DOI 10.15276/imms.v11.no1-2.63
УДК 0004.942

Informatics and Mathematical Methods in Simulation
Vol. 11 (2021), No. 1-2, pp.63-73

ДОСЛІДЖЕННЯ МЕТОДУ СТАБІЛІЗАЦІЇ ДИСПЕРСІЇ ВИПАДКОВИХ ВЕЛИЧИН НА ОСНОВІ СТЕПЕНЕВИХ ПЕРЕТВОРЕНЬ ДЛЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Фомін О.О.[1], Масрі М.М.[2], Туманов О.А.[2], Сєрбова О.В.[2], Ткаченко В.В.[2]

[1]Одеський національний політехнічний університет,
пр. Шевченка. 1, Одеса, 65044; Україна; fomin@opu.ua
[2]Приватний заклад вищої освіти «Одеський технологічний університет «Шаг»»,
вул. Садова, 3, Одеса, 65023; Україна; dr.mohanadmasri@gmail.com

В роботі розвязуєтьсязавдача побудови моделей машинного навчання. Метою роботи є підвищення точності моделей машинного навчання для випадку розподілу даних навчальної вибірки відмінного від нормального. Поставлена мета досягається шляхом застосування статечних перетворень до ознак навчальної вибірки для стабілізації дисперсії випадкових величин. Найбільш суттєві результати: досліджено ефективність методу статистичної класифікації на основі баєсова підходу для побудови моделей у вигляді поліномів другого порядку, що дозволяє підвищити точнысть моделювання у випадку розподілу даних навчальної вибірки відмінного від нормального, проведені експериментальні дослідження і сформульовані рекомендації щодо ефективності застосування статечних перетворень даних навчальної вибірки при різних розподілах даних навчальної вибірки для наступних законів розподілу випадкових величин: рівномірний, біноміальний, експоненціальне. Значимість отриманих результатів: застосування сформульованих рекомендацій дозволяє забезпечити підвищення точності моделей машинного навчання в умовах розподілу даних навчальної вибірки відмінного від нормального. Запропоновані рекомендації апробовані на даних тестового завдання. Завдання демонструє підвищення точності моделей машинного навчання для більшості розглянутих законів розподілу випадкових величин.

Ключові слова: статистичні характеристики,випадкова величина, степеневі перетворення, моделі машинного навчання.

Вступ

Великі обсяги даних та значні успіхи в області DataScience і BigData забезпечують нові можливості при рішенні прикладних задач, пов’язані з підвищеннямточностіпобудови моделей машинного навчання [1, 2].Ефективне функціонування сучасних інформаційних систем в таких умовах забезпечується зростаючими вимогами до процесів обробки вихідної інформації, в тому числі, до методів статистичної обробки даних.

Одними з поширених в теоретичних розробках є задачі, пов'язані з використанням теорії перевірки статистичних гіпотез. Незважаючи на те, що в загальному випадку теорія перевірки статистичних гіпотез розроблена без накладення обмежень на щільності розподілу розглянутих випадкових величин, на практиці, у зв'язку зі спостереженням нормального закону розподілу випадкових величин в багатьох природних процесах, зручністю використання математичного апарату і наявністю ефективних інструментальних засобів комп'ютерного моделювання найбільш широко використовуються гаусові моделі випадкових величин, що складають навчальну вибірку. Це не завжди адекватно відображає реальні природні процеси, що, в загальному випадку, призводить до зниження точності побудови моделей машинного навчання[3].

Проблема подальшого розвитку даного напрямку полягає в тому, що обробка негауссовских процесів не в повній мірі досліджена і потребує подальшого своєму вивченні. Тому, подальший розвиток моделей машинного навчання та методів їх побудови, в разі розподілу даних навчальної вибірки відмінного від нормального, є актуальним напрямком досліджень.

Мета роботи

Метою роботиє підвищення точності моделей машинного навчання у випадку розподілу даних навчальної вибірки відмінного від нормального шляхом застосування степеневих перетворень до ознак навчальної вибірки для стабілізації дисперсії випадкових величин.

Аналіз останніх досліджень і публікацій

Развитие теории и методов машинного обучения при работе с негауссовскими процессами характеризуется несколькими научными школами и направлениями, в частности применением марковских процессов и использованием полигауссовских моделей [4, 5].

Были разработаны общие подходы статистических методов обработки данных, нашедшие отражение в работах [6–9]. Ці наукові напрямки отримали свій розвиток, що відображено в безлічі публікацій і послужило основою для створення методів побудови моделей об'єктів у випадку розподілу даних навчальної вибірки відмінного від нормального і самих технічних систем їх реалізації. Незважаючи на те, що зазначені напрями принципово відрізняються один від одного, їх об'єднує те, що в їх основі лежить імовірнісний підхід до опису випадкових процесів, що грунтується на використанні щільності ймовірності розподілу для опису випадкових величин. Однак такі підходи характеризуються рядом проблем, пов'язаних з необхідністю оперування самими плотностями ймовірностей розподілу, що не завжди є можливим, а також складністю практичної реалізації.

