УДК 004.032.26:004.048

А. Дорошенко
Національний університет ”Львівська політехніка”,
кафедра автоматизованих систем управління

МЕТОДИ КЛАСИФІКАЦІЇ НА ОСНОВІ МОДЕЛІ ГЕОМЕТРИЧНИХ ПЕРЕТВОРЕНЬ ДЛЯ ЗАВДАНЬ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Проаналізовано особливості постановки та підходи до розв’язку задач класифікації для випадків великорозмірних завдань інтелектуального аналізу даних. Подано основи розроблених методів класифікації, вдосконалених завдяки використанню оптимізації відпалом металу.

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 are proposed. This methods of classification are improved by simulated annealing algorithm.

Вступ

Розвиток сучасних інформаційних технологій надає можливість створювати сховища даних, що містять величезні обсяги даних з усіх сфер 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). Однак разом із збільшенням розміру сховищ даних ускладнюється можливість їх аналізу, який є головною метою накопичення 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

Більшість відомих методів класифікації не є адаптованими для задач інтелектуального аналізу даних, а саме до таких основних характерних особливостей завдань видобутку даних, як: об'ємність завдань та велика розмірність даних, виродженість задач, суперечливість, неповнота та неоднорідність даних.

Коротко розглянемо кожну з цих особливостей окремо:

1. Об'ємність завдань.

Якість інтелектуального аналізу даних певною мірою залежить від розмірів сховища даних, що обробляється. Відповідно для розв'язання задач видобутку даних із максимальною точністю обробляють вибірки великого обсягу та розмірності. Однак, це не лише підвищує якість видобутку даних, але й ускладнює його процес, робить його обчислювально складнішим.

2. Виродженість задач

Оскільки вхідні дані в задачах інтелектуального аналізу даних, як правило, є корельованими між собою, то в багатьох випадках такі задачі є виродженими.

Крім того, системи рівнянь, що описують залежність між вхідними та вихідними даними, часто є погано зумовленими (або надчутливими), тобто навіть незначна зміна коефіцієнтів задачі здатні призвести до великих змін її розв’язку. Відповідно, погано зумовлена задача не може бути розв’язана класичними методами (методом найменших квадратів тощо). Для розв’язання вироджених задач можливе застосування методу сингулярної декомпозиції, однак цей метод є дуже об’ємним та обчислювально складним. Наприклад, порівняно із QR-факторизацією, сингулярна декомпозиція

вимагає в 5-10 разів більше арифметичних операцій. Крім того, сингулярну декомпозицію неможливо ефективно обновити, коли змінюються дані. Тому, враховуючи розміри вибірок, що обробляються при інтелектуальному аналізі даних, метод сингулярної декомпозиції не підходить для розв'язання задач видобутку даних.

3. Суперечливість, неповнота даних

Задачі інтелектуального аналізу даних переважно полягають у видобуванні знань з даних, зібраних на основі опитувань, анкет, експериментальних даних тощо. Тобто такі дані не описуються чіткими математичними залежностями, а є емпіричними і, відповідно, можуть бути суперечливими. Відповідно, дані, зібрані таким чином можуть мати пропуски, спричинені відмовою вимірювальної техніки, небажанням респондентів давати відповіді на окремі питання, помилкою оператора тощо, тобто бути неповними.

4. Неоднорідність даних

Інформація, що обробляється під час інтелектуального аналізу даних, є різнорідною (кількісною, якісною, текстовою). Це ускладнює процес автоматизованої обробки даних, вимагає їх передобробки.

5. Нерівномірність представлення даних

Ця особливість є характерною для такої задачі видобутку даних, як класифікація. Кількість екземплярів одного класу може на порядок відрізнятися від кількості екземплярів інших класів, що значно ускладнює процес класифікації. Крім того, густина розподілу екземплярів різних класів в просторі ознак також може бути різною.

6. Різна вага помилок

Залежно від умови задачі, що розв'язується, кожен тип помилки може мати свою вагу. Найбільш часто такі умови ставлять для задач класифікації. Відповідно, під час розв'язання задачі класифікації необхідно враховувати не лише точність розпізнавання кожного з класів, а і їх взаємозалежність, для того, щоб вага загальної кількості помилок була мінімальною

Крім того, 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, в багатьох випадках для систем, представлених в умовах невизначеності, існує взаємне перекриття класів. Це спричиняється неповнотою інформаційного базису, суперечливістю даних та іншими факторами.

Сукупність цих особливостей робить складним або й неможливим застосування класичних методів класифікації (нейронні мережі, машини опорних векторів тощо) та спонукає шукати нові методи класифікації для розв'