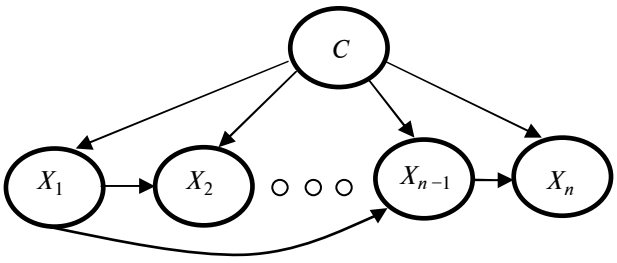

Рис. 2. Структура доповненого деревом байєсівського класифікатора

Ключовою особливістю доповненої деревом моделі є її деревоподібна структура. Для того щоб побудувати дерево, необхідно спочатку визначити батька кожної змінної. Крім того, тільки змінні з максимальною кореляцією повинні бути з'єднані одна з одною.

Ще одним важливим поняттям, що відіграє ключову роль при побудові дерева, є взаємна інформація [10]. Для того щоб побудувати дерево, необхідно оцінити кореляцію між кожною парою змінних у системі і додати ребро тільки між тими змінними, які найбільше корелюють. Якщо у досліджуваній системі є N змінних, то відповідний граф матиме N вузлів. Для того щоб отримати деревоподібну структуру, яка з'єднує всі вузли в графі, необхідно додати $N-1$ ребро. Крім того, сума ваг усіх цих ребер повинна бути максимальною вагою серед усіх таких деревоподібних структур. Міра кореляції між двома змінними X і Y називається взаємною інформацією і обчислюється за виразом

$$I_p(X;Y) = \sum_{x,y} P(x,y) \cdot \log\left(\frac{P(x,y)}{P(x)P(y)}\right).$$

Для пари змінних ця функція показує, скільки інформації про одну змінну містить інша. Для того щоб побудувати дерево для моделі TAN, використовується умовна взаємна інформація між двома змінними. Це необхідно для того, щоб визначити ребра, які будуть належати до дерева. Умовна взаємна інформація розраховується за формулою

$$I_p(X;Y|Z) = \sum_{x,y} P(x,y,z) \cdot \log\left(\frac{P(x,y|z)}{P(x|z)P(y|z)}\right).$$

Алгоритм побудови дерева для моделі TAN складається з п'яти кроків [8].

1. Обчислити взаємну інформацію $I_p(X_i; X_j|C)$ між кожною парою змінних $i \neq j$.

2. Побудувати повний неорієнтований граф, у якому вершини є параметрами $X_1, X_2, ..., X_n$, і знайти ваги ребер, що з'єднують пари X_i і X_j, з використанням значення взаємної інформації $I_p(X_i; X_j|C)$.

3. Побудувати максимально зважене дерево.

4. Для того щоб перетворити отримане неорієнтоване дерево на орієнтоване, необхідно задати цільову зміну як кореневу вершину та визначити напрямки всіх ребер, що виходять назовні від неї.

Обчислення складності байєсівських класифікаторів

Нехай C – загальна кількість станів цільової змінної в наборі даних; N – загальна

кількість вхідних змінних у наборі даних; $\{S_1,...,S_N\}$ – загальна кількість значень, яких можуть набувати відповідні змінні $\{A_1,...,A_N\}$; $S_{\max}$ – максимальне значення $\{S_1,...,S_N\}$; R – кількість спостережень навчальної вибірки.

Припустимо, що підрахунок кількості станів цільової змінної і значень вхідних змінних може бути зроблений протягом одного сканування навчальної вибірки. Позначимо складність реалізації цього процесу для наївного та доповненого деревом класифікаторів величиною $O(R)$. Для розрахунку апріорної ймовірності потрібно знайти ймовірність появи кожного стану (значення) цільової змінної для навчальної вибірки даних. Складністю реалізації цього процесу для обох класифікаторів є $O(C)$.