Іншим напрямком процесу побудови моделей машинного навчання у випадку розподілу даних навчальної вибірки відмінного від нормального є вибір ознак навчальної вибірки і визначення способу їх виділення (вимірювання) [10–13]. Згідно цьому підходу після накопичення статистичних даних необхідно розглянути можливість виключення частини ознак, задовольняючи при цьому існуючим вимогам до точності моделювання.

Рішення задачі вибору сукупності ознак може бути досягнуто двома шляхами.

Перший з них заснований на зважуванні тим або іншим способом різних систем ознак з метою оцінки їх інформативності при моделюванні, наприклад, за результатами рішення задачі класифікації об'єктів екзаменаційної вибірки за допомогою побудованого одним з алгоритмів навчання вирішального правила [14, 15]. Набори ознак, для яких критерій оцінки точності малий, відкидаються, і в якості системи ознак обирається набір, для якого додавання будь-якої нової ознаки не збільшує або збільшує небагато його інформативність. Якщо параметри статистично незалежні, то систему ознак можна сформувати шляхом оцінки інформативності кожного параметра і відкидання деякої кількості найменш інформативних з них. При такій процедурі формування простору ознак значення самих ознак не змінюються, а тільки зменшується їх кількість.

Для вирішення різноманітних завдань математичної статистики отримав свій подальший розвиток інший підхід, де в якості апріорного опису випадкових величин використовується не щільність ймовірності розподілу випадкових величин, а деякий перетворення вихідних даних з метою отримання щільності розподілу випадкової величини, близької до нормальної, що дозволяє істотно спростити рішення конкретних практичних завдань при досягненні високої ефективності обробки.

Цей шлях вирішення задачіскорочення інформації при моделюванніоб'єкту полягає в знаходженні оптимального, в деякому сенсі, перетворення *A* вихідного простору векторів вимірювань **x** розмірністю *n* в простір зображень меншої розмірності *Y* (розмірністю *m*<*n*) [16, 17]. В якості такого перетворення може використовуватися вейвлет-перетворення і розкладання Карунена-Лоєва [18]. Однак на відміну від описаного вище методу вибору діагностичних ознак цей метод не передбачає скорочення кількості вимірювань. В цьому випадку нові ознаки виявляються відірваними від конкретного фізичного змісту і мають тільки абстрактне інформаційне значення.

Ця актуальна та перспективна задача може бути вирішена за рахунок побудови моделей істотно меншої розмірності, що забезпечує зниження вимог до вимірювань і зберігання даних, зменшення обчислювального навантаження при комп'ютерній реалізації моделей машинного навчання, підвищення оперативності моделювання при збереженні високої його точності.

Незважаючи на великий інтерес до цього напрямку і достатня кількість публікацій, до сьогоднішнього дня не в повній мірі не вирішеною залишається завдання проведення систематизованих досліджень ефективності подібних перетворень на випадки найбільш поширених розподілів випадкових величин. Вирішенню цього завдання присвячена дана робота.

Дана робота присвячена вирішенню цього завдання, а саме: дослідженню методів стабілізації дісперсії Випадкове величин на основе степеневих Перетворення для моделей машинного навчання.

Основна частина

Ефективність застосування методів машинного навчання для розв'язання задач регресії, класифікації, кластерізації тощо у великій мірі залежить від інформативності ознак навчальної вибірки. Якщо обрані ознаки досить повно характеризують внутрішню структуру объекта контроля, то ідентичні за структурою об'єкти з'являться в просторі цих ознак у вигляді щільної множини точок. Об'єктам з особливостями структури – дефектним – будуть відповідати точки, що відхиляються від цієї щільної безлічі.

Вибір сукупності ознак навчальної вибірки робить вирішальний вплив на точність моделі машинного навчання. Визначення точності моделімашинного навчання передбачає використання конкретної моделі, оскільки тільки в її рамках має сенс точністьмоделювання.

Алгоритм машинного навчання.

Як показав огляд, для задач моделюванняоб'єктів навколишнього світу більшість традиційних методів побудови моделей машинного навчання, що базуються на припущенні про нормальний закон розподілення даних в навчальній вибірці, є мало ефективні внаслідок наявності у вибірці негаусівських розподілів випадкових величин[19, 20].

В даній роботі використано метод статистичної класифікації на основі баєсова підходу (максимальної правдоподібності) для побудови діагностичних моделей у вигляді поліномів другого порядку.

