МАТЕМАТИЧНІ МЕТОДИ, МОДЕЛІ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЕКОНОМІЦІ

УДК 330.46

Класифікація об'єктів консалтинговою компанією з урахуванням асиметричності інформації

Бачало І.Б.
аспірант кафедри економічної кібернетики
Львівського національного університету імені Івана Франка

У статті розглянуто використання методів економіко-математичного моделювання для управління ризиками консалтингової діяльності в умовах асиметричної інформації на прикладі вирішення задачі класифікації економічних суб'єктів на вибірці з незбалансованими класами. Детально описано метрики, що можуть застосовуватися для оцінювання роботи бінарного класифікатора та використовуються на незбалансованих даних, а також наведено приклад використання алгоритмів із відновлення міжкласової рівноваги для виявлення шахрайств на даних однієї з українських страхових компаній. Автор пропонує використовувати поєднання алгоритмів генерування мінорантного класу та видалення мажорантного класу для реалізації стратегії страхової компанії з оптимізації залучення нових клієнтів.

Ключові слова: асиметрична інформація, консалтинг, незбалансовані дані, класифікація, страхова компанія, виявлення шахрайства, SMOTE, ENN, BalanceCascade, EasyEnsemble.

Бачало И.Б. КЛАССИФИКАЦИЯ ОБЪЕКТОВ КОНСАЛТИНГОВОЙ КОМПАНИЕЙ С УЧЕТОМ АСИММЕТРИЧНОСТИ ИНФОРМАЦИИ

В статье рассмотрено использование методов экономико-математического моделирования для управления рисками консалтинговой деятельности в условиях асимметричной информации на примере решения задачи классификации экономических субъектов на выборке с несбалансированными классами. Подробно описаны метрики, которые могут применяться для оценки работы бинарного классификатора и используются на несбалансированных данных, а также приведен пример использования алгоритмов по восстановлению межклассового равновесия для выявления мошенничеств на данных одной из украинских страховых компаний. Автор предлагает использовать сочетание алгоритмов генерации минорантного класса и удаления мажорантного класса для реализации стратегии страховой компании по оптимизации привлечения новых клиентов.

Ключевые слова: асимметричная информация, консалтинг, несбалансированные данные, классификация, страховая компания, выявления мошенничества, SMOTE, ENN, BalanceCascade, EasyEnsemble.

Bachalo I.B. THE CLASSIFICATION OF OBJECTS BY A CONSULTING COMPANY WITH TAKING INTO ACCOUNT THE ASYMMETRY OF INFORMATION

The use of methods of economic-mathematical modeling for risk management consulting activity under conditions of asymmetric information on the example of solving the task of economic subjects' classification on a sample with unbalanced classes is reviewed in the article. The work describes in detail the metrics that can be used to evaluate the productivity of the binary classifier and are used on unbalanced data. There is also the example of using algorithms to restore interclass equilibrium to detect fraud on the data of one Ukrainian insurance company. The author suggests using a combination of algorithms for generating a minority class and the removal of the majority class to implement the strategy of the insurance company to optimize the attraction of new customers.

Keywords: asymmetric information, consulting, imbalanced data, classification, insurance company, fraud detection, SMOTE, ENN, BalanceCascade, EasyEnsemble.

Постановка проблеми у загальному вигляді. Консалтингові компанії, що мають справу з аналізом даних та математичним моделюванням, під час надання своїх рекомендацій можуть стикатися із завданням класифікації економічних суб'єктів. Як правило, ці суб'єкти – це клієнти, постачальники, працівники або партнери компанії-клієнта, яка звернулася за вирішенням своєї проблеми до консалтерів. Необхідність вирішення задачі класифікації виникає через потребу управління ризиками з урахуванням інформаційної асиметрії, ситуації, коли компанія-клієнт намагається працювати із суб'єктами економічної діяльності, що є краще поінформованими за неї. Як приклад тут можна навести

кредитні відділи, що намагаються класифікувати позичальників, чи ті будуть виплачувати кредит, чи ні; страхові компанії, що хочуть класифікувати страхувальників за різними групами ризику; цифрові рекламні агенції, що бажають визначити користувачів, які купуватимуть рекламовані продукти, тощо. За умови наявності достатнього обсягу даних та їхньої збалансованості (пропорції кожного з класів у даних не надто відрізняються) такі задачі легко вирішуються методами економіко-математичного моделювання. Проте якщо різниця у пропорціях розподілу класів у наявній статистиці надто велика, тобто дані мінорантного класу є надто розрідженими, то точність прогнозу приналежності нового спостереження до класу меншості буде низькою, що за умови високої ціни помилки може призвести до фінансових збитків або втрати репутації консалтингової компанії.

Аналіз останніх досліджень і публікацій. У комп'ютерних науках математичним розв'язанням задачі класифікації займалися К. Миругін [1–3], Е. Настенко [4], А. Кардаш [5], Н. Волошин [6] та ін. Дослідженнями використання класифікації для вирішення бізнесових питань займалися багато науковців, зокрема: М. Жук [7], І. Пістунов [8], Д. Мойсеєнкова [9], А. Литвин [10], С. Бакун [11], І. Гюнтер [12], О. Мінц [13], А. Матвійчук [14], О. Солошенко [15], О. Кожухівська [16], Н. Кузнєцова [17]. Проте у роботах цих учених не було досліджено вплив розріджених даних на якість проведеної класифікації. Розв'язки задач розріджених даних шукали Лю Нін [18], В. Нітеш [19], О. Отман [20], Д. Гуан [21], Р. Шапіре [22] та ін., але у їхніх роботах не приділялася увага застосуванню описаних методів до бізнесових завдань. Щодо застосування методів роботи з розрідженими даними до прикладних проблем, то часто такі задачі розв'язуються у медицині, біоінформатиці та задачах кредитного скорингу, якими займалися Н. Паклін [23], В. Вінціотті [24], але ці дослідження були орієнтованими на вузький спектр вирішуваних задач і не можуть покрити потреби вирішення завдань консалтингових компаній, які часто стикаються з асиметричністю інформації та нестачею даних по мінорантних класах.

Формулювання цілей статті (постановка завдання). Метою статті є розроблення математичного методу врахування асиметричності інформації консалтинговою компанією, що базується на використанні класифікаторів у поєднанні з техніками роботи з розрідженими даними, та практична демонстрація його роботи на прикладі виявлення шахрайства у страховій компанії.

