biomedical engineering. – 2005. – Vol. 52, n. 12. – P. 2115–2121. 4. Мельник Р.А., Алексеев О.А. Кластеризація мікрообразів для кодування зображень // Пр. міжнар. конф. "Укробраз'2004". – К., 2004. – С. 81–85. 5. Melnyk R., Tushnytskyy R. Decomposition of Visual Patterns // Досвід розробки та застосування приладо-технологічних САПР в мікроелектроніці: Матеріали IX Міжнар. наук.-техн. конф. CADSM 2007. – Львів. – С. 278–279. 6. Мельник Р.А., Алексеев О.А. Кластеризація ключів образів на основі декомпозиції їх множини // Відбір і обробка інформації. – 2006. – Вип. 24(100). – С. 110–114.

УДК 004.032.26:004.048

Р.О. Ткаченко, А.В. Дорошенко
Національний університет "Львівська політехніка",
кафедра автоматизованих систем управління

ВДОСКОНАЛЕННЯ НЕЙРОМЕРЕЖНИХ МЕТОДІВ КЛАСИФІКАЦІЇ В ЗАВДАННЯХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ЗА ДОПОМОГОЮ МЕТОДУ ІМІТАЦІЇ ВІДПАЛУ МЕТАЛУ

Проаналізовано особливості постановки та підходи до розв'язання задач класифікації для випадків великорозмірних завдань інтелектуального аналізу даних. Подано основи нейромережного методу класифікації, вдосконаленого завдяки оптимізації методу імітації відпалу металу, та результати його апробації.

The article analyses the features of the target setting and the approach to solving a problem of classification task for Data Mining tasks where data are high-dimensional. Essential principles of the method of classification on the base of neural networks are proposed. This method of classification is improved by simulated annealing algorithm. The results of classification are proposed.

Вступ. Сучасні інформаційні технології надають можливість накопичувати величезні обсяги даних в усіх сферах людської діяльності, від повсякденних та ділових (дані про транзакції в супермаркетах, записи про використання кредитних карток, інформація про телефонні дзвінки та урядову статистику) до наукових (такі, як зображення астрономічних тіл, база даних молекул чи медичних записів). Враховуючи гігантські обсяги накопичених даних, було розроблено концепцію сховищ даних, що втілила досконалішу технологію запису і збереження інформації. Однак головною метою накопичення даних є не стільки їхнє збереження, скільки аналіз. З одного боку, збільшення обсягів даних надає більші можливості для отримання нової корисної інформації та знань, з іншого – зменшується можливість аналізу даних. Для точного й оперативного аналізу даних виникли нові, відмінні від традиційно статистичних, методи, об'єднані під назвою видобуток даних, або інтелектуальний аналіз даних.

Постановка задачі. В основу багатьох методів класифікації покладено гіпотезу компактності [1], яка передбачає, що об'єкти, які належать до одного класу, формують певні кластери в просторі ознак, а отже, можуть бути розділені гіперповерхнями простого вигляду. Однак у багатьох випадках для систем, представлених в умовах невизначеності, існує взаємне перекриття класів. Це спричиняється неповнотою інформаційного базису, суперечливістю даних та іншими факторами. Тому класичні методи класифікації, які ґрунтуються на основі гіпотези компактності, є неефективними, а отже, необхідно знаходити нові.

Пропонуємо розглянути поєднання методу кусково-лінійної класифікації на основі моделі геометричних перетворень (нейронні мережі типу "функціонал на множині табличних функцій"), який

дає змогу роз'язувати задачі інтелектуального аналізу даних в умовах неповноти інформаційного базису та враховує такі особливості задач видобутку даних, як: об'ємність завдань, виродженість задач, суперечливість та неповнота даних, неоднорідність представлення даних, нерівномірність даних у просторі, різна вага помилок, та методу глобальної оптимізації – алгоритму імітації відпалу металу.

Розглянемо їх застосування для розв'язання задачі, описаної в [2] – оцінки ризиків втрати коштів в процесі online-торгівлі. В online-магазинах, окрім передоплати, використовують велику кількість різних способів оплати замовлення – від відкриття покупцю рахунку до зняття грошей з його розрахункового рахунку чи кредитної картки. Із збільшенням популярності online-торгівлі збільшується кількість клієнтів і зростають обсяги продажів, що приваблює до інтернет-магазинів велику кількість дрібних та крупних шахраїв. Відповідно, для оцінки ризиків втрати коштів необхідно класифікувати клієнтів на надійних (з високою ймовірністю оплати замовлення) та ненадійних покупців.