Діагностичні моделі $d(\mathbf{x})$на основі методу максимальної правдоподібностібудуються за виразом (1). Запис діагностичні моделі у векторній формі має вигляд:

$$d(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{S}_1^{-1}(\mathbf{x}-\mathbf{m}_1)+\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T\mathbf{S}_2^{-1}(\mathbf{x}-\mathbf{m}_2)+\frac{1}{2}\ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}(1),$$

де $\mathbf{S}_i$ – коваріаційна матриця розмірності nхn вимірюваних параметрів стану для i-го класу; $|\mathbf{S}_i|$ – визначник матриці $\mathbf{S}_i$; $\mathbf{S}_i^{-1}$– матриця, зворотна матриці $\mathbf{S}_i$.

У випадку незалежних p ознак$\mathbf{S}_i=\sigma_i E_0$, де E_0 – одинична матриця розміру pхp, σ_i – дисперсія ознак у i-му класі навчальної вибірки,вираз (2) для діагностичних моделей спрощується:

$$d(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T(\mathbf{x}-\mathbf{m}_1)/\sigma_1^2+\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T(\mathbf{x}-\mathbf{m}_2)/\sigma_2^2+\frac{1}{2}\ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|} \quad (2).$$

Така модель дає оптимальний за кількістю помилок результат у випадку нормального розподілу даних навчальної вибірки тадобрепрацює при дотриманні вимог симетричного і унімодального розподілу даних, що не завждизабезпечується реальним світом.Для усунення зазначеного недоліку можна використовувати два шляхи: адаптація метода побудови моделі на випадок негаусівського розподілу ознак навчальної вибірки або адаптиациянавчальної вибірки під требования алгоритма.

Критерій точності моделювання.

В якості критерія точності моделювання об’єкту використовуються експериментальні оцінки: помилки класифікації та ймовірності правильного розпізнавання.

Помилки класифікації визначаються за виразом:

$$\Delta_{ij} = \frac{L_{ij}}{N_i}, \quad (3)$$

де L_{ij} – кількість об’єктів i–го класу в екзаменаційній вибірці, помилкововіднесених доіншого класу $j(j\neq i)$; N_i – кількість элементов i–го класу в екзаменаційній вибірціi=1, 2,…, M; M – кількість класів станів ОД.

Імовірність правильного распізнавання P, средняя по всім класам:

$$P = 1 - \frac{\sum_{i=1}^{M}\sum_{j\neq i} L_{ij}}{\sum_{i=1}^{M} N_i} \quad (4)$$

На практиці частіше використовують перетворення випадкової величини, які наближають її функії щільності розподілу до нормального закону (процес нормалізація випадкової величини). Використання такого підходу може забезпечити високу точністьмашинного моделювання, навіть, при використанні математичного апарату таинструментальных средств комп’ютерного моделювання, що розраховано на випадки гаусівських моделейвипадкових величин.Це важливо, так як алгоритми машинного навчанняпрацюють значно кращез нормалізованими даними.

Постановка експерименту.

Для підвищення якості моделювання в задачі класифікації в разі розподілу ознак в навчальній вибірці відмінного від нормального пропонується використання перетворень, спрямованих на усунення зв'язку між дисперсією і математичним очікуванням. Таким чином, дисперсія стає постійною по відношенню до середнього. Таку стабілізацію пропонується здійснювати шляхом застосування степеневих перетворень випадкової величини$\mathbf{x}$=(x_1, …, x_n), зокрема логарифмування зі стабілізацією:

$$x_j^0 = log_a(x_j + b) \quad (5)$$

де a – основа логарифма; b – стабілізуючий коефіцієнт.

Для дослідженняефективности такого перетворенняпри використанні різноманітних розподілів ознакнавчальної вибірки поставленоі виконанотестовий експеримент.

Не зважаючи на те, що існує багато різноманітних законів розподілу випадкових величин, на практиці частіше за інші використовуються наступні випадки: рівномірний, біноміальний та экспоненціальний закон (Таблиця 1).

Таблиця 1.