Виклад основного матеріалу дослідження. Консалтингові компанії, що працюють із даними, можуть уважатися посередниками, що усувають асиметричність інформації між компанією-клієнтом та середовищем, яке потрібно дослідити консалтерам [25]. Цим середовищем може бути як набір процесів та інформаційних потоків усередині компанії, так і її зовнішнє оточення. Під час проведення консалтингового дослідження, частиною якого може бути класифікація економічних суб'єктів, що взаємодіють із компанією-клієнтом, може виникнути проблема нестачі статистичних даних. Така ситуація може виникати навіть тоді, коли самих даних багато, але немає достатньо інформації про мінорантний клас, який важливо класифікувати. Неможливість проведення класифікації таких даних може породити ризики для консалтерів і бути причиною краху консалтингового проекту і, відповідно, причиною втрати репутації консалтинговою компанією. Для керування такими ризиками необхідно використовувати методи роботи з розрідженими даними.

Проте для того щоб оцінити якість класифікації, що проводитиметься на вибірці з розрідженими даними, потрібно використовувати метрику, яка б могла відобразити ефективність класифікатора під час роботи з мінорантним класом.

Наведемо гіпотетичний приклад роботи класифікатора на розріджених даних, щоб пояснити використання різних метрик його оцінювання. Нехай є навчальна вибірка з 100 тис. клієнтів страхової компанії і лише 1 тис. з них є шахраями, які робили фальшиві заяви щодо настання страхового випадку. В результаті отримуємо ситуацію міжкласового дисбалансу [26] з пропорцією 100:1 (варто сказати, що часто такі пропорції у даних можуть становити 1000:1, 10000:1 та більше). Припустимо, що був використаний бінарний класифікатор для виявлення шахрайства, який не був готовий до роботи з розрідженими даними. На даних було отримано прогнозні значення, де було вірно класифіковано 300 шахраїв з 1 тис., та 98 100 добросовісних клієнтів з 99 тис. Таблично дану ситуацію зображено рис. 1.

Отримана матриця має назву таблиці помилок (англ. Confusion Matrix) та використовується в задачах бінарної класифікації для відображення вірно та помилково класифікованих значень (рис. 2).

Якщо скористатися загальноприйнятою метрикою вимірювання точності роботи кла-

		Фактичний клас	
		Шахрай	**Добросовісний клієнт**
Прогнозований клас	**Шахрай**	300	900
	Добросовісний клієнт	700	98100

Рис. 1. Приклад розподілу клієнтів компанії після класифікації

Джерело: побудовано автором

		Фактичний клас	
		Позитивне значення класу (P)	**Негативне значення класу (N)**
Прогноз	**Позитивне передбачення класу**	Вірно передбачений клас (TP)	Невірно позитивне значення (FP)
	Негативне передбачення класу	Невірно негативне значення (FN)	Вірно знехтуваний клас (TN)

Рис. 2. Таблиця помилок для бінарної класифікації

Джерело: узагальнено автором

сифікатора (англ. Accuracy) (1), то отримане значення покаже достатньо хорошу якість проведеної класифікації – 0,984. Але майже таке ж число (0,981) могло бути досягнутим, якщо б абсолютно всіх клієнтів називали добросовісними.

$$ACC = \frac{TP + TN}{FP + FN + TP + TN}. \quad (1)$$

Проте якщо маємо ціль визначити саме шахраїв, то дана метрика не є репрезентативною, адже частка правильно класифікованих шахраїв, що були виявлені серед усіх шахраїв, надто мала. Дане співвідношення можна зобразити за допомогою метрики відгуку, яку ще називають чутливістю (англ. Recall / Sensitivity) (2), яка для нашого прикладу становитиме 0,3.

$$REC = \frac{TP}{P} = \frac{TP}{FN + TP}. \quad (2)$$

Оберненим показником до цього є похибка другого роду, яку позначають β та ще іноді називають пропуском події (*англ.* False Negative Rate) (3):

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - REC. \quad (3)$$

У нашому випадку її значення становитиме 0,7.

Також обчислюють показник прецизійності (англ. Precision) (4), який показує відношення вірно класифікованих шахраїв до усіх клієнтів, що були позначені класифікатором як шахраї.

$$PRE = \frac{TP}{TP + FP}. \quad (4)$$

Для наведених даних це значення становитиме 0,25.

Досить часто показники точності та чутливості об'єднують у метрику F1 (англ. F1-score), яка є їхнім середнім гармонійним та обчислюється за формулою (5):

$$F1 = 2\frac{PRE \times REC}{PRE + REC}. \quad (5)$$

Для нашого класифікатора її значення становитиме 0,273. Варто зазначити, що чим ближче значення F1 до 1, тим краще класифікатор справився з поставленим завданням. Однак дана метрика не бере до уваги значень вірно знехтуваного класу (TN), що потрібно враховувати в деяких дослідженнях.

Також розраховують середнє геометричне для точності та чутливості й позначають його G (англ. G-measure) (6):

$$G = \sqrt{PRE \times REC}. \quad (6)$$

У нашому прикладі цей показник приймає значення 0,274.

Якщо б якість роботи класифікатора на мінорантному класі поліпшилася удвічі, наприклад, показник вірно класифікованих шахраїв збільшився з 300 до 600, за інших рівних умов, то показник точності майже не зміниться, з 0,984 до 0,987, проте значення чутливості зросте з 0,3 до 0,6, похибка другого роду, відповідно, зменшиться з 0,7 до 0,4, а точність зросте з 0,25 до 0,4. Усереднені метрики теж зростуть: показник F1 – з 0,273 до 0,48, а G – з 0,274 до 0,49.

Як можна побачити, використання метрик, що враховують незбалансованість даних між

класами, дає змогу краще оцінювати роботу класифікатора, націленого на виявлення мінорантного класу.