Складність наївного класифікатора. Для розрахунку умовної ймовірності кожної змінної потрібно знайти ймовірність появи кожного значення всіх параметрів, обумовлених кожним станом цільової змінної. Отже, загальна кількість обчислень при реалізації цього процесу визначається за такою формулою:

$$\textit{Загальна\_кількість\_обчислень} =$$
$$= \sum_{i=1}^{N} C \cdot S_i = C \cdot S_1 + C \cdot S_2 + ... + C \cdot S_N \le$$
$$\le C \cdot (S_{\max} + S_{\max} + ... + S_{\max}) = N \cdot C \cdot S_{\max}.$$

Таким чином, складність навчання класифікатора визначається обчисленням умовних ймовірностей, для чого необхідно виконати $N \cdot C \cdot S_{\max}$ обчислень. Для класифікації тестової вибірки даних необхідно знайти ймовірності належності кожного значення вибірки до кожного стану цільової змінної. Для розрахунку цих значень необхідно перебрати кожне значення цільової змінної та всі вхідні змінні для кожного тестового об'єкта. Отже, цей процес має часову складність $N \cdot C \cdot R$. Таким чином, часова складність наївного байєсівського класифікатора обумовлюється насамперед складністю знаходження умовних ймовірностей.

Обчислювальна складність доповненого деревом класифікатора. Загальна кількість пар для N вхідних змінних дорівнює $N(N-1)/2$. Для кожної пари необхідно обчислити ймовірність появи усіх комбінацій для всіх значень, яких може набувати кожен параметр. Позначимо максимальну кількість значень кожного параметра величиною $S_{\max}$. Кількість операцій, які необхідно виконати для того, щоб знайти взаємну інформацію, визначається за формулою

$$\textit{Кількість\_операцій} \le \frac{N(N-1)}{2} \cdot S_{\max}^2.$$

Кількість обчислень для знаходження умовних ймовірностей для TAN відрізняється від наївного класифікатора. Для моделі TAN необхідно знайти умовну ймовірність для кожної вхідної змінної, враховуючи її батьківську вершину і цільову змінну. Таким чином, також буде враховуватись кількість станів батьків. Це випливає з того, що для $i = 1$ до N необхідно обчислювати $P(A_i = x \,|\, A_j = y, C = z)$, де A_j є батьком A_i, x набуває значення від 1 до S_i, а y – від 1 до S_j.

Отже, загальна кількість обчислень для знаходження умовних ймовірностей доповненого деревом класифікатора визначається за формулою

$$\textit{Загальна\_кількість\_обчислень} =$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{N} C \cdot S_i \cdot S_j \le N \cdot C \cdot S_{\max}^2.$$

Можна зробити висновок, що складність моделі TAN істотно залежить від розрахунку взаємної інформації, що зростає зі збільшенням кількості вхідних змінних. Окрім того, можна відзначити, що складність доповненої деревом моделі є поліноміальною, а складність наївної моделі – експоненціальною. Таким чином, обмежуючи рівень залежності між вхідними змінними, можна знайти простішу модель. Як правило, це вибір між складністю і адекватністю. Моделі, які мають вищі рівні залежності між змінними, як правило, мають вищу адекватність, але є дуже складними.

Практична реалізація наївного і доповненого деревом класифікаторів

Для порівняння ефективності та складності функціонування наївного і TAN класифікаторів використано аналітичну систему SAS Enterprise Miner, що є спеціалізованим інструментом для аналізу даних при створенні високоточних інтелектуальних і описових моделей, заснованих на великих масивах даних у масштабі підприємства [10].

У 2015 р. компанія SAS запропонувала новий компонент HP Bayesian Network Classifier, що надає можливість будувати мережі Байєса різних типів: наївну мережу, дереводоповнену, наївну розширену мережу Байєса (BAN), мережу батько–нащадок (PC) і покриття Маркова (MB).