Розповсюдженізакони розподілу випадкових величин

№.	**Закон розподілу**	**Формальний опис**
1	Рівномірний	$f(x) = \begin{cases} 1/(b-a), a \le x \le b \\ 0, x < a, x > b \end{cases}$
2	Біноміальний	$f(x) = C_n^k p^k (1-p)^{n-k}, k = 0, \dots, n$
3	Експоненціальний	$f(x) = \begin{cases} \lambda e^{-\lambda x}, x \ge 0 \\ 0, x < 0 \end{cases}$

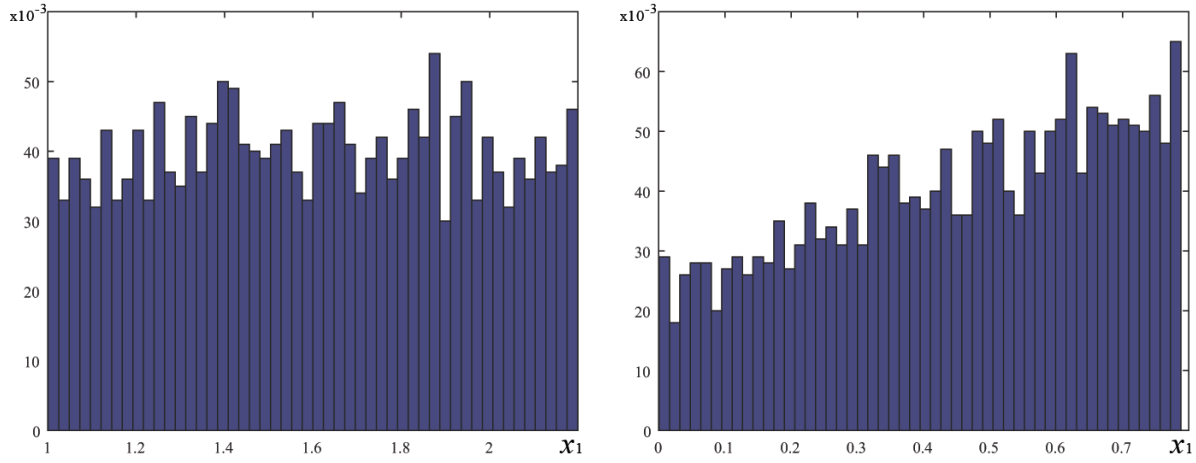
Навчальна вибірка для експерименту являє собою матрицю розмірністюmx2 (m – кількість вимірювань).Кожнастрока матриці є вектором **x**=(x_1, x_2).має при різних законах розподілувипадкової величини (рівномірному, біноміальному, експоненціальному) та розділена на 2 класи.

Для кожного випадку вирішуються задачі побудови моделі (вирішального правила) об’єктів методом статистичної класифікації на основі баєсова підходу:

- на основі початкової навчальної вибірки;
- на основі навчальної вибірки, отриманої шляхом степеневого (5) перетворення її ознак;

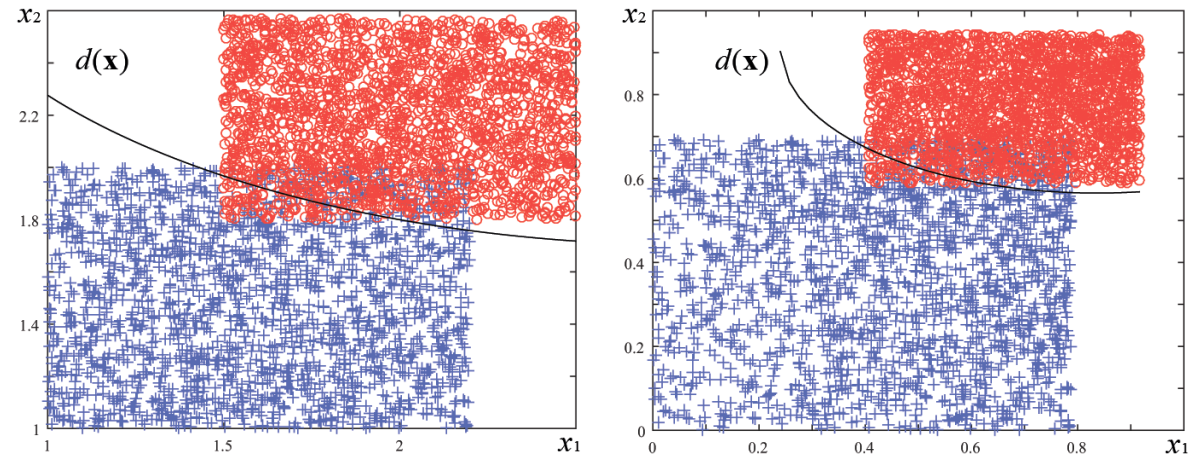
Випадок рівномірногорозподілу випадкової величини

На рис. 1та 2 представлено результатирозв’язання задачі класифікаціїдля рівномірного законурозподілу даних навчальної вибірки. Експериментально збудована функціярозподілуймовірностей для рівномірного законурозподілу ознаки x_1наведено на рис. 1, а.Скоригована за допомогою степеневих перетворень навчальної вибірки функція розподілу ймовірностей ознаки x_1наведено на рис. 1, б.На рис. 2 наведено вирішальні правилакласифікації для рівномірного закону розподілу ознак (x_1, x_2): для початкової навчальної вибірки (риа. 2, а) та скоригованої за допомогою степеневих перетворень навчальної вибірки (риа. 2, б).