Загалом у комп'ютерних науках, зокрема у галузях знань, що працюють із машинним навчанням (англ. Machine Learning), для роботи з незбалансованими даними використовують методи збільшення мінорантного класу (англ. Oversampling) та видалення значень мажорантного класу (англ. Undersampling) [27]. Загальноприйняті методики передбачають випадкове видалення значень мажорантного класу або випадкове дублювання даних для мінорантного класу. Проте існують техніки, що використовують складніші алгоритми для роботи із задачею розрідженості даних. Зокрема можна виділити такі алгоритми для зменшення мажорантного класу: NearMiss-1, NearMiss-2, NearMiss-3, Condensed Nearest Neighbor (CNN), Edited Nearest Neighbor (ENN), Tomek Link Removal та ін. У літературі також описуються техніки, що базуються на використанні ансамбельних методів для зменшення кількості даних превалюючого класу: EasyEnsemble та BalanceCascade. Для штучної генерації мінорантного класу використовують такі алгоритми: Synthetic Minority Oversampling Technique (SMOTE) та його модифікації Borderline-SMOTE та Adaptive Synthetic Sampling (ADASYN), а також кластерні методи збільшення даних мінорантного класу Cluster-Based Oversampling (CBO) [26].

Автор дослідження пропонує провести тестування кількох із вищеописаних алгоритмів на справжніх даних однієї з українських страхових компаній та порівняти їх між собою, а також співставити отримані результати з базовим класифікатором, який працюватиме без застосування згаданих методів.

Для тесту була використана база даних із 10 тис. клієнтів однієї з українських страхових компаній, яким проводилися страхові виплати та серед яких були присутні 500 шахраїв (співвідношення 1:20). Таблиця складалася з 18 змінних: вік, стать страхувальника, тип страховки, тривалість страховки, вартість страховки, вартість застрахованого майна, франшиза, категорія застрахованого майна, місце укладання угоди, час від купівлі страховки до настання страхового випадку, час початку страхування, час закінчення страхування, місце настання страхового випадку, тип настання інциденту, причина настання інциденту, час від настання інциденту до звернення у страхову компанію, місце настання інциденту, бінарна змінна, чи страхувальник був клієнтом страхової компанії в минулому.

Вибірку було поділено випадковим чином на тренувальну та тестову у відношенні 70:30 відповідно. Базовими класифікаторами, що використовувалися, були дерево рішень (англ. Decision Tree) та ансамбельний метод AdaBoost, який базувався на тому ж таки дереві рішень.

Далі було проведено тестування об'єднаного алгоритму збільшення даних мінорантного класу SMOTE в парі з алгоритмом зменшення даних мажорантного класу ENN та застосовано класифікатор AdaBoost, що базувався на деревах рішень, до новоутвореного набору даних. Також було здійснено порівняння продуктивності попереднього методу з методами EasyEnsemble та BalanceCascade, що базувалися на побудованому раніше ансамбельному методі AdaBoost.

Підбір оптимальних параметрів для класифікаторів здійснювався за допомогою вбудованої функції GridSearchCV в бібліотеку для наукових обчислень scikit-learn у мові програмування Python. Параметрами що оптимізувалися, були: інформаційний критерій розділення – коефіцієнт Джині або показник ентропії; вибір точки розділення – випадково або з допомогою інформаційного критерію; кількість побудованих дерев рішень від 1 до 100.

Ціллю для оптимізації параметрів була метрика F1, вибірка для побудови моделі розбивалася на п'ять частин за крос-валідації, кількість проведених випробувань для кожної моделі – 8.

У тренувальній вибірці після розділення випадково було взято 7 тис. клієнтів, серед яких був 351 шахрай (співвідношення 1:20 збереглося). У тестовій вибірці, на якій показуватимуться результати роботи моделей, залишилося 3 тис. клієнтів, серед яких було 149 шахраїв (співвідношення класів відповідно теж збереглося).

Для візуалізації отриманих результатів класифікації багатовимірного набору даних із метою зменшення розмірності використовувався метод головних компонент (англ. Principal Component Analysis), за допомогою якого 18-вимірний набір даних перетворювався на 2-вимірний та візуалізувався за допомогою точкової діаграми. На всіх малюнках жовтими кругами позначався мажорантний клас, синіми – мінорантний, а червоними трикутниками – невірні спрацювання класифікатора. Весь набір даних виглядає так (рис. 3).