Розглянемо постановку задачі і дані, надані в межах конкурсу Data Mining Cup 2005 (http://www.data-mining-cup.com/2005). Тренувальна вибірка складається з даних про 30000 замовлень, в яких на основі спостережень за 4 тижні визначено їхню належність до одного з двох класів. Необхідно розробити таку систему, яка б дала змогу передбачати факт втрати платежів за замовлення, які надходитимуть, та відносити їх до одного з двох класів відповідно до наведеної матриці вартостей. Розроблену систему тестуємо на виборці з 20000 замовлень, належність яких до одного з двох класів є невідомою, однак може бути перевіреною.

За умовами конкурсу класифікація замовлень на високоризикові чи низькоризикові здійснюється відповідно до матриці ваг (табл. 1).

Таблиця 1

Матриця ваг для нарахування балів вартості

Замовлення	Втрата платежів	Звичайне замовлення
Замовлення, класифіковане як високоризикове	2	-25
Замовлення, класифіковане як низькоризикове	13	15

Основною метою класифікації тестових даних є мінімінізація штрафних балів, які, відповідно до матриці ваг, нараховують так: якщо замовлення є низькоризиковим, а класифікується як високоризикове – нараховуємо 27 штрафних балів, якщо замовлення є високоризиковим, а класифікується як низькоризикове – 2 штрафні бали.

Розв'язання задачі. У результаті класифікації цих даних за допомогою нейромережі, побудованої на основі МГП, без будь-яких додаткових перетворень було отримано результати, наведені в табл. 2.

Таблиця 2

Результати класифікації із застосуванням моделі геометричних перетворень

	Тестові дані		Штрафні бали		∑ штрафів
	так	ні	так → ні (×27)	ні → так (×2)	
Реальні значення	1156	18844			
Правильно спрогнозовані значення	849	11984			
Неправильно спрогнозовані значення	307	6860	8289	13720	**22009**

Метод кусково-лінійної класифікації у поєднанні із методом штрафних функцій. Пропонується застосувати кусково-лінійний підхід у поєднанні із методом штрафних функцій для розв'язання задачі, що розглядається.

Для цього формуємо матрицю штрафних функцій (табл. 3).

Таблиця 3

Матриця штрафних функцій

Замовлення	Втрата платежів	Звичайне замовлення
Замовлення, класифіковане як високоризикове	k	-k
Замовлення, класифіковане як низькоризикове	- k	k

Після визначення коефіцієнтів матриці штрафних функцій замінюємо символьні визначники належності кожного клієнта до одного з класів (“так” – для замовлень з високим ризиком, “ні” – для замовлень з низьким ризиком) на відповідні пари значень з матриці штрафних функцій – (k; -k) для “так” та (-k; k) для “ні”.

Після цього задачу класифікації розглядаємо як задачу прогнозування: на основі вхідних даних прогнозуємо значення пари виходів. Спрогнозовані пари значень аналізуємо так: якщо перше значення в парі більше ніж друге – присвоюємо виходу цього вектора значення ”так”, у протилежному випадку – значення ”ні”. Порівнюємо отримані значення виходів із значеннями виходів, відомими для тренувальної вибірки, та підраховуємо кількість штрафних балів.

Змінюємо значення коефіцієнта k та повторюємо процедуру, доки не отримаємо мінімальне значення штрафних балів. Вибірки із спрогнозованими виходами (як тренувальну, так і тестову) – ділимо на 2 кластери: вектори, розпізнані як ”так”, та вектори, розпізнані як “ні”. Після кластеризації в тренувальні вибірки, отримані для кожного кластера, підставляємо реальні значення виходів. Також окремо для кожного кластера підбираємо значення коефіцієнта k, для якого сума штрафів в цьому кластері є мінімальною.

Якщо для певного кластера оптимальним є значення k, за якого нейромережа передбачає для тренувальної вибірки лише один клас (”так” чи “ні”) – алгоритм кластеризації для цього кластера зупиняється, інакше – продовжуємо виконувати кластеризацію (рис. 1). Після зупинки обробки кожного з кластерів аналізуємо тестові дані та обчислюємо суму штрафних балів за кожним з кластерів. Отримані результати наведено у табл. 4.

