Економічні науки

УДК 519.86:336.717

Бень Владислав Петрович
провідний спеціаліст
Акціонерне товариство «МОТОР СІЧ»
Бень Владислав Петрович
ведущий специалист
Акционерное общество «МОТОР СИЧ»
Ben' Vladyslav
Leading Specialist
The Company «MOTOR SICH»

ЗАСТОСУВАННЯ АЛГОРИТМУ БУСТІНГУ ПРИ СТВОРЕННІ АНСАМБЛЮ МОДЕЛЕЙ ДЛЯ ВИЗНАЧЕННЯ СКОРИНГОВОЇ ОЦІНКИ ПОЗИЧАЛЬНИКІВ-ФІЗИЧНИХ ОСІБ
ПРИМЕНЕНИЕ АЛГОРИТМА БУСТИНГА ПРИ СОЗДАНИИ АНСАМБЛЯ МОДЕЛЕЙ ДЛЯ ОПРЕДЕЛЕНИЯ СКОРИНГОВОЙ ОЦЕНКИ ЗАЕМЩИКОВ-ФИЗИЧЕСКИХ ЛИЦ
APPLICATION OF BOOSTING ALGORITHM IN CREATING ANSEMBLE MODELS FOR MEASURE SCORING OF BORROWERS-INDIVIDUALS

Анотація. *Стаття присвячена застосуванню алгоритму бустінгу на основі використання нейромереж для розв'язання задачі скорингової оцінки позичальників банку.*

Ключові слова*: скоринг, бустінг, нейромережі.*

Аннотация. *Статья посвящена применению алгоритма бустинга на основе использования нейросетей при решении задачи скоринговой оценки заемщиков банка.*

***Ключевые слова**: скоринг, бустинг, нейросети.*

***Summary.** The article is devoted to the application of the bootstrap algorithm based on the use of neural networks in solving the problem of scoring the borrowers of the bank.*

***Key words:** scoring, boosting, neural networks.*

Постановка проблеми. Оцінка надійності позичальника банку є однією з важливих задач банківських установ. Недосконалі механізми оцінки кредитоспроможності здатні призвести до суттєво негативних наслідків. В найгіршому випадку – до банкрутства банківської установи. В інших, менш катастрофічних випадках, наслідком може бути таке негативне явище, як порушення ліквідності банку.

Найчастіше, для оцінки кредитоспроможності позичальників, банки застосовують скорингові моделі. Скоринговою моделлю називають математичну або статистичну модель, яка, на основі наявної інформації про попередній кредитний досвід позичальників банку, дає змогу оцінити кредитоспроможність потенційного клієнта [1]. При розробці скорингової оцінки застосовується широкий клас математичних моделей. Одним з видів моделей, що дають високу якість результатів є нейромережі. Однак, при розрахунках скорингової оцінки, в якості вхідних даних доцільно використовувати інформацію, що може бути представлена як кількісними, так і якісними показниками. До останніх відносяться анкетні дані позичальника, інформація від операторів мобільного зв'язку, відомості про активність потенційного клієнта у соціальних мережах тощо. Обробка такої специфічної інформації є складною проблемою. Тому не завжди вдається на її основі отримати достатній рівень точності класифікації позичальників. Саме з метою пошуку шляхів підвищення ефективності моделей пов'язана ідея створення ансамблів (комітетів) моделей.

Створення ансамблю покликано за допомогою кількох моделей, кожна з яких окремо має досить низьку точність оцінювання, утворити таку процедуру проведення класифікації, що дасть можливість отримати більш високий рівень загальної точності класифікації ансамблем моделей [2].

Аналіз останніх досліджень і публікацій. Теоретико-методологічна база для аналізу та дослідження загальних питань кредитного ризику та зокрема розробки скорингових моделей розвинута вітчизняними вченими Вітлінським В.В., Камінським А.Б., Кишакевичем Б.Ю., Пернарівським О.В., Писанцем К.К. Сучасний математичний інструментарій – методи нечіткої логіки та нейронних мереж в управлінні діяльності комерційного банку застосовано та розвинуто в роботі Великоіваненко Г.І., Трокоз Л.О. [3]. Однак потреби врахування специфіки інформаційного забезпечення процесу оцінювання кредитоспроможності клієнтів банку вимагають подальших досліджень. Одним з підходів в даному напрямку є використання ансамблів моделей.