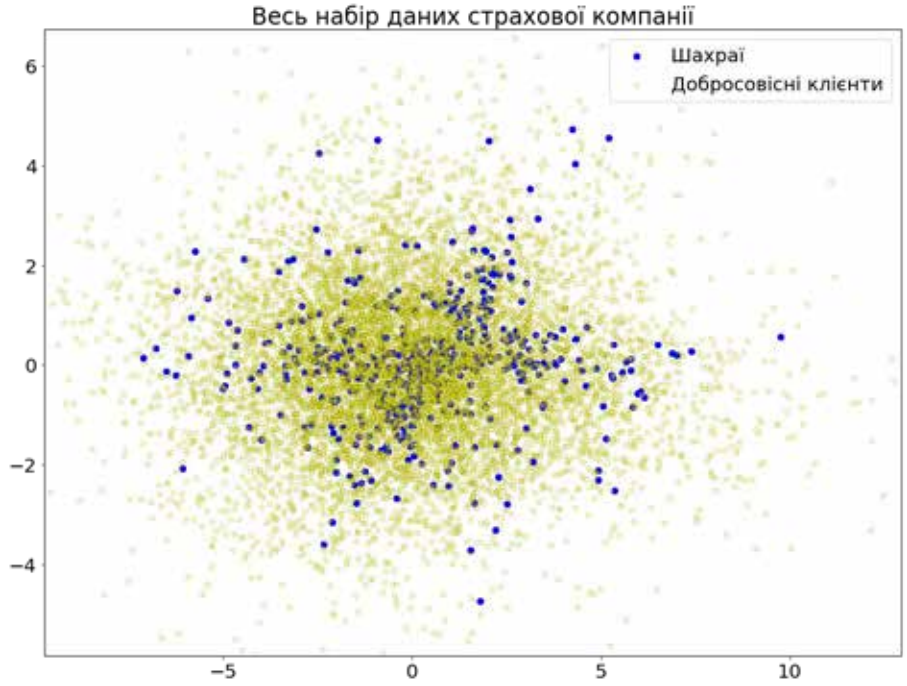

Рис. 3. Візуалізація даних страхової компанії у двовимірному просторі

Джерело: побудовано автором

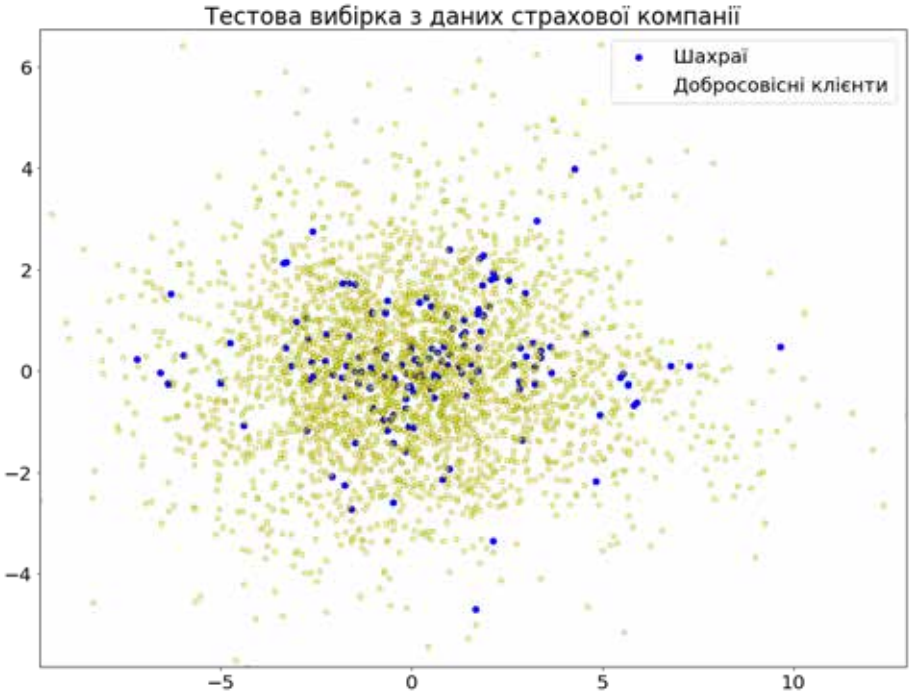

Рис. 4. Візуалізація тестової вибірки з даних страхової компанії

Джерело: побудовано автором

Візуалізація тестової вибірки зображена на рис. 4.

Результати роботи дерева рішень – Decision Tree та ансамбельного методу AdaBoost без використання алгоритмів роботи з незбалансованими даними були такими (табл. 1).

Як бачимо, дерево рішень дало кращий результат класифікації, ніж ансамбельний метод. Проте чутливість роботи класифі-

катора до даних мінорантного класу надто мала, а висока точність роботи досягається за рахунок великої кількості даних мажорантного класу, які легко класифікувати. Графічне зображення результатів роботи дерева рішень на мінорантному класі зображено на рис. 5.

Наступним кроком було використання методу EasyEnsemble для роботи з проблемою розріджених даних. EasyEnsemble відбирає декілька незалежних підмножин даних N_1, N_2, … N_T з усієї вибірки мажорантного класу N. Для кожної підмножини $N_i\,(1 \leqslant i \leqslant T)$ тренується класифікатор H_i з використанням N_i і P – множини даних мінорантного класу.

Алгоритм EasyEnsemble можна описати такими кроками [28]:

1. Вхідні дані: множина даних мінорантного класу P, множина даних мажорантного класу N, $|P| < |N|$, і T, кількість підмножин, що повинні бути вибрані з N.
2. Присвоюємо $i \leftarrow 0$.
3. Початок циклу:
4. $i \leftarrow i+1$.
5. Випадково генеруємо підмножину N_i з N, таку, що $N_i = P$.
6. Навчаємо H_i, використовуючи P та N_i. H_i у нашому випадку був AdaBoost ансамблем зі «слабкими» класифікаторами $h_{i,j}$ та відповідними вагами $\alpha_{i,j}$, тобто

$$H_i(x) = sgn\left(\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x) - \theta_i\right).$$

7. Повторюємо доки $i = T$.
8. Вихідні дані: ансамбль

$$H(x) = sgn\left(\sum_{i=1}^{T}\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x) - \sum_{i=1}^{T}\theta_i\right).$$

Автором роботи на тренувальній вибірці було згенеровано сім підмножин мажорантного класу та проведено навчання ансамбельного класифікатора на основі дерева рішень на кожній із підмножин у парі з множиною мінорантного класу.

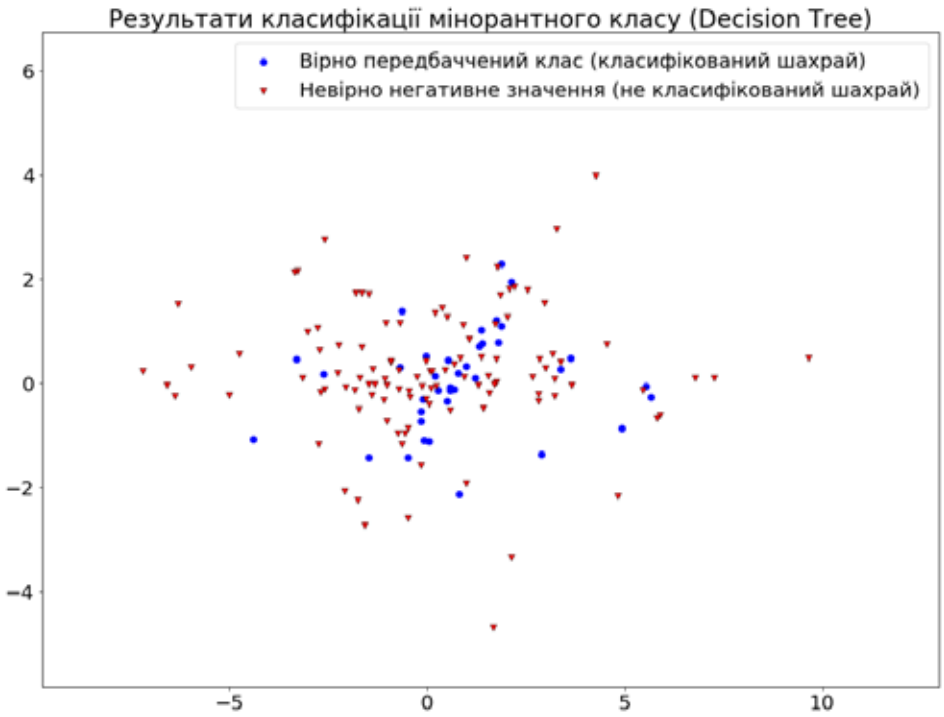

Рис. 5. Візуалізація роботи дерева рішень на мінорантному класі тестової вибірки з даних страхової компанії у двовимірному просторі

Джерело: побудовано автором

Таблиця 1

Оцінка роботи дерева рішень та AdaBoost

Алгоритм	Прицезійність	Чутливість	F1	Точність
Decision Tree	0,27	0,26	0,26	0,93
Ada Boost	0,25	0,23	0,24	0,93

Джерело: побудовано автором

На тестовій вибірці алгоритм показав такі результати (табл. 2).