Для використання HP Bayesian Network Classifier необхідна наявність категоріальної цільової змінної та однієї або більше вхідних незалежних змінних. Вхідні змінні можуть бути категоріальними або інтервальними, але значення інтервальних змінних розбиваються на рівні діапазони, кількість яких вказується в ідентифікаторі Number of Bins [11].

Застосування класифікаторів для прогнозування кредитоспроможності фізичних осіб

Для порівняння наївного і TAN класифікаторів використано стандартизовану відому серед фахівців вибірку даних German Bank Accepts, що описує процес кредитування фізичних осіб на суму до 120 тис. дол. Вибірка містить 20 змінних (19 вхідних і одну цільову) та 5837 записів (спостережень). Опис використаних змінних подано у табл. 1.

Порівняємо дві поданих вище моделі, використовуючи SAS Enterprise Miner. Діаграма обчислювальних процесів у цій системі зобра-

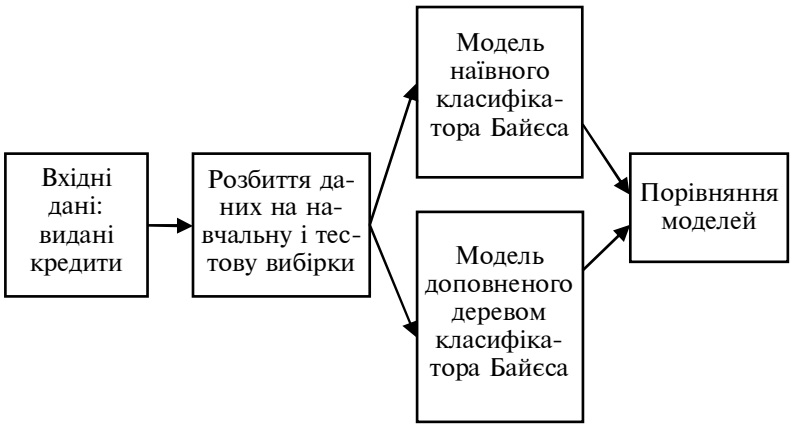

Рис. 3. Діаграма технологічного процесу побудови мережі Байєса в комп'ютерній програмі SAS Enterprise Miner

***Таблиця* 1.** Опис змінних із вибірки даних German Bank Accepts

Назва змінної	Опис
Y	Цільова змінна (0 – клієнт повернув кредит, 1 – не повернув)
X_1	Час з моменту відкриття першої кредитної лінії
X_2	Наявність непогашених кредитів (1 – клієнт повернув усі кредити, 0 – не повернув кредит)
X_3	Скоринговий бал від бюро кредитних історій
X_4	Щомісячна сума виплат
X_5	Сума кредиту
X_6	Строк видачі кредиту (міс.)
X_7	Строк співпраці з клієнтом (міс.)
X_8	Середній місячний доход
X_9	Повна сума товару, на який береться кредит
X_{10}	Тип кредитного продукту (лізинг, кредит)
X_{11}	Використана сума кредиту
X_{12}	Коефіцієнт обмеження
X_{13}	Загальний доход
X_{14}	Загальна кількість відкритих кредитних ліній
X_{15}	Загальна сума зобов'язань по всіх кредитних лініях
X_{16}	Загальна сума погашених кредитів по закритих лініях
X_{17}	Кількість погашених кредитних ліній
X_{18}	Загальна кількість кредитних ліній
X_{19}	Індекс використання кредитів (1 – повністю використаний кредит, 0 – частково)

жена на рис. 3. За допомогою вузла Data Partition вхідна вибірка буде розділена на навчальну і тестову у співвідношенні 70 і 30 % відповідно. Вузол HP BN Classifier відповідає за побудову наївного байєсівського класифікатора, а вузол HP BN Classifier (2) – за побудову TAN-класифікатора. За допомогою вузла Model Comparison можна порівняти побудовані моделі за певними статистичними характеристиками.