Підходи до створення та застосування ансамблевих структур почали розвиватись у кінці минулого століття з роботи Р. Шепайре [4] де було вперше запропоновано ідею бустінгу. Успішність подальших модифікацій початкової ідеї бустінгу та створення нових алгоритмів роботи комітетів моделей дали поштовх до використання ансамблів у різних сферах досліджень. Однак, лише останнім часом комітети моделей почали використовувати також і для розв'язання задачі кредитного скорингу. Тому кількість публікацій за даною темою досить обмежена. В роботі [5] розглянуто застосування алгоритму бустінгу для проведення кредитного скорингу на основі дерев рішень. Використання дерев рішень є одним з найбільш простих методів при розробці скорингових моделей, тому недоцільно обмежуватись лише результатами таких досліджень. Цікавою є робота [6], в якій описано процедуру розробки ансамбля для розв'язання задачі проведення поведінкового скорингу. В роботі наведено основні

аспекти, що впливають на підвищення точності роботи ансамблевих структур та розглянуто реалізацію одного з таких аспектів, а саме – метод узагальнення результатів окремих моделей ансамблю. Отже залишаються відкритими для дослідження інші питання. Наприклад, проблеми вибору окремих моделей, які складатимуть комітет.

Постановка завдання. Метою роботи є дослідження процесу застосування базового варіанту алгоритму бустінгу на основі використання нейромереж в якості окремих моделей ансамблю при розв'язанні задачі скорингової оцінки позичальників-фізичних осіб.

Виклад основного матеріалу. На сьогодні розроблено та описано значну кількість різноманітних видів ансамблів, які різняться за алгоритмами побудови. В процесі удосконалення таких алгоритмів деякі з них, наприклад бустінг, мають кілька модифікацій та вже перетворились у окремі сімейства алгоритмів.

За алгоритмом бустінгу (підсилення) моделі ансамблю будуються послідовно таким чином, щоб кожна наступна модель проводила класифікацію тих прикладів, які не змогли класифікувати моделі на попередніх кроках. Для цього спеціальною процедурою проводиться формування навчальних вибірок. Експерти комітету навчаються послідовно на різних масивах початкових даних. Однаковим для всіх експертів є лише обсяг навчальної вибірки. Таким чином, кожен з експертів не може повторити помилки попереднього, що забезпечує незалежність похибок окремих моделей комітету. А це, в свою чергу, є однією з умов підвищення ефективності роботи комітету [2].

Статистичною базою дослідження є дані з кредитних заявок позичальників-фізичних осіб комерційного банку та відомостей щодо виконання ними зобов'язань за отриманими кредитами. Представлена інформація містить дані за 6 чинниками та нараховує 2175 спостережень.

Досліджувався базовий варіант алгоритму бустінгу [2]. Всі наявні дані було розділено на дві частини – навчальну та тестову вибірки. Навчальна вибірка складається з 600 значень, решта 1575 використовуються як тестова вибірка. Незначний розмір навчальної вибірки обумовлений недостатньо великим обсягом наявного масиву даних, щоб забезпечити паритет надійних та дефолтних позичальників. Зауважимо, що у літературі [2, 7] вказується на необхідність значного обсягу даних для застосування даного алгоритму. Деталізовану схему алгоритму бустінгу для досліджуваних даних проілюстровано на рис. 1.

На першому кроці навчається перший експерт комітету, для якого обсяг навчальної вибірки сформований із 600 спостережень. За результатами класифікації першим експертом формується навчальна вибірка для другого експерта. Вона повинна мати той самий обсяг, що і початкова вибірка, та містити рівно половину її прикладів, всі з яких були точно класифіковані першим експертом. Тобто, в даному випадку слід обрати по 150 надійних (0) та дефолтних (1) позичальників. Інші 300 прикладів для навчальної вибірки другого експерта випадковим чином обираються з тестового масиву. Залишок елементів з тестового масиву першого експерта стає тестовим масивом для другого експерта (в нашому дослідженні тестова вибірка для другого експерта складатиметься з 1275 значень).

За результатами обчислень, проведених першим та другим експертами, формується навчальна вибірка для третього експерта. Вона має містити також 600 прикладів, до яких із застосованих раніше мають увійти лише ті, за якими результати класифікації першим та другим експертами відрізняються. В начальній вибірці таких розбіжностей серед надійних позичальників виявлено 75, а з-поміж дефолтних – 115. Отже, для забезпечення потрібного обсягу навчальної вибірки для третього експерта необхідно взяти із тестової вибірки другого експерта ще 225 прикладів

надійних позичальників (які характеризуються нульовими значеннями) та 185 – дефолтних (що представлені одиницями). Залишок тестової вибірки другого експерта, який склав 865 елементів, стає тестовою вибіркою для третього експерта.