З таблиці та рис. 6 можна побачити, що чутливість роботи класифікатора щодо мінорантного класу значно краща відносно дерева рішень, проте це сильно відбилося на точності роботи класифікатора загалом.

Для порівняння було використано алгоритм BalanceCascade, який також базується на побудові моделей класифікації на попередньо розбитій вибірці мажорантного класу та усьому мінорантному класі. Опісля цієї процедури побудовані моделі також об'єднуються в ансамбль моделей. Відмінність цього методу від EasyEnsemble полягає у тому, що підмножини тут вибираються не повністю випадковим чином, у результаті чого одна і та ж точка в багатовимірному просторі може класифікуватися кілька раз, а використовуються методи виключення, коли після проведення класифікації на кожній із підмножин мажорантного класу правильно класифіковані значення виключаються з наступної вибірки N_i. Ансамблем, що використовувався для здійснення класифікації, тут, як і в минулих прикладах, використовувався AdaBoost, що базувався на деревах рішень.

Загалом алгоритм BalanceCascade можна описати так [28]:

1. Вхідні дані: множина мінорантного класу прикладів P, множина мажорантного класу прикладів N, $|P| < |N|$, і T, кількість підмножин, що повинні бути обрані з N.
2. Присвоюємо $i \leftarrow 0$.
3. Початок циклу:
4. $i \leftarrow i + 1, f \leftarrow \sqrt[T-1]{\frac{|P|}{|N|}}$.
5. Випадково вибираємо підмножину N_i з N, $|N_i| = |P|$.
6. Навчити H_i, використовуючи P та N_i. H_i, у нашому випадку був AdaBoost ансамблем зі слабкими класифікаторами $h_{i,j}$ та відповідними вагами $\alpha_{i,j}$, тобто

$$H_i(x) = sgn\left(\sum_{j=1}^{s_i} \alpha_{i,j} h_{i,j}(x) - \theta_i\right).$$

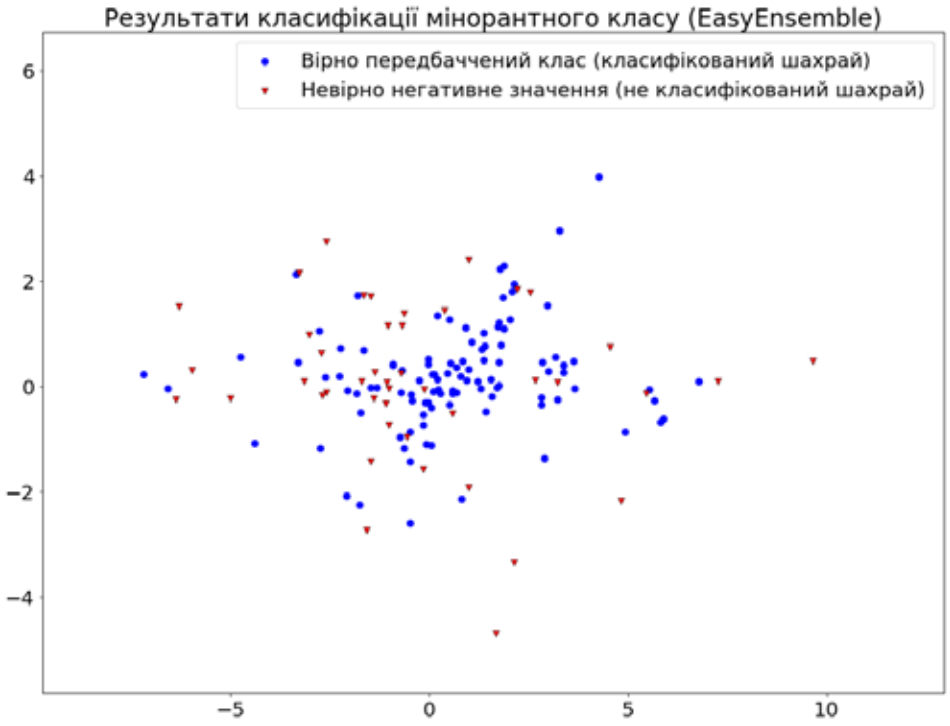

Рис. 6. Візуалізація роботи дерева рішень на мінорантному класі тестової вибірки після застосування алгоритму EasyEnsemble

Джерело: побудовано автором

Таблиця 2

Оцінка роботи EasyEnsemble

Алгоритм	Прицезійність	Чутливість	F1	Точність
EasyEnsemble	0,1	0,62	0,17	0,71

Джерело: побудовано автором

7. Наблизити θ_i так, щоб невірно позитивне значення (FP) ансамблю H_i було рівним f.

8. Видалити з N усі приклади, які правильно класифікував H_i.

9. Повторюємо, доки $i = T$.

10. Вихідні дані: ансамбль

$$H(x) = sgn\left(\sum_{i=1}^{T}\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x) - \sum_{i=1}^{T}\theta_i\right).$$

Після застосування алгоритму та використання класифікатора було отримано такі результати (табл. 3).

Можна побачити, що алгоритм чутливіший за EasyEnsemble у разі класифікації мінорантного класу та показує кращі результати щодо загальної точності роботи. Візуально результати класифікації можна побачити на рис. 7.