Побудована наївна мережа показана на рис. 4.

Найбільшу кількість інформації для прийняття рішення стосовно видачі кредиту містять такі характеристики: X_3 (скоринговий бал бюро кредитних історій); X_{16} (загальна сума погашених кредитів по закритих лініях); X_1 (термін з моменту відкриття першої кредитної лінії).

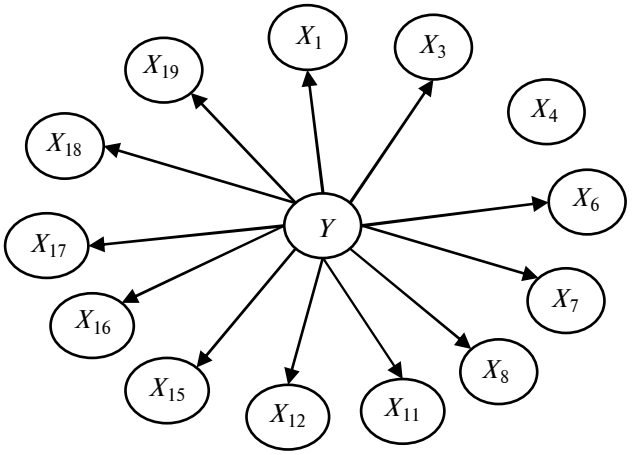

Рис. 4. Наївна байєсівська мережа для вибірки даних German Bank Accepts

На рис. 5 наведена топологія побудованої TAN-моделі.

Дерево вхідних змінних зображено на рис. 6.

Основні характеристики, які впливають на рішення про кредитування особи, збігаються з тими, що наведені для наївної мережі: X_3 (скоринговий бал бюро кредитних історій), X_{16} (загальна сума погашених кредитів по закритих лініях), X_1 (термін з моменту відкриття першої кредитної лінії).

Порівняємо наївний та доповнений деревом класифікатори на основі таких статистичних характеристик: відсоток неправильно класифікованих спостережень – кількість неправильно класифікованих спостережень відносно загальної кількості спостережень. Чим менший цей відсоток, тим кращою є модель.

ROC-індекс – числовий показник площі під ROC-кривою. Своєю чергою ROC-крива показує залежність кількості правильно класифікованих позитивних спостережень від кількості неправильно класифікованих негативних спостережень. Допустимими для моделі є значення від 0,7 до 1. Чим більшого значення набуває цей індекс, тим кращою є модель.

Коефіцієнт GINI – відношення площі фігури, утвореної ROC-кривою і кривою Лоренца, до площі одиничного квадрата. Допустимими для моделі є значення від 0,4 до 1. Чим більше значення коефіцієнта, тим кращою є

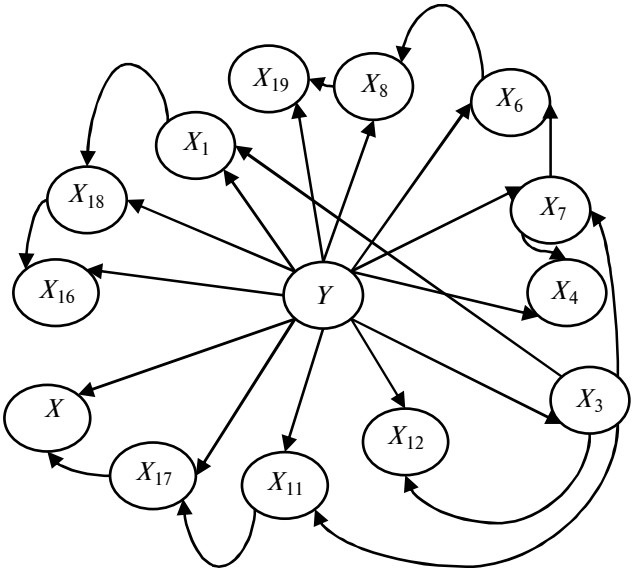

Рис. 5. Доповнена деревом байєсівська мережа для набору даних German Bank Accepts