а б

Рис. 1. Функції розподілу ймовірностейознаки x_1: а –рівномірний закон розподілу, б – скоригований за допомогою степеневих перетворень навчальної вибірки закон розподілу

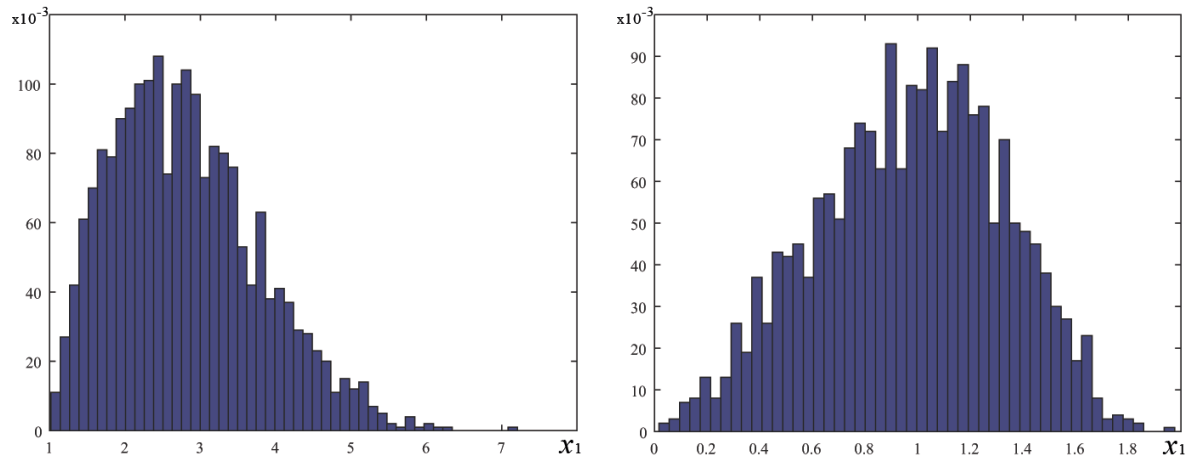


а б

Рис. 2. Вирішальні правила $d(x)$ класифікації в просторі ознак (x_1, x_2): а – початкова навчальна вибірка з рівномірним законом розподілу, б – скоригований за допомогою степеневих перетворень навчальна вибірка

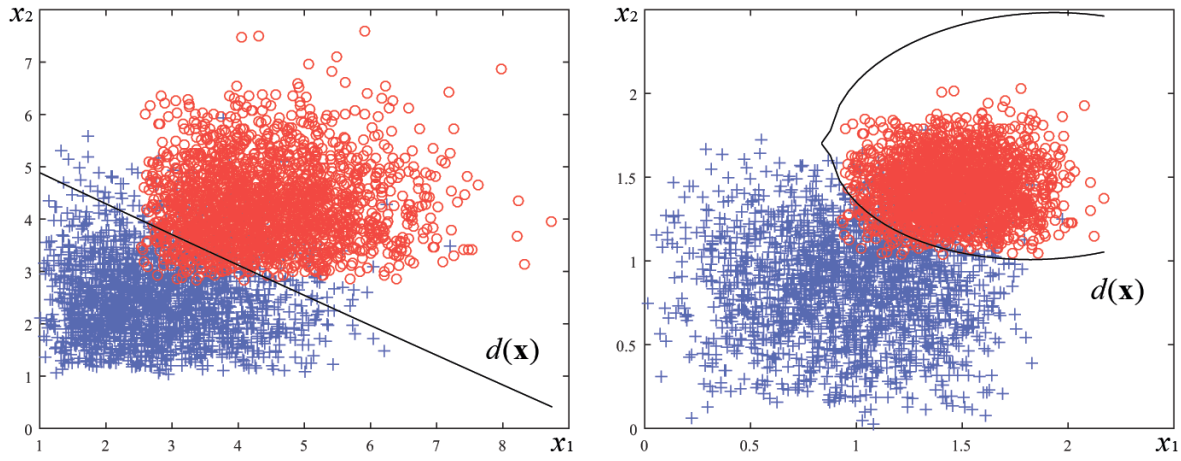
Випадок біноміальногорозподілу випадкової величини

На рис. 3та 4 представлено результати розв'язання задачі класифікації для рівномірного закону розподілу даних навчальної вибірки. Експериментально збудована функція розподілу ймовірностей для рівномірного закону розподілу ознаки x_1наведено на рис. 3, а. Скоригована за допомогою степеневих перетворень навчальної вибірки функція розподілу ймовірностей ознаки x_1наведено на рис. 3, б. На рис. 4 наведено вирішальні правила класифікації для рівномірного закону розподілу ознак (x_1, x_2): для початкової навчальної вибірки (риа. 4, а) та скоригованої за допомогою степеневих перетворень навчальної вибірки (риа. 4, б).