Як можна побачити, використання алгоритмів для роботи з розрідженими даними передбачає певний компроміс чутливості класифікатора до мінорантного класу та його чутливості до мажорантного класу, що відображається на точності роботи побудованої моделі. Відповідь на питання, який тип моделі повинен обрати економічний агент (бізнес), повинна давати консалтингова компанія залежно від розробленої стратегії роботи на ринку та прийнятої системи управління ризиками. В умовах асиметричної інформації, якщо компанія-клієнт бажає максимізувати прибуток, доцільно врахувати вартість помилок першого та другого роду, наприклад скільки компанія потенційно втратить грошей, якщо помилково відкине хорошого клієнта, коли класифікатор спрацював помилково (похибка першого роду), або які збитки може нанести компанії шахрай, якщо класифікатор пропустить його (похибка другого роду). Проте якщо компанія-клієнт бажає залучити якнайбільше покупців, то консалтерам варто подумати про застосування стратегії масового залучення користувачів зі зменшенням кількості відмов хорошим клієнтам, що може потягнути за собою збільшення кількості шахрайств, утри-

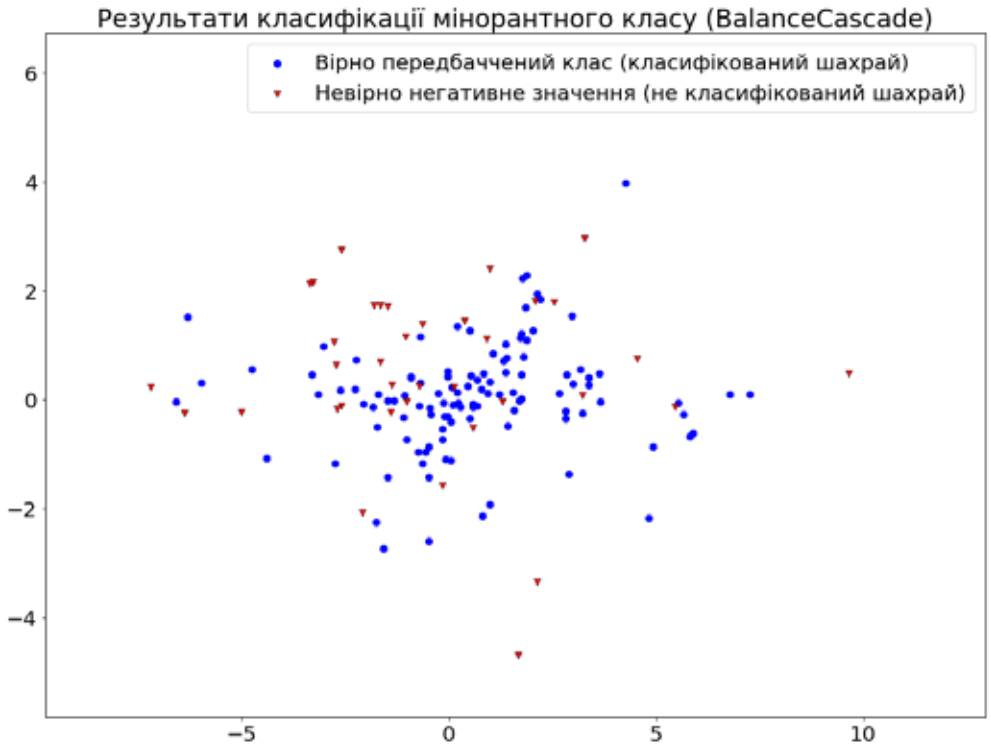

Рис. 7. Візуалізація роботи дерева рішень на мінорантному класі тестової вибірки після застосування алгоритму BalanceCascade

Джерело: побудовано автором

Таблиця 3

Оцінка роботи BalanceCascade

Алгоритм	Прицезійність	Чутливість	F1	Точність
BalanceCascade	0,14	0,75	0,24	0,76

Джерело: побудовано автором

манням кількості яких на заданому рівні повинна займатися побудована система управління ризиками. Також цього можна досягнути зміною цільової метрики оптимізації роботи моделі або використанням додаткових алгоритмів по роботі з даними. У цій ситуації автор статті пропонує використовувати поєднання техніки SMOTE (англ. Synthetic Minority Oversampling Technique) для збільшення кількості значень мінорантного класу у поєднанні з ENN (англ. Edited Nearest Neighbor) для зменшення кількості значень мажорантного класу з подальшим застосуванням класифікатора на новому отриманому наборі даних.

Метод SMOTE створює додаткові дані мінорантного класу, генеруючи «синтетичні» точки даних у багатовимірному просторі [29]. Алгоритм створення додаткових даних базується на методі k-найближчих сусідів та може бути записаний таким псевдокодом:

1. Вхідні дані: кількість точок мінорантного класу T, відсоток, на який повинна бути збільшена множина мінорантного класу, – N %, кількість найближчих сусідів – k.
2. Якщо $N < 100$,
3. тоді випадково вибираємо значення з мінорантного класу T.
4. $T = (N/100) * T$.
5. $N = 100$.
6. Кінець циклу.
7. $N = (\text{int}) N/100$.
8. k = Кількість найближчих сусідів.
9. a = Кількість атрибутів.
10. S: масив оригінальних даних мінорантного класу.
11. $n \leftarrow 0$: зберігає кількість згенерованих синтетичних даних.
12. Y: масив синтетичних значень.
13. Для $i \leftarrow 1$ до T.
14. Обчислити k-найближчих сусідів для i та зберегти індекси в масив K.
15. Заповнити масив (N, i, K).
16. Кінець циклу.

Заповнення масиву (N, i, K) (Функція для генерації синтетичних даних):

17. Поки $N \uparrow 0$.
18. Вибрати випадкове число між 1 і k, назвати його nn. Цей крок вибирає одного з k найближчих сусідів i.
19. Для $attr \leftarrow 1$ до a.
20. Обчислити:

$$dif = S[K[nn]][attr] - S[i][attr].$$

21. Обчислити:

gap = випадкове число між 0 та 1.

22. $Y[n][attr] = S[i][attr] + gap * dif$.
23. Кінець циклу.
24. n=n+1.
25. $N = N - 1$.
26. Кінець циклу.
27. Повернути масив (N, i, K).

Вихідні дані: $(N/100) * T$ синтетичних значень мінорантного класу

Метод ENN працює за схожим алгоритмом, використовуючи метод k-найближчих сусідів, але вже для видалення значень мажорантного класу. Псевдокод методу може бути записаний так [30]:

1. Вхідні дані: X – масив оригінальних даних мажорантного класу, N – кількість точок даних у X, кількість найближчих сусідів – k.
2. Для $i \leftarrow 1$ до N.
3. Знайти k-найближчих сусідів для X_i серед $\{X_1, X_2, ... X_{i-1}, X_{i+1}, ..., X_N\}$.
4. Знайти клас θ, який відповідає найбільшому числу точок серед k-найближчих сусідів.
5. Редагувати множину $\{(X_i, \theta_i)\}$ шляхом видалення (X_i, θ_i), якщо θ_i не рівне найбільшому числу з k-найближчих сусідів, як означено вище.