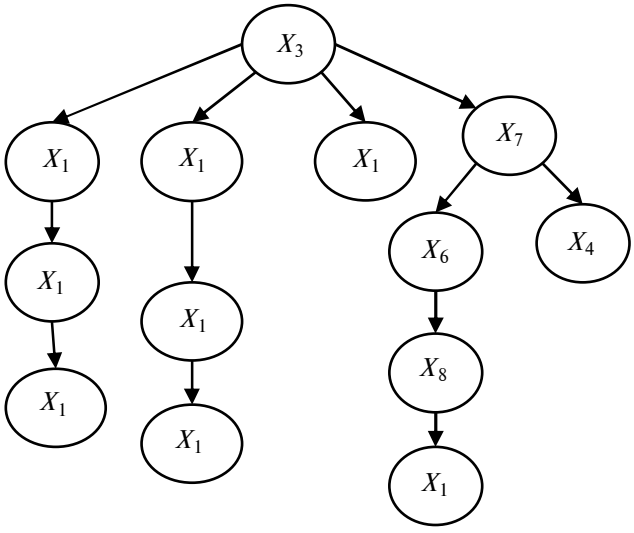

Рис. 6. Дерево вхідних змінних для вибірки даних German Bank Accepts

модель. Значення статистичних характеристик для моделей наведені у табл. 2 і 3.

Отже, доповнена деревом модель і на навчальній, і на тестовій вибірках має кращі показники описаних вище характеристик. ROC-індекс і коефіцієнт GINI перебувають у межах допустимих значень, що свідчить про середню якість моделей. Слід зазначити, що для тестової вибірки наївної моделі ці характеристики набувають граничних значень.

***Таблиця* 2.** Значення характеристик навчальної вибірки наївної та доповненої деревом моделей, застосованих до даних German Bank Accepts

Модель	Відсоток неправильно класифікованих спостережень	ROC-індекс	Коефіцієнт GINI
Наївна	33,78	0,751	0,502
TAN	27,22	0,784	0,568

***Таблиця* 3.** Значення статистичних характеристик тестової вибірки наївної та доповненої деревом моделей, отриманих на даних German Bank Accepts

Модель	Відсоток неправильно класифікованих спостережень	ROC-індекс	Коефіцієнт GINI
Наївна	34,93	0,702	0,405
TAN	32,53	0,726	0,452

Для порівняння складності алгоритмів і витрат часу на побудову класифікаторів на практиці використаємо такі бази даних від SAS Enterprise Miner: CS_ACCEPTS, HMEQ, DMAGECR.

База даних CS_ACCEPTS описує процес кредитування 3000 фізичних осіб українського банку ВаБанк і містить 27 змінних (з яких 26 вхідних змінних і одна цільова). HMEQ містить інформацію щодо дефолту 5960 клієнтів німецького банку; для опису процесу використано тринадцять змінних (з яких дванадцять незалежних і одна цільова).

База DMAGECR містить фактичні статистичні дані за 2007 рік щодо кредитування 1000 фізичних осіб одного з найбільших банків Німеччини. Для опису процесу використано двадцять одну змінну (з яких двадцять вхідних і одна цільова). Всі вони містять інформацію про кредитування осіб за певною кількістю характеристик, що наведені в табл. 1.

У табл. 4 відображено витрати часу на навчання і максимальні загальні витрати часу для ймовірнісних моделей обох типів. Очевидно, що загальні витрати часу на реалізацію моделі у формі розширеного деревом (TAN) класифікатора у 10–15 разів перевищують аналогічні витрати на реалізацію простого класифікатора.

