МОДЕЛЮВАННЯ ПРОЦЕСІВ І СИСТЕМ

УДК 004.032.26:004.048

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

КЛАСИФІКАЦІЯ ДАНИХ В УМОВАХ НЕВИЗНАЧЕНОСТЕЙ ЗА ДОПОМОГОЮ НЕЙРОПОДІБНИХ СТРУКТУР НА ОСНОВІ МОДЕЛІ ГЕОМЕТРИЧНИХ ПЕРЕТВОРЕНЬ

Проаналізовано особливості постановки та підходи до розв'язання задач класифікації для випадків великорозмірних завдань інтелектуального аналізу даних. Подано основи розроблених нейромережних методів класифікації, результати проведених експериментів.

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 methods of classification on the base of neural networks and the results of experiments are proposed.

Вступ

Численні успішні приклади застосувань засобів інтелектуального опрацювання інформації як у наукових дослідженнях, так і у різноманітних бізнес-ужитках призвели до того, що все більше компаній з різних галузей хочуть за допомогою методів інтелектуального аналізу даних видобувати знання з величезних сховищ даних, що накопичились в них завдяки розвитку інформаційних технологій та впровадження їх у всі сфери людської діяльності. Однак, надзвичайно великий обсяг сховищ даних, що використовуються для пошуку знань, а також високі вимоги до достовірності отриманих знань змушують шукати нові або вдосконалювати існуючі методи інтелектуального аналізу даних.

Постановка задачі

Серед завдань інтелектуального аналізу даних розглядаються задачі класифікації, кластеризації, асоціації та прогнозування. Як приклад розглянемо задачу класифікації, сформульовану організаторами провідної німецької лотереї South German Class Lottery в межах конкурсу Data Mining Cup 2008 (http://www.data-mining-cup.com/2008). Особливістю цієї лотереї є те, що кількість та розмір призів, що розігрується, визначена та оголошена наперед і не залежить від того, скільки білетів продано. Тривалість лотереї – шість місяців, в кожному з яких проводиться окремий розіграш. Взяти участь у кожному з розіграшів може лише той учасник, який брав участь у всіх попередніх розіграшах. Отже, організатори лотереї ще до її початку потребують інформації про те, скільки білетів буде продано та на який прибуток вони можуть розраховувати.

Виходячи із цих потреб, було сформульовано завдання – розділити всіх гравців на 5 класів: ті, хто брав участь лише в одному розіграші, але не платив за білет; ті, хто брав участь лише в одному розіграші, але платив за білет; ті, хто брав участь принаймні у двох розіграшах; той, хто брав участь у всіх розіграшах, але не збирається грати в наступній лотереї; ті, хто брав участь у всіх розіграшах і купив принаймні один білет наступної лотереї.

Кожна помилка класифікації має свою вагу та оцінюється згідно з таблицею, сформованою організаторами лотереї (табл. 1).

Таблиця 1

Матриця ваг для нарахування балів вартості

Прогнозована належність клієнтів до класів	**Існуюча належність клієнтів до класів**				
	1	**2**	**3**	**4**	**5**
1	20	0	-10	-20	-40
2	5	20	0	-10	-20
3	0	5	20	0	-10
4	-5	0	5	20	0
5	-10	-5	0	5	20

Тренувальна та тестова вибірки складаються з 113456 записів про гравців, кожен з яких має 70 атрибутів, зокрема: стать, вік та сімейний стан гравця, інформація про його банк та марку автомобіля, кредитну привабливість тощо.

Складність такої задачі полягає в тому, що через невизначеності, спричинені пропусками в даних, їх суперечливістю тощо, для неї не виконується гіпотеза компактності, що покладено в основу багатьох методів класифікації [1], різні класи перекриваються між собою, що унеможливлює їх розділення гіперповерхнями простого вигляду.

Для розв'язання цієї проблеми пропонується застосувати кусковий метод побудови розділяючих поверхонь на основі моделі геометричних перетворень [2], модифікований для задачі класифікації на більше ніж два класи, який дає змогу врахувати нелінійність задач видобування даних, але не вимагає великої кількості часу для виконання. Окрім цього, ще однією перевагою застосування кускового методу побудови розділяючих поверхонь є те, що за його допомогою, завдяки розділенню загальної вибірки на кластери, можна опрацьовувати всю вибірку за прийнятний час.

З використанням дерева поділу на класи можна об'єднувати в окремі кластери вектори даних, що мають схожі вхідні показники та аналізувати їх незалежно один від одного. Після того, як отримано значення штрафних балів за кожним з кластерів, вони підсумовуються. Такий підхід дає змогу істотно підвищити загальну точність класифікації.