а б

Рис. 3. Функції розподілу ймовірностейознаки x_1: а – біноміальний закон розподілу, б – скоригований за допомогою степеневих перетворень навчальної вибірки закон розподілу

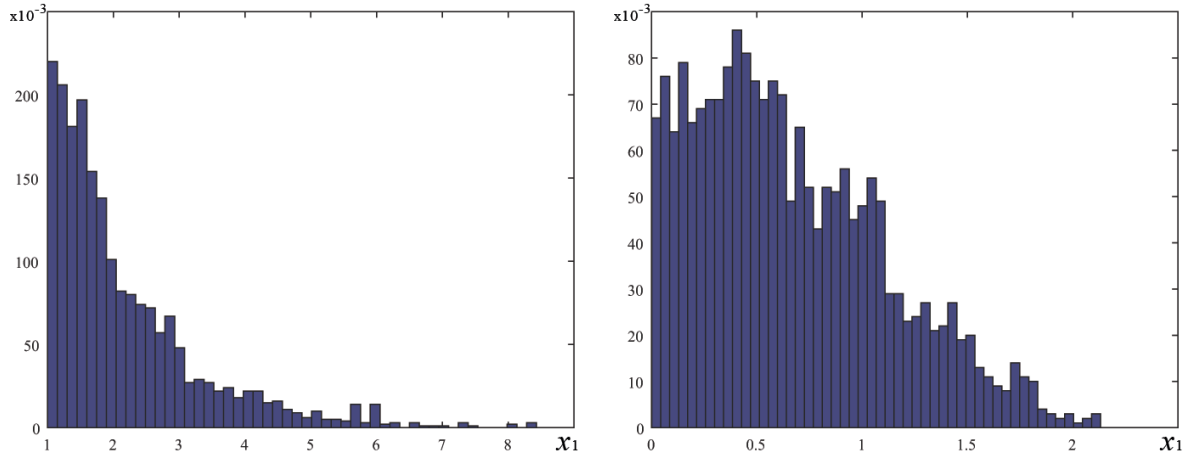


а б

Рис. 4. Вирішальні правила $d(x)$ класифікації в просторі ознак (x_1, x_2): а – початкова навчальна вибірка з биномыальним законом розподілу, б – скоригований за допомогою степеневих перетворень навчальна вибірка

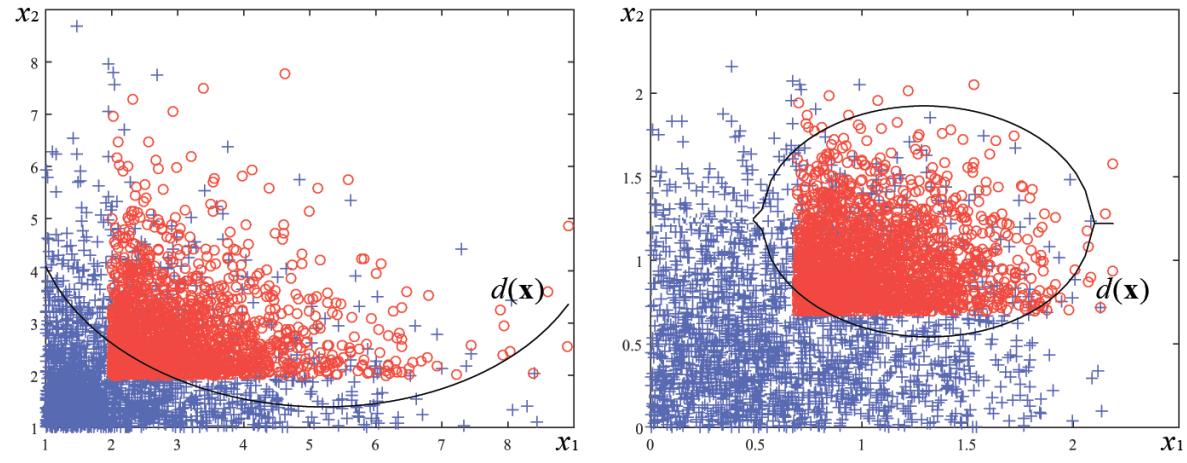
Випадок експоненціальногорозподілу випадкової величини

На рис. 5та 6 представлено результати розв'язання задачі класифікації для рівномірного закону розподілу даних навчальної вибірки. Експериментально збудована функція розподілу ймовірностей для рівномірного закону розподілу ознаки x_1наведено на рис. 5, а. Скоригована за допомогою степеневих перетворень навчальної вибірки функція розподілу ймовірностей ознаки x_1наведено на рис. 5, б. На рис. 6 наведено вирішальні правила класифікації для рівномірного закону розподілу ознак (x_1, x_2): для початкової навчальної вибірки (риа. 6, а) та скоригованої за допомогою степеневих перетворень навчальної вибірки (риа. 6, б).