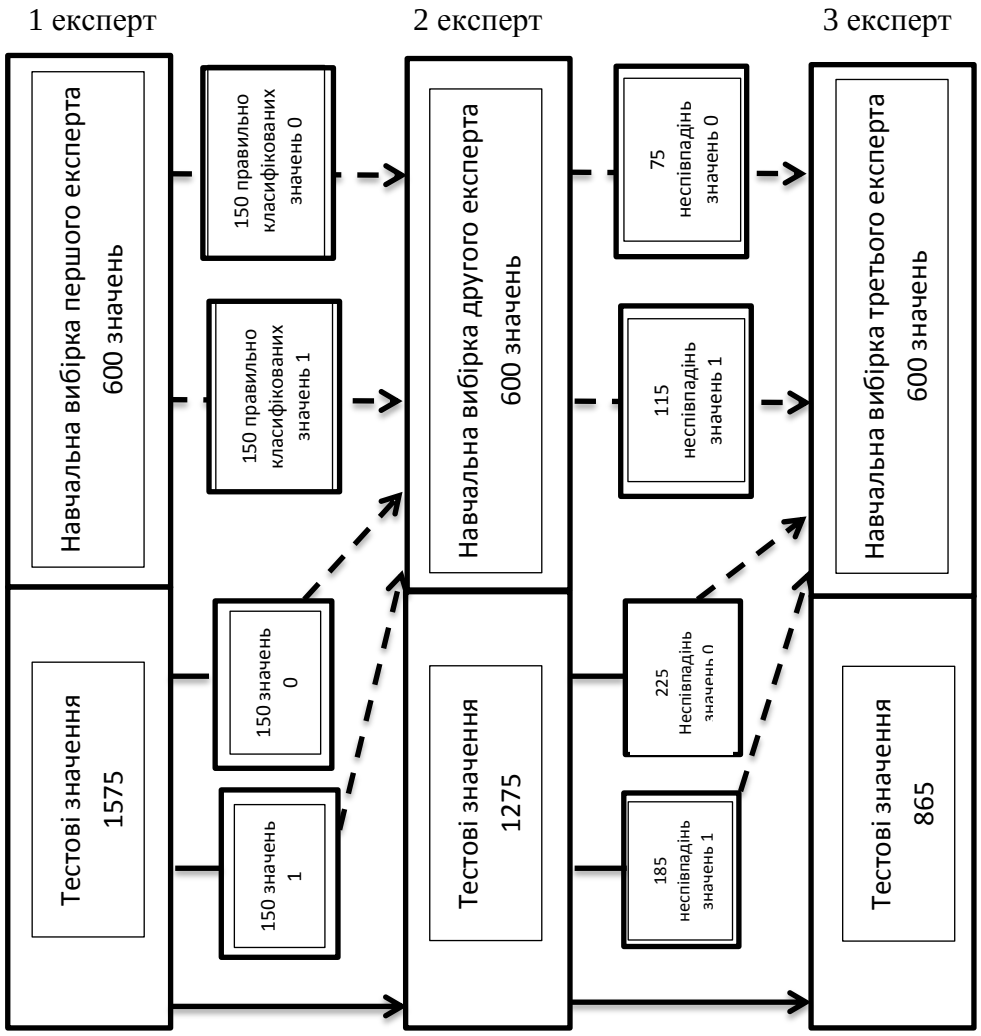

Рис. 1. Схема алгоритму бустінгу для навчальної вибірки з 600 значень

Джерело: розробка автора

Першим експертом перш за все було взято нейромережу радіально-базисної архітектури з 124 нейронами проміжного шару. Така мережа в проведеному раніше дослідженні [8] демонструвала найкращий результат для незначних обсягів навчальної вибірки. З метою обґрунтованого вибору

першого експерта разом з цією моделлю було розглянуто ще дві нейромережі з високою точністю класифікації (див. табл. 1).

Таблиця 1

Розрахунки ефективності нейромереж для вибору параметрів моделі-першого експерта за алгоритмом бустінгу

Архітектура мережі, кількість входів, кількість нейронів проміжного шару	Відсоток правильно класифікованих спостережень у навчальній вибірці, %	Відсоток правильно класифікованих спостережень у тестовій вибірці, %	Узагальнені дані по всьому масиву (без поділу на навчальну та тестову вибірки)		Правильно класифікованих, %
			Клас	Всього	
Тришаровий персептрон, 6, 6	75,5	52,5	0	1131	56,5
			1	1044	61,3
Радіально-базисна, 6, 124	81,7	53	0	1131	58,4
			1	1044	63,3
Радіально-базисна, 6, 187	84,3	50	0	1131	57,7
			1	1044	62,5

Джерело: розробка автора

На основі даних табл. 1 можна зробити висновок, що обраний обсяг навчальної вибірки не надає можливості якісного навчання моделі. Про це свідчить значна розбіжність між показниками точності класифікації для навчальної (близько 80 %) та тестової (не перевищує 53 %) вибірок. Тобто модель, налаштована на невеликій навчальній вибірці, не буде достатньо гнучкою для ефективного моделювання усього різноманіття варіантів із тестової вибірки. Для підвищення точності класифікації можна збільшити обсяг навчальної вибірки, однак в даному дослідженні такої можливості не було через незначний обсяг наявних даних.