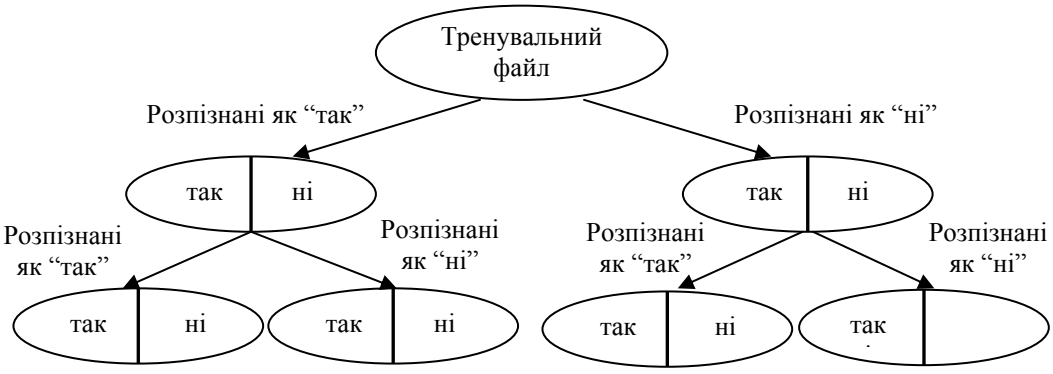

Рис. 1. Схема виконання алгоритму кластеризації

За допомогою методу кусково-лінійної класифікації на основі моделі геометричних перетворень до початкових даних отримано кластери різного розміру, які містять подібні вектори даних та з високою ймовірністю належать до одного класу. Застосування окремо до кожного з цих класів методу штрафних функцій дало змогу значно підвищити точність класифікації.

Таблиця 4

Результати класифікації із застосуванням методу штрафних функцій на основі нейромережної реалізації

	Тестові дані		Штрафні бали		Σ штрафів
	так	ні	т → н (×27)	н → т (×2)	
Реальні значення	1156	18844			
Правильно спрогнозовані значення	627	15624			
Неправильно спрогнозовані значення	529	3220	14283	6440	**20723**

Метод імітації відпалу металу. Пропонується поєднати метод штрафних функцій на основі нейромережної реалізації із оптимізаційним методом імітації відпалу металу для подальшого збільшення точності класифікації.

На рис. 2 зображено структурну схему розробленої нейромережі на основі моделі геометричних перетворень, де $x_1, x_2, ..., x_n$ – первинні ознаки об'єктів класифікації – вхідні дані, $ГК_1, ГК_2, ..., ГК_n$ – головні компоненти, отримані на основі вхідних даних, $w_1, w_2, ..., w_n$ – вагові коефіцієнти, $\tilde{y}$ – вихід, що задає належність до визначених класів.

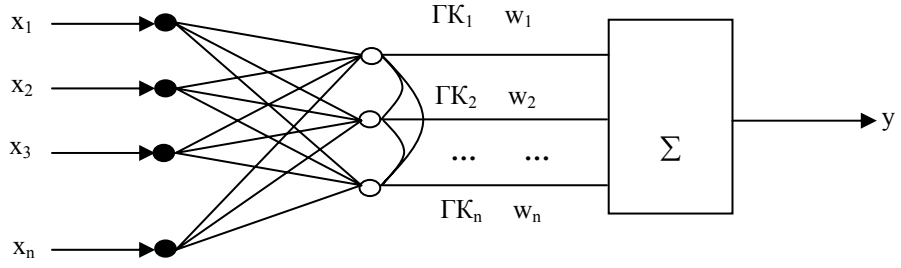

Рис. 2. Структурна схема нейромережі на основі ФМТФ

Функціонування такої нейронної мережі можна описати формулою $y = \sum_{i=1}^{n} ГК \cdot w_i$. Метод імітації відпалу металу пропонується застосовувати для оптимізації вагових коефіцієнтів так, щоби результуюча сума штрафних балів була мінімальною.