***Таблиця* 4.** Складність і витрати часу на побудову класифікаторів

База даних	Максимальна загальна кількість обчислень		Час навчання, с	
	наївна	TAN	наївна	TAN
German Bank Accepts	420	4200	0,04	0,06
CS_ ACCEPTS	540	5400	0,05	0,17
HMEQ	312	4056	0,03	0,04
DMAGECR	600	9000	0,02	0,03

Висновки

Подано математичне обґрунтування та опис алгоритмів реалізації наївного і доповненого деревом байєсівських класифікаторів. Ключовою особливістю доповненої деревом моделі є її уточнена деревоподібна структура. Для побудови дерева необхідно спочатку визначити батьків кожної змінної моделі. Крім того, тільки змінні з максимальною кореляцією повинні бути з'єднані дугами одна з одною. Ще одним важливим поняттям, що відіграє ключову роль при побудові дерева, є взаємна інформація. Вона необхідна для того, щоб визначити ребра, які будуть належати до дерева. Для того щоб побудувати дерево, оцінюється кореляція між усіма парами змінних у системі і додається ребро тільки між тими змінними, які найбільше корелюють. Для отримання деревоподібної структури, яка з'єднує всі вузли в графі, необхідно додати $N-1$ ребро, тобто на одиницю менше від кількості змінних. Крім того, сума вагових коефіцієнтів всіх цих ребер повинна мати максимальне значення серед всіх таких (можливих) деревовидних структур. Розроблено зручний для практичного застосування алгоритм побудови дерева, що складається з п'яти кроків, і продемонстровано його використання із залученням до аналізу фактичних даних.

Запропоновано підхід до визначення обчислювальної складності наведених алгоритмів.

На основі розрахунку умовних імовірностей усіх змінних імовірнісної моделі знайдено ймовірність появи кожного значення для всіх параметрів, обумовлених кожним станом цільової змінної. Це надало можливість визначити загальну кількість обчислень, необхідних для реалізації класифікаційних моделей обох типів. Через виконання обчислювальних експериментів визначено витрати часу на навчання і максимальні загальні витрати часу для обох типів імовірнісних моделей. Очевидно, що загальні витрати часу на реалізацію моделі у формі розширеного деревом (TAN) класифікатора перевищують у 10–15 разів аналогічні витрати на реалізацію простого класифікатора, але загалом час виконання складнішого алгоритму є цілком прийнятним для практичного використання.

Для визначення ефективності практичного застосування двох побудованих класифікаційних моделей виконано необхідні обчислювальні експерименти на чотирьох вибірках фактичних статистичних даних стосовно кредитування фізичних осіб (бази даних від SAS Enterprise Miner). Результати побудови і практичного застосування класифікаційних моделей у формі байєсівських мереж показали, що класифікатор, доповнений деревом, характеризується вищою складністю і витратами часу, але забезпечує більшу точність результатів у випадку, якщо існують кореляції між незалежними змінними. Так, відсоток неправильно класифікованих спостережень становив 34,93 % для спрощеної моделі і 32,53 % для моделі, розширеної деревоподібною структурою. Тобто похибка класифікації зменшилась на 2,4 %, що є добрим результатом. Також встановлено, що у випадку незначної кореляції між атрибутами (характеристиками) позичальників кредитів точність класифікації є практично однаковою для обох моделей.

Дослідження виконано в рамках проекту NATO “Безпека заради миру” NUKR.SFPP G4877 “Моделювання та попередження соціальних лих, спричинених катастрофами та тероризмом”, що виконується на базі Науково-навчального комплексу Інститут прикладного системного аналізу НТУУ “КПІ”.

У подальших дослідженнях планується застосування описаних підходів до моделювання соціально-економічних процесів з метою побудови уточнених математичних моделей, призначених для оцінювання прогнозів, які можна використати в органах місцевого самоврядування та державного управління з метою прийняття рішень з урахуванням невизначеностей різної природи і типів. Перспективним підходом є комбіноване застосування ідеологічно різних методів, наприклад регресійного аналізу та інтелектуального аналізу даних, регресійного аналізу та нейронечітких структур тощо.