а б

Рис. 5. Функції розподілу ймовірностейознаки x_1: а – експоненціальний закон розподілу, б – скоригований за допомогою степеневих перетворень навчальної вибірки закон розподілу



а б

Рис. 6. Вирішальні правила $d(x)$ класифікації в просторі ознак (x_1, x_2): а – початкова навчальна вибірка з експоненцыальним законом розподілу, б – скоригований за допомогою степеневих перетворень навчальна вибірка

Отримані результати оцінки якості моделювання з використанням степеневих перетворень для нормалізації законів розподілу ознак навчальної вибірки, а саме, помилки класифікації Δ_{ij} (3) та ймовірності правильного розпізнавання P (4) зведені в Таблицю 2.

Таблиця 2.

Результати оцінки якості моделювання з використанням степеневих перетворень для нормалізації законів розподілу ознак навчальної вибірки

№	Закон розподілу	Якість моделювання						Відносна помилка δ(P)
		Без корегування			З корегуванням			
		Δ_1	Δ_2	P	Δ_1	Δ_2	P	
1	Рівномірний	0.079	0.060	0.86	0.061	0.060	0.88	2.02
2	Біноміальний	0.093	0.038	0.87	0.034	0.036	0.93	5.85
3	Експоненціальний	0.129	0.051	0.82	0.095	0.34	0.86	4.11

Аналіз даних Таблиці 2 демонструє переваги степеневих перетворень для біноміального та експоненціального законів розподілу даних навчальної вибірки. Для

рівномірного закону розподілу даних навчальної вибірки ефективність степеневих перетворень не є досить переконливою: перевага методу проявляється на рівні статистичної помилки. Тому, в якості рекомендації, можна стверджувати, що степеневі перетворення ознак навчальної вибірки доцільно використовувати при законах разподілу ознак біноміальному (та таких, що походять від нього: Пуассона, логнормального, хі-квадратного) та експоненціальному (та таких, що походять від нього: геометричного, Вейбула).

Висновки

В роботі успішно вирішена задача підвищення точності моделей машинного навчання для випадку розподілу даних навчальної вибірки відмінного від нормального шляхом застосування статечних перетворень до ознак навчальної вибірки для стабілізації дисперсії випадкових величин.

Досліджено ефективність методу статистичної класифікації на основі баєсова підходу для побудови діагностичних моделей у вигляді поліномів другого порядку шляхом застосування степеневих перетворень ознак навчальної вибірки для стабілізації їх дисперсії та подальшого використання методу максимальної правдоподібності, що дозволяє підвищити достовірність діагностування у випадку розподілу даних навчальної вибірки відмінного від нормального.

Розглянутий метод стабілізації дисперсії випадкових величин на основі степеневих перетворень апробовано на тестовых данных. На основі проведенихекспериментальнихдослідженьотриманорезультатиефективностізастосуванн я степеневих перетворень даних навчальної вибірки при різноманітнихрозподілах її ознак.Сформульовано рекоммендації по застосуванню степенного перетворення для розглянутих законов розподілення ознак навчальної вибірки: степеневі перетворення доцільно використовувати при законах разподілу ознак біноміальному (та таких, що походять від нього: Пуассона, логнормального, хі-квадратного) та експоненціальному (та таких, що походять від нього: геометричного, Вейбула).

Застосування запропонованих рекомендацій дозволяє забезпечити підвищення достовірності моделей машинного навчання в умовах розподілу даних навчальної вибірки відмінного від нормального.

Список літератури

1. Yan H., Wan J., Zhang C., Tang S., Hua Q., Wang Z. Industrial big data analytics for prediction of remaining useful life based on deep learning. *IEEE Access*. 2018. Vol. 6. P.17190–17197.
2. Zhao R., Yan R., Chen Z., Mao K., Wang P., Gao R. X. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*. 2019. Vol. 115. P. 213–237.
3. Шелухин О.И., Беляков И.В. Негауссовские процессы. СПб: Политехника, 1992. 312с.
4. Кунченко Ю.П., Палагин В.В. Синтез алгоритмов проверки простых статистических гипотез, основанных на использовании стохастических полиномов, оптимальных по критерию асимптотической нормальности. Труды 51-й научной сессии, посвященной дню радио. Москва, 1996.Т.2, С.128-129.
5. ЧабдаровШ.М., СафиуллинН.З., ФеоктистовА.Ю. Основы статистической теории радиосвязи: Полигауссовы модели и методы. Казань: КАИ, 1983. 87 с.
6. Guyon I., Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003. No 3. P. 1157–1182.