Метод імітації відпалу є алгоритмічним аналогом фізичного процесу керованого охолодження і дає змогу практично знаходити глобальний мінімум функції декількох змінних. Алгоритм імітації відпалу побудовано на ідеї, запозиченій із статичної механіки. Він відображає поведінку матеріального тіла під час затвердіння із застосуванням процедури відпалу – керованого охолодження при температурі, що послідовно знижується до нуля. Відповідно до проведених науковцями досліджень, під час затвердіння розплавленого матеріалу його температура має знижуватись поступово до моменту повної кристалізації. Якщо процедура остигання відбувається занадто швидко, то утворюються значні нерегулярності структури матеріалу, які викликають внутрішнє напруження. В результаті загальний енергетичний стан тіла, що залежить від його внутрішньої напруженості, залишається на набагато вищому рівні, ніж у разі повільного охолодження [4].

Швидка фіксація енергетичного стану тіла на рівні, вищому за нормальний, аналогічна до збіжності оптимізаційного алгоритму до точки локального мінімуму. Енергія стану тіла відповідає цільовій функції, а абсолютний мінімум цієї енергії – глобальному мінімуму. Однак допускаються ситуації, в яких енергія може на деякий час збільшуватись. Це забезпечує вихід із пасток локальних мінімумів, які виникають при реалізації процесу. Лише зменшення температури до абсолютного нуля робить неможливим будь-яке самостійне збільшення його енергетичного рівня.

На ефективність роботи алгоритму імітації відпалу надзвичайно великий вплив має вибір таких параметрів, як початкова температура $T_{\max}$, коефіцієнт зменшення температури r та

кількість циклів L, що виконуються на кожному температурному рівні. Необхідно зазначити, що для кожного кластера значення цих параметрів можуть бути різними.

Модифікований алгоритм імітації відпалу металу у поєднанні із методом штрафних функцій. 1. Запустити процес з початкової точки w, обраної випадковим чином при заданій початковій температурі $T = T_{\max} = 20895$, що дорівнює максимальному значенню штрафних функцій в початковій точці.

2. Доки $T > 0.5$, повторити $L=100$ разів такі дії:

- обрати новий розв'язок w' з околу w;
- розрахувати зміну цільової функції $\Delta = E(w') - E(w)$, де значенням цільової функції є сума штрафних функцій;
- якщо $\Delta \le 0$ – прийняти $w = w'$; інакше, при $\Delta > 0$, прийняти $w = w'$ з ймовірністю $\exp(-\Delta / T)$ шляхом генерації випадкового числа R з інтервалу (0,1) з подальшим порівнянням його із значенням $\exp(-\Delta / T)$; якщо $\exp(-\Delta / T) > R$, прийняти новий розв'язок $w = w'$; у протилежному випадку – проігнорувати його.

3. Зменшити температуру $(T = rT)$ з використанням коефіцієнта зменшення r, що обирається з інтервалу (0,1), та повернутися до пункту 2. Пропонується використовувати значення r = 0,9

Отримані результати наведено у табл. 5.

Таблиця 5

Результати класифікації
із застосуванням методу штрафних функцій та методу імітації відпалу металу

	Тестові дані		Штрафні бали		∑
	так	ні	т → н (×27)	н → т (×2)	штрафів
Реальні значення	1156	18844			
Правильно спрогнозовані значення	670	15105			
Неправильно спрогнозовані значення	486	3739	13122	7746	**20600**

Висновки. Проведені дослідження та отримані результати експериментів свідчать про доцільність вдосконалення нейромережних методів класифікації в завданнях видобутку даних шляхом застосування методу кусково-лінійної класифікації та методу імітації відпалу металу, оскільки це дає можливість отримати мінімум функції, наближений до глобального, а відповідно й мінімальну кількість штрафних балів, тобто підвищити точність класифікації.

1. Васильев В.И., Коноваленко В.В., Горелов Ю.И. Имитационное управление неопределенными объектами. – К.: Наук. думка, 1989. – 216 с. 2. *Дорошенко А.В. Нейромережний розв'язок задач класифікації в умовах неповноти інформаційного базису // Моделювання та керування станом еколого-економічних систем регіону: Зб. наук. пр. – К., 2006. – Вип. 3. – С. 115–122. 3. Tkachenko R., Tkachenko P., Tkachenko O., Schmitz J. Geometrical Data Modelling // Зб. матеріалів між нар. Наук. конф. "Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій" (ISDMIT' 2006). – Т. 2. – С. 279–283. 4. Хайкин С. Нейронные сети: полный курс: Пер с англ. – М.: Вильямс, 2006. – 1104 с. 5. Осовский С. Нейронные сети для обработки информации / Пер. с польск. И.Д. Рудинского. – М.: Финансы и статистика, 2004. – 344 с.*