Список літератури

1. *Shim J.K., Siegel J.G.* Schaum’s Outline of Theory and Problems of Financial Management. – New York: McGraw-Hill, 1998. – 517 p.
2. *Bouchard J.Ph., Potters M.* From Statistical Physics to Risk Management. – Cambridge: Cambridge University Press, 2000. – 218 p.
3. *Gallati R.* Risk Management and Capital Adequacy. – New York: McGraw-Hill, 2003. – 577 p.
4. *Darwiche A.* Modeling and Reasoning with Bayesian Networks. – Cambridge: Cambridge University Press, 2009. – 548 p.
5. *Korb K.B., Nicholson A.E.* Bayesain Artificial Intelligence. – London: CRC Press Company, 2004. – 365 p.
6. *Байєсівські* мережі в системах підтримки прийняття рішень / М.З. Згуровський, П.І. Бідюк, О.М. Терентьєв, Т.І. Просянкіна-Жарова. – К.: ТОВ “Видавниче підприємство “Едельвейс”, 2015. – 300 с.
7. *Neal R.M.* Probabilstic Inference Using MCMC Methods. – Toronto: University of Toronto, 1993. – 144 p.
8. *Padmanaban H.,* Comparative analysis of naïve Bayes and tree augmented naive bayes models: M.S. thesis, San José State University. – 2014. – 65 p.
9. *Naveen K., Sagar N., Deekshitulu Y.* Implementation of naive Bayesian classifier and ada-boost algorithm using maize expert system // Int. J. Inform. Sci. Techniques. – 2012. – № 3. – P. 63–75.
10. *Терентьев А.Н., Домрачев В.Н., Костецкий Р.И.* SAS BASE: Основы программирования. – К.: ТОВ “Видавниче підприємство “Едельвейс”, 2014. – 304 с.
11. *SAS* Enterprise Miner 13.2: Reference Help. SAS Documentation. – SAS Institute Inc., Cary, 2015. – 320 p.

References

1. J.K. Shim and J.G. Siegel, *Schaum’s Outline of Theory and Problems of Financial Management.* New York: McGraw-Hill, 1998, 517 p.
2. J.Ph. Bouchard and M. Potters, *From Statistical Physics to Risk Management.* Cambridge: Cambridge University Press, 2000, 218 p.

3. R. Gallati, *Risk Management and Capital Adequacy*. New York: McGraw-Hill, 2003, 577 p.
4. A. Darwiche, *Modeling and Reasoning with Bayesian Networks*. Cambridge: Cambridge University Press, 2009, 548 p.
5. K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*. London, UK: CRC Press Company, 2004, 365 p.
6. M. Zgurovsky *et al.*, *Bayesian Networks in Decision Support Systems*. Kyiv, Ukraine: Edelveis Publ., 2015, 300 p. (in Ukrainian).
7. R.M. Neal, *Probabilstic Inference Using MCMC Methods*. Toronto: University of Toronto, 1993, 144 p.
8. H. Padmanaban, "Comparative analysis of Naive Bayes and tree augmented naıve Bayes models", M.S. thesis, San José State University, 2014.
9. K. Naveen *et al.*, "Implementation of naïve Bayesian classifier and ada-boost algorithm using maize expert system", *Int. J. Inform. Sci. Techniques*, no. 3, pp. 63–75, 2012.
10. A. Terentyev *et al.*, *SAS BASE: Programming Basics*. Kyiv, Ukraine: Edelveis Publ., 2015, 304 p. (in Russian).
11. *SAS Enterprise Miner* 13.2: *Reference Help*, SAS Documentation, SAS Institute Inc., Cary, 2015, 320 p.