Відповідно до табл. 1, найефективнішою моделлю можна вважати нейромережу з радіально-базисною архітектурою та 124 нейронами проміжного шару як за показниками класифікації окремо для навчальної та тестової вибірок, так і за узагальненими даними по всьому масиву значень.

При виборі другого та третього експертів проводились аналогічні попередньому випадку експериментальні розрахунки.

У якості другої моделі-експерта було обрано нейромережу з радіально-базисною архітектурою та 13 нейронами проміжного шару. Вона має вищі показники точності класифікації порівняно з іншими моделями як у розрізі навчальної та тестової вибірок, так і за узагальненими даними. В якості третього експерта було обрано нейромережу з радіально-базисною архітектурою та 28 нейронами проміжного шару.

Загальний результат роботи комітету моделей визначається простим голосуванням: приклад відноситься до того класу, до якого його віднесено трьома чи двома експертами. На всьому масиві даних (з 1465 значень) правильно класифікованих комітетом надійних позичальників, що представлені нулями, виявилось 58,68 %, дефолтних (одиниць) – 56 %.

Для можливості співставлення ефективності роботи комітету з окремими його моделями в табл. 2 наведено показники точності класифікації першого, другого та третього експертів окремо та спільно для масиву даних із 1465 значень.

Таблиця 2

Показники ефективності класифікації окремими експертами та комітетом, сформованим за алгоритмом бустінгу

Архітектура мережі, кількість входів, кількість нейронів проміжного шару	Узагальнені дані по всьому масиву (без поділу на навчальну та тестову вибірки)		Правильно класифікованих, %
	Клас	Всього	
Радіально-базисна, 6, 124	0	756	49
	1	709	61,6
Радіально-базисна, 6, 13	0	756	53,6
	1	709	52,8
Радіально-базисна, 6, 28	0	756	55
	1	709	55
Комітет моделей	0	756	58,6
	1	709	56

Джерело: розробка автора

Дані табл. 2 підтверджують, що точність класифікації комітетом є вищою за окремі його моделі. Однак, все одно відсоток правильно класифікованих комітетом прикладів є досить невисоким.

На нашу думку, головною проблемою в даному випадку є неефективність використання комітетів моделей для вибірок малого обсягу. Як зазначається в [2, 7], призначення комітетів моделей – обробка даних, що нараховують 100 тисяч і більше записів. У даному випадку малий обсяг навчальної вибірки не дає достатніх можливостей для ефективного навчання навіть першого експерта. А при подальшій послідовній зміні навчальної вибірки з неї взагалі вилучаються певні її значущі частини, що призводить до її виродження. Це є відомим недоліком даного методу [7]. Крім того, в дослідженні було реалізовано найпростіший варіант з усіх модифікацій сімейства алгоритмів бустінгу.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі. На сьогоднішній день, для проведення скорингової оцінки позичальників банку, слід враховувати великі обсяги різнорідної інформації. Таке завдання вимагає пошуку нових методів та підходів щодо обробки масивів даних з особливою специфікою. Одним з напрямків розв'язання цієї проблеми є застосування ансамблевих структур. В якості окремих моделей ансамблю можна використовувати різні види математичних моделей. Процес відбору окремих моделей до комітету вимагає проведення ряду експериментальних розрахунків, оскільки вибір окремих експертів значною мірою впливатиме на ефективність його роботи. Результати розрахунків підтверджують, що реалізація алгоритму бустінгу дозволяє підвищити точність класифікації позичальників банку. Однак для коректного застосування алгоритму бустінгу слід використовувати початкові данні більшої розмірності.

Отримані результати досліджень можуть бути основою для реалізації інших модифікацій алгоритму бустінгу або інших алгоритмів при створенні анасамблів моделей.