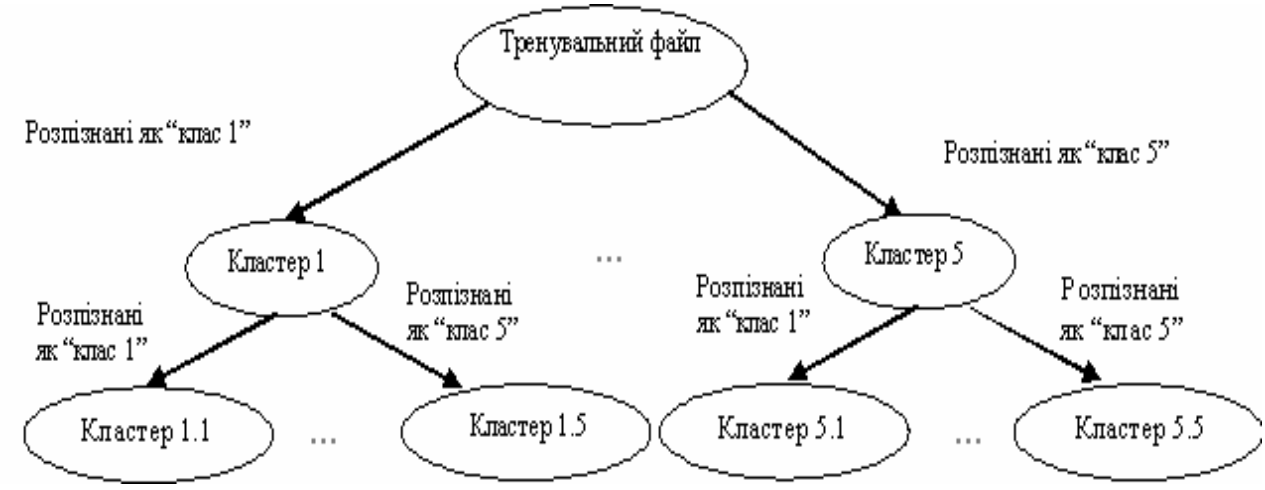

Рис. 1. Дерево поділу на кластери для кускового методу побудови розділяючих поверхонь

Також, для врахування різної ваги помилок (табл. 1), для кожного з кластерів використаємо метод штрафів та заохочень [3]. Відповідно, у цьому випадку задача класифікації зводиться до задачі максимізації суми заохочувальних балів.

Таблиця 2

Матриця штрафів та заохочень

	Вектор розпізнано як клас 1	Вектор розпізнано як клас 2		Вектор розпізнано як клас K
Вектор належить до класу 1	a_{11}	a_{12}	...	a_{1K}
Вектор належить до класу 2	a_{21}	a_{22}	...	a_{2K}
…	...	...	...	...
Вектор належить до класу K	a_{K1}	a_{K2}	...	a_{KK}

Отже, запропонований нами алгоритм поєднання використання моделі геометричних перетворень із методом штрафів та заохочень має вигляд:

Алгоритм класифікації із використанням методу штрафів та заохочень

1. У навчальній вибірці замінюємо ідентифікатори класів відповідними коефіцієнтами:

Клас 1 → (a_{11}; a_{12};…; a_{1K})

Клас 2 → (a_{21}; a_{22};…; a_{2K})

…

Клас K → (a_{K1}; a_{K2};…; a_{KK})

2. На отриманій навчальній вибірці вчимо нейроподібну структуру на основі моделі геометричних перетворень (МГП) такої структури [3,4]:

Рис. 2. Нейроподібна структура МГП

Через навчену нейронну мережу пропускаємо тестові дані.

3. Аналізуємо коефіцієнти (a_1,…, a_k), отримані на виходах МГП для кожного вектора вхідних даних з тестового файла за правилом «переможець забирає все».

4. Для тестової вибірки підраховуємо кількість штрафних балів відповідно до матриці штрафів.

5. Основною метою алгоритму є максимізація заохочувальних балів.

Як правило, попередньо визначають або суму штрафів, яка є прийнятною для даної задачі, або час, впродовж якого буде виконуватись мінімізація – це умови зупинки виконання алгоритму.

Розглянемо результати експериментів, проведених для сформульованої вище задачі.

У табл. 3 наведено результати навчання нейроподібної структури МГП, навченої на 10000 навчальних векторах та протестованої на 2000 випадкових тестових векторах.

Таблиця 3

Результати класифікації за допомогою нейроподібної структури МГП

	Сума набраних балів
Під час тренування	62745
Під час тестування	11915

Після застосування кускового методу побудови розділяючих поверхонь на основі моделі геометричних перетворень із застосуванням методу штрафів та заохочень було отримано результати, подані в табл.4.