Під час застосування запропонованого методу поєднання алгоритмів вхідним параметром було взято кількість найближчих сусідів – 5 (для обох алгоритмів), а також встановлено бажане значення показника присутності мінорантного класу – 0,35 (на початку було 0,05). У результаті використання алгоритму ENN частка мінорантного класу зросла до 0,44. Щоб продемонструвати роботу алгоритмів, результати їх роботи були візуалізовані на усіх даних страхової компанії (рис. 8).

Після застосування алгоритмів на тренувальних даних, на яких навчалася модель класифікації, кількість спостережень змінилася з 351 для шахраїв та 6 649 для добросовісних клієнтів до 4 395 та 5 162 відповідно. Під час валідації моделі на тестових даних було отримано такі результати (табл. 4).

Графічне зображення застосування запропонованого автором поєднання алгоритмів можна побачити на рис. 9.

Можна помітити, що порівняно з попередніми алгоритмами поєднання запропонованого методу робить класифікатор менш чутливим до даних мінорантного класу, проте збільшує прицезійність та точність класифікатора загалом.

Висновки з цього дослідження. Консалтингова компанія може достатньо ефективно застосовувати економіко-математичне моделювання для врахування асиметричності інформації між економічними агентами, навіть

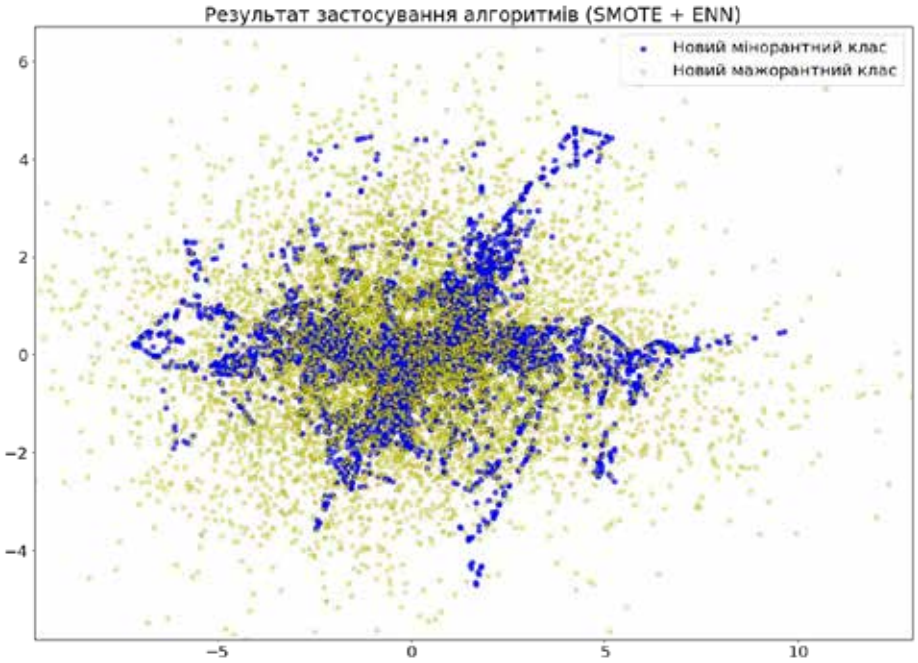

Рис. 8. Візуалізація роботи алгоритмів SMOTE та ENN на всіх даних страхової компанії

Джерело: побудовано автором

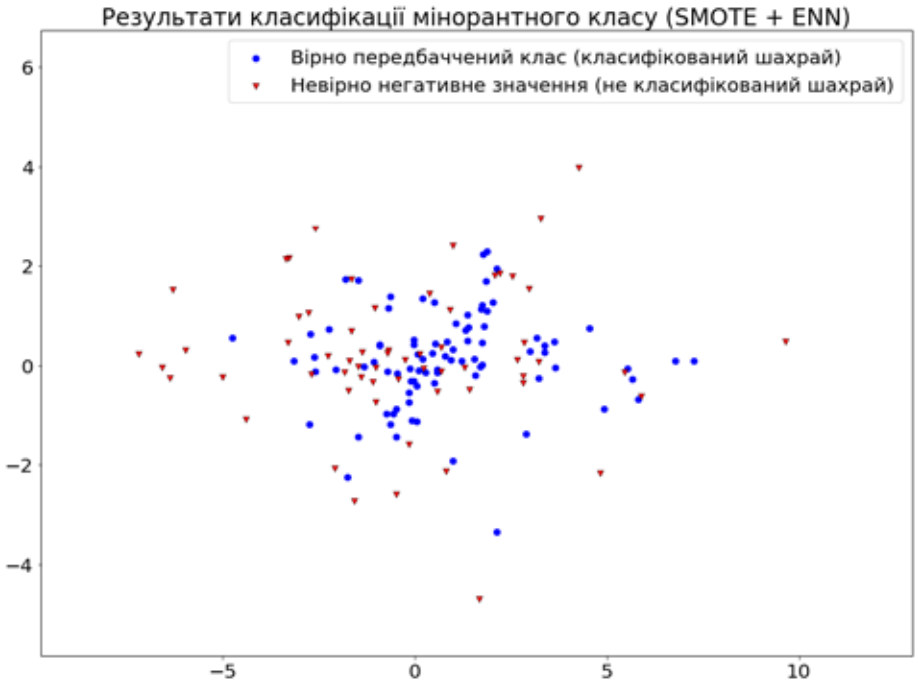

Рис. 9. Візуалізація роботи дерева рішень на мінорантному класі тестової вибірки після застосування алгоритмів SMOTE та ENN

Джерело: побудовано автором

Таблиця 4

Оцінка роботи BalanceCascade

Алгоритм	Прицезійність	Чутливість	F1	Точність
SMOTE+ENN	0,18	0,58	0,28	0,85

Джерело: побудовано автором

якщо дані в її розпорядженні є незбалансованими. Використання математичних моделей класифікаторів дає змогу передбачати, як економічні суб'єкти поводитимуться в майбутньому, а застосування алгоритмів у роботі з розрідженими даними дає змогу оптимізувати процес прийняття рішень та керувати ризиками, беручи до уваги розроблену стратегію роботи компанії-клієнта на ринку.