О.М. Терентьєв, В.Е. Кириченко, П.І. Бідюк, Т.І. Просянкіна-Жарова

ПРОГНОЗУВАННЯ ФІНАНСОВИХ РИЗИКІВ З ВИКОРИСТАННЯМ НАЇВНОГО І ДОПОВНЕНОГО ДЕРЕВОМ КЛАСИФІКАТОРІВ НА ОСНОВІ БАЙЄСІВСЬКИХ МЕРЕЖ

Проблематика. Побудова та дослідження характеристик наївного і доповненого деревом класифікаторів у формі байєсівських мереж при розв'язанні задачі оцінювання кредитного ризику.

Мета дослідження. Визначення точності класифікації позичальників кредиту банку за допомогою байєсівських класифікаторів двох типів. Математичне обґрунтування та опис алгоритмів реалізації обох моделей, а також обчислення їх алгоритмічної складності.

Методика реалізації. Розробка необхідного математичного апарату та виконання обчислювальних експериментів з метою побудови класифікаторів у формі байєсівських мереж на основі фактичних статистичних даних щодо кредитоспроможності фізичних осіб.

Результати дослідження. Створено методику побудови і використання наївного та доповненого деревом байєсівських класифікаторів при розв'язанні практичних задач оцінювання кредитоспроможності позичальників кредитів; виконано аналіз алгоритмічної складності розроблених алгоритмів; побудовано класифікаційні моделі у формі байєсівських мереж на основі фактичних статистичних даних із банківської системи та виконано порівняльний аналіз результатів застосування розроблених класифікаторів.

Висновки. Встановлено, що доповнений деревом класифікатор має більшу обчислювальну складність, ніж наївний байєсівський класифікатор, але він показує кращу точність результатів класифікації позичальників кредитів на дві групи: групу тих, хто повертає кредит, і проблемну.

Ключові слова: інтелектуальний аналіз даних; мережі Байєса; кредитний скоринг; фінансовий аналіз; макроекономічні показники.

А.Н. Терентьев, В.Э. Кириченко, П.И. Бидюк, Т.И. Просянкина-Жарова

ПРОГНОЗИРОВАНИЕ ФИНАНСОВЫХ РИСКОВ С ИСПОЛЬЗОВАНИЕМ НАИВНОГО И ДОПОЛНЕННОГО ДЕРЕВОМ КЛАССИФИКАТОРОВ НА ОСНОВЕ БАЙЕСОВСКИХ СЕТЕЙ

Проблематика. Построение и исследование характеристик наивного и дополненного деревом классификаторов в форме байесовских сетей при решении задачи оценивания кредитного риска.

Цель исследования. Определение точности классификации заемщиков кредита банка с помощью байесовских классификаторов двух типов. Математическое обоснование и описание алгоритмов реализации обеих моделей, а также вычисление их алгоритмической сложности.

Методика реализации. Разработка необходимого математического аппарата и выполнение вычислительных экспериментов с целью построения классификаторов в форме байесовских сетей на основе фактических статистических данных относительно кредитоспособности физических лиц.

Результаты исследования. Создана методика построения и использования наивного и дополненного деревом байесовских классификаторов при решении практических задач оценки кредитоспособности заемщиков кредитов; выполнен анализ алгоритмической сложности разработанных алгоритмов; построены классификационные модели в форме байесовских сетей на основе фактических статистических данных из банковской системы и выполнен сравнительный анализ результатов применения разработанных классификаторов.

Выводы. Установлено, что дополненный деревом классификатор имеет большую вычислительную сложность, чем наивный байесовский классификатор, но он показывает лучшую точность результатов классификации заемщиков кредитов на две группы: группу тех, кто возвращает кредит, и проблемную.

Ключевые слова: интеллектуальный анализ данных; сети Байеса; кредитный скоринг; финансовый анализ; макроэкономические показатели.

Рекомендована Радою
Навчально-наукового комплексу
"Інститут прикладного системного
аналізу" НТУУ "КПІ"

Надійшла до редакції
4 березня 20165 року