Література

1. Сорокин А. С. Построение скоринговых карт с использованием модели логистической регрессии. [Электронный ресурс] / А.С. Сорокин / Интернет-журнал «Науковедение». – 2014. – Вып. 2. – С. 1–29. – Режим доступу: http://naukovedenie.ru/PDF/180EVN214.pdf.
2. Научная сессия МИФИ-2007. IX Всероссийская научно-техническая конференция «Нейроинформатика-2007»: Лекции по нейроинформатике / [авт.тексту С.А.Терехов]. – Часть 2. – М.: МИФИ, 2007. – 148 с.
3. Великоіваненко Г. І. Нейро-нечітка модель оцінювання прострочених позик комерційного банку / Г.І. Великоіваненко, Л.О. Трокоз // Нейро-нечіткі технології моделювання в економіці. - 2014. - № 3. - С. 23-66.
4. Robert E. Schapire. Theoretical views of boosting and applications / Algorithmic Learning Theory, 10th International Conference, ALT '99, Tokyo, Japan, December 1999, Proceedings [Електронний ресурс]. — Режим доступу : http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG445/literatur/schapire_99a.pdf.
5. Bastoi J. Credit scoring with boosted decision trees / MPRA Paper No. 8156, posted 8. April 2008. [Електронний ресурс]. — Режим доступу : https://mpra.ub.uni-muenchen.de/8156/1/paper.pdf
6. Кузнецов И.А. Разработка ансамбля алгоритмов классификации с использованием энтропийного показателя качества для решения задачи поведенческого скоринга / И.А.Кузнецов, В.С.Киреев // Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным

использованием данных», Ершово, 11-14 октября 2016. [Електронний ресурс]. — Режим доступу : http://ceur-ws.org/Vol-1752/paper07.pdf

7. Паклин Н. Б. Бизнес-аналитика: от даннях к знаниям: Учебное пособие. 2-е изд. испр. / Н.Б. Паклин, В.П. Орешков. – СПб: Питер. 2013. – 704 с.
8. Савіна С. С. Вибір архітектури нейромережі для розв'язання задачі класифікації надійності позичальників-фізичних осіб / С. С. Савіна, В. П. Бень // Нейро-нечіткі технології моделювання в економіці. - 2015. - № 5. - С. 123-151.

References

1. Sorokin, A. S. (2014). Postroenie skorinhovih kart s ispol'zovaniem modeli logisticheskoy regressii. Internet-zhurnal «Naukovedenie», 2. Retrieved from http://naukovedenie.ru/PDF/180EVN214.pdf.
2. Nauchnaya sessiya MIFI-2007. IX Vserossiyskaya nauchno-tekhnicheskaya konferentsiya «Neyroinformatika-2007»: Lektsii po neyroinformatike. / [avt.tekstu S.A.Terekhov]. – Chast 2. – M.: MIFI, 2007. – 148 S.
3. Velykoivanenko, H. I., & Trokoz, L. O. (2014). Nejro-nechitka model' otsiniuvannia prostrochenykh pozyk komertsijnoho banku. Nejro-nechitki tekhnolohii modeliuvannia v ekonomitsi (Neuro-Fuzzy Modeling Technigues in Economics), 3, 23—66.
4. Robert E. Schapire. Theoretical views of boosting and applications // Algorithmic Learning Theory, 10th International Conference, ALT '99, Tokyo, Japan, December 1999, Proceedings [Elektronnyj resurs]. — Rezhym dostupu: http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG445/literatur/schapire_99a.pdf.

5. Bastoi J. Credit scoring with boosted decision trees // MPRA Paper No. 8156, posted 8. April 2008. [Elektronnyj resurs]. — Rezhym dostupu: https://mpra.ub.uni-muenchen.de/8156/1/paper.pdf.
6. Kuznetsov I.A. Razrabotka ansamblya algoritmov klassifikatsii s ispolzovaniem entropiynogo pokazatelya kachestva dlya resheniya zadachi povedencheskogo skoringa / I.A.Kuznetsov, V.S.Kireev // Trudy XVIII Mezhdunarodnoy konferentsii DAMDID/RCDL'2016 «Analitika i upravlenie dannymi v oblastyakh s intensivnym ispolzovaniem dannykh», Yershovo, 11-14 oktyabrya 2016. [Yelektronniy resurs]. — Rezhim dostupu : http://ceur-ws.org/Vol-1752/paper07.pdf.
7. Paklin N. B. Biznes-analitika: ot dannyakh k znaniyam: Uchebnoe posobie. 2-e izd. ispr. / N.B.Paklin, V.P.Oreshkov – SPb: Piter. 2013. – 704 s.
8. Savina S. S. Vybir arkhitektury nejromerezhi dlja rozv'jazannja zadachi klasyfikaciji nadijnosti pozychaljnykiv-fizychnykh osib / S. S. Savina, V. P. Benj // Nejro-nechitki tekhnologhiji modeljuvannja v ekonomici. — 2015. — # 5. — S. 123—151.