Таблиця 4

Результати класифікації після застосування кускового методу побудови розділяючих поверхонь поєднано із методом штрафів та заохочень

	Сума набраних балів
Під час тренування	67515
Під час тестування	13615

Вдосконалення методу штрафів та заохочень шляхом застосування модифікованого алгоритму імітації відпалу металу

Пропонуємо розглянути поєднання методу кускової побудови розділяючих поверхонь на основі моделі геометричних перетворень та методу глобальної оптимізації – алгоритму імітації відпалу металу.

Пропонується поєднати метод штрафів та заохочень на базі нейромережної реалізації із оптимізаційним методом імітації відпалу металу для подальшого збільшення точності класифікації.

На рис. 3 зображено структурну схему розробленої нейроподібної структури 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, y – вихід, що задає належність до визначених класів.

Функціонування такої нейроподібної структури можна описати формулою $y = \sum_{i=1}^{n} ГК \cdot w_i$.

Метод імітації відпалу металу пропонується застосовувати для оптимізації вагових коефіцієнтів так, щоб результуюча сума заохочувальних балів була максимальною.

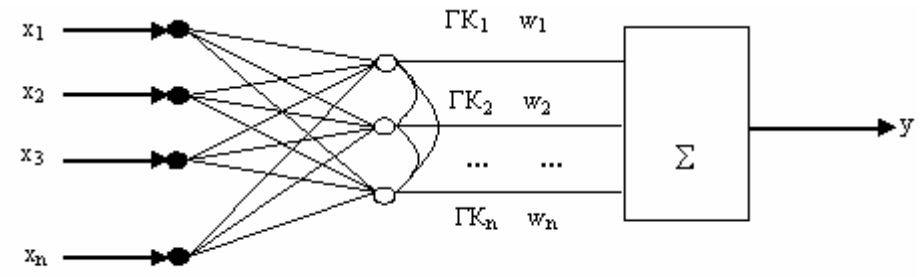

Рис.3. Структурна схема нейроподібної структури на основі МГП

Модифікований алгоритм імітації відпалу металу поєднано із методом штрафів та заохочень

1. Запустити процес з початкової точки w, обраної випадково при заданій початковій температурі $T = T_{\max}$, що дорівнює мінімальному значенню заохочувальних балів у 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 заохочувальних балів;

§ 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.

Розроблений модифікований алгоритм імітації відпалу металу поєднано із методом штрафів і заохочень застосовується для покращання результатів класифікації для кожного з кластерів (рис. 1).

У табл. 4 наведено результати експерименту після оптимізації результатів за допомогою застосування алгоритму імітації відпалу металу.

Таблиця 4

Результати класифікації після застосування кускового методу побудови розділяючих поверхонь із застосуванням алгоритму імітації відпалу металу

	Сума набраних балів					
	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5	Разом
Під час тренування	62400	0	33425	54205	117090	267120
Під час тестування	4345	0	1657	3112	8723	17837

Висновки

Розглянуто методи класифікації на основі моделі геометричних перетворень, орієнтовані на розв'язання задач інтелектуального аналізу даних, перевагою яких є висока швидкодія . Описані методи враховують основні особливості завдань видобування даних та дають змогу опрацьовувати великі обсяги даних за невеликий час. Крім того, вдосконалення нейромережних методів класифікації в завданнях видобування даних шляхом застосування методу кускової побудови розділяючих поверхонь та методу імітації відпалу металу дає можливість отримати максимум функції, наближений до глобального, а відповідно й максимальну кількість балів, тобто підвищити точність класифікації.

1. Васильев В.И., Коноваленко В.В., Горелов Ю.И. Имитационное управление неопределенными объектами. – К.: Наукова думка, 1989. – 216с. 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. Ткаченко Р.О. Модель нейронних мереж // Вісник Держ. ун-ту "Львівська політехніка": Комп'ютерна інженерія та інформаційні технології. – 1998. – № 349. – С.83–86. 4. Ткаченко Р.О. Нейронні мережі з нелінійними синаптичними зв'язками // Вісник Держ. ун-ту "Львівська політехніка": Комп'ютерні системи проектування. Теорія і практика. – 1999. – № 373. – С.20–22. 5. Ткаченко Р.О., Ткаченко П.Р. Багатошаровий перцептрон з неітеративним навчанням // Збірник матеріалів міжнародної наукової конференції "Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій" (ISDMIT' 2005). – Т.5. – С.69–73. 6. Tkachenko R., Tkachenko P., Tkachenko O.,Schmitz J. Geometrical Data Modelling // Збірник матеріалів міжнародної наукової конференції "Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій" (ISDMIT' 2006). – Т.2. – С.279–283. 7. Хайкин С. Нейронные сети: полный курс: Пер с англ. – М.: "Вильямс", 2006. – 1104 с. 8. Осовский С. Нейронные сети для обработки информации / Пер. с польск. И.Д. Рудинского. – М.: Финансы и статистика, 2004. – 344 с.