Автор дослідження пропонує використовувати звичайний класифікатор під час роботи з незбалансованими даними лише тоді, коли ціна прийняття невірного рішення стосовно економічного суб'єкта з міноритарного класу є низькою. Однак якщо ціна помилки другого роду є високою, потрібно застосовувати методи роботи з незбалансованими даними. Залежно від того, яка стратегія роботи з клієнтами є прийнятою у компанії – максимальне залучення клієнтів або ж зменшення втрат, – потрібно обирати відповідні цільові метрики та методики роботи з даними.

Розглянуті методи EasyEnsemble та BalanceCascade дають змогу максимізувати чутливість роботи класифікатора на мінорантному класі, а запропоноване автором поєднання алгоритмів SMOTE та ENN дасть змогу консалтинговій компанії забезпечити компроміс між точністю математичної моделі та її чутливістю до мінорантного класу.

ЛІТЕРАТУРА:

1. Мурыгин К. Обнаружение автомобильных номеров на основе смешанного каскада классификаторов / К. Мурыгин // Штучний інтелект. – 2010. – № 2. – С. 147–152.

2. Мурыгин К. Особенности реализации алгоритма AdaBoost для обнаружения объектов на изображениях / К. Мурыгин // Штучний інтелект. – 2009. – № 3. – С. 573–581.

3. Murygin K. Detection and Recognition of Objects on Images Based on MKV-Classifiers // Штучний інтелект. – 2013. – № 1. – P. 209–217.

4. Настенко Е. Синтез логистической регрессии на принципах самоорганизации моделей / Е. Настенко // Кибернетика и вычисл. техника. – 2015. – Вып. 182. – С. 85–93.

5. Кардаш А. Задача розпізнавання людських обличь методами штучного інтелекту / А. Кардаш // Інформаційні технології та комп'ютерна інженерія. – 2013. – № 1. – С. 84–87.

6. Волошин Н. Моделирование и распознавание информативных участков в автоматизированных системах иридодиагностики / Н. Волошин // Восточноевропейский журнал передовых технологий. – 2011. – № 2/2(50). – С. 65–69.

7. Жук М. Аналіз платоспроможності позичальника – представника домогосподарства за допомогою економетричних моделей бінарного вибору / М. Жук // Регіональна економіка. – 2013. – № 3. – С. 114–122.

8. Пістунов І. Визначення ймовірності неповернення кредиту особами, що не мають кредитної історії / І. Пістунов // Економічний вісник. – 2014. – № 2. – С. 101–108.

9. Мойсеєнкова Д. Модель прийняття комплексного скорингового рішення / Д. Мойсеєнкова // Економічний вісник НТУУ «КПІ». – 2015. – № 12. – С. 495–502.

10. Литвин А. Побудова моделей прогнозування банкрутства страхових компаній України в післякризовий період / А. Литвин // Економічний аналіз. – 2013. – Т. 14. – № 1. – С. 282–300.

11. Бакун С. Методика побудови скорингових карт із використанням платформи SAS / С. Бакун // Наукові вісті НТУУ «КПІ». – 2016. – № 2. – С. 23–32.

12. Гюнтер И. Скоринг как основа минимизации кредитного риска / И. Гюнтер // Науковий вісник Полтавського університету економіки і торгівлі. – 2011. – № 6(51). – Ч. 2. – С. 271–273.

13. Минц А. Современные методы анализа данных в финансово-кредитной сфере / А. Минц // Вісник Приазовського державного технічного університету. Економічні науки. – 2011. – № 2(22). – С. 149–156.

14. Матвійчук А. Використання logit- та probit-регресій для оцінки кредитоспроможності позичальника / А. Матвійчук // Вісник Національного банку України. – 2015. – Травень. – С. 37–41.

15. Солошенко О. Розробка методу k-plus-найближчих сусідів для задач машинного навчання кредитного скорингу / О. Солошенко // Восточноевропейский журнал передовых технологий. – 2015. – № 3/9(75). – С. 29–38.

16. Кожухівська О. Розроблення системи підтримки прийняття рішень для оцінювання фінансових ризиків / О. Кожухівська // Вісник Черкаського державного технологічного університету. Технічні науки. – 2014. – № 1. – С. 51–56.

17. Кузнєцова Н. Інтегрований підхід до оцінювання кредитних ризиків / Н. Кузнєцова // Труды Одесского политехнического университета. – 2010. – № 1–2. – С. 187–192.

18. Liu Ning Ensemble classification algorithm based improved SMOTE for imbalanced data // Науковий вісник НГУ. – 2016. – № 2. – P. 122–127.

19. Nitesh V. Chawla, Special Issue on Learning from Imbalanced Data Sets. – ACM SIGKDD Explorations Newsletter. Volume 6, Issue 1. – 2004. – P. 1–6.

20. Osama Othman Preceding Rule Induction with Instance Reduction Methods. – Machine Learning and Data Mining in Pattern Recognition. MLDM 2013. Lecture Notes in Computer Science, v. 7988. Springer. – 2013. – P. 209–218.

21. Donghai Guan Nearest neighbor editing aided by unlabeled data. – Information Sciences 179. – 2009. – P. 2273–2282.

22. Schapire R. The Strength of Weak Learnability. – Machine Learning. – 1990. – № 5. – P. 197–227.

23. Паклин Н. Построение классификаторов на несбалансированных выборках на примере кредитного скоринга / Н. Паклин // Искусственный интеллект». – 2010. – № 3. – С. 528–534.

24. Vinciotti V. Scorecard construction with unbalanced class sizes // Journal of The Iranian Statistical Society, Volume 2, Issue 2. – 2003. – P. 189–205.

25. Leland H. Informational Asymmetries, Financial Structure, and Financial Intermediation. – The Journal of Finance, v. 32, № 2. – 1977. – P. 371–387.

26. Haibo He Learning from Imbalanced Data. – Transactions on Knowledge and Data Engineering. – 2009. – vol. 21. – № 9. – P. 1263–1284.

27. Damodaran R. Predicting Rare Events Using Specialized Sampling Techniques in SAS. – SAS Institute, Paper 11140. – 2016. – P. 1–7.

28. Xu-Ying Liu Exploratory Under-Sampling for Class-Imbalance Learning. – IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) vol. 39, Issue 2. – 2009. – P. 539–550.

29. Nitesh V. Chawla, SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. – 2002. – № 16. – P. 321–357.

30. Wilson D. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. – IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. SMC-2, № 3. – 1972. – P. 408–421.