7. Gantz J., E. Reinsel. Extracting Value from Chaos.*IDC's Digital Universe Study*, 2011. 12 p.
8. Кунченко Ю.П., Палагін В.В. Розробка теорії нелінійних методів опрацювання негауссівських сигналів на основі застосування стохастичних поліномів. *Труды IX Международной научно-практической конференции «Системы и средства передачи и обработки информации»* (ССПОИ-2005). Черкассы: ЧГТУ, 2005. С. 9-11.
9. Fainzilberg L.S. Mathematical methods for evaluating the usefulness ofdiagnostic features. Kiev: Osvita Ukrainy, 2010. 152 p. [in Russian]
10. Tang, J., Alelyani, S., Liu, H. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*. CRC Press, 2014. P. 37–64.
11. Shahana A.H., Preeja V. Survey on feature subset selection for high dimensional data. *International Conference on Circuit, Power and Computing Technologies* (ICCPCT), Nagercoil, India, 2016.P. 1–4.
12. Jain D., Singh V. Feature selection and classification systems for chronic disease prediction: A review.*Egyptian Informatics Journal*. 2018. Vol. 19, Issue 3. P. 179–189.
13. Liu H., Hiroshi M.Feature Selection for Knowledge Discovery and Data Mining. *The Springer International Series in Engineering and Computer Science*. 1998. 214 p.
14. Kohavi R., JohnG. Wrappers for feature selection. *Artificial Intelligence*. 1997. Vol. 97, Issues 1–2. P. 273–324.
15. Gopika N., Meena kowshalaya M.E. Correlation Based Feature Selection Algorithm for Machine Learning Proceedings. *International Conference on Communication and Electronics Systems* (ICCES), 2018. P. 692–695.
16. Medvedew A., Fomin, O., Pavlenko, V., Speranskyy, V. Diagnostic features space construction using Volterra kernels wavelet transforms.*Proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications* (IDAACS), 2017. P. 1077–1081.
17. Derevianchenko O., Fomin O., Skrypnyk N. Improvement of the Quality for Cutting Tool Monitoring by Optimizing the Features of the State Space. In: Tonkonogyi V. et al. (eds) Advanced Manufacturing Processes II.*Lecture Notes in Mechanical Engineering*. Springer,InterPartner, 2020.
18. Dumas S. Karhunen-Loeve transform and digital signal processing – part 1. SETI League, 2016. 39 p.
19. Вагис Г.А., Гупал А.М., Сергиенко И.В. Эффективность байесовской процедуры распознавания. *Кибернетика и системный анализ*. 2001. №1. С. 71-77.
20. Bernard C. Picinbono. Measurements of Second-Order Properties of Point Processes. *IEEE Trans. Instrumentation and Measurement*.Vol. 57, №3, 2008. P.548-555.

ANALYSIS OF A METHOD FOR STABILIZING THE VARIANCE OF RANDOM VARIABLES BASED ON POWER TRANSFORMATIONS FOR MACHINE LEARNING MODELS

O.O. Fomin[1], M.M. Masri[2], O.A. Tumanov[2], O.V. Serbova[2], V.V. Tkachenko [2]

[1]Odesa national polytechnic university,
Shevchenko ave., 1, Odesa, 65044; Ukraine; fomin@opu.ua
[2]Private Higher Educational Institution «Odesa Technological University «STEP»»,
вул. Садова, 3, Одеса, 65023; Україна; dr.mohanadmasri@gmail.com

In the work the task of creating machine learning models is solved. The aim of the work is to increase the accuracy of machine learning models for the case of a distribution of data of the educational selection other than normal. This goal is achieved by using statistical rearrangements to the features of the educational selection for stabilization of dispersion of incidental values. The most significant results: The efficacy of the method of statistical classification on the basis of the basis of the basic approach for creating models in the form of polynomials of another order, which allows increasing the accuracy of modeling in the case of distribution of data of the educational sample different from the normal one, was investigated, Experimental studies and recommendations on the efficiency of using statistical transformations of the educational selection data at different distributions of the educational selection data for the following laws of distribution of incidental values were made: Equinomial, Binomial, Exponential. Significance of the results: the application of the formulated recommendations allows us to improve the accuracy of machine learning models under conditions of a distribution of educational selection data different from the normal. The provided recommendations were tested on the test assignment data. The task demonstrates an increase in the accuracy of machine learning models for most of the considered laws of distribution of incidental quantities.

Keywords: statistical characteristics, random variable, power transformations, models of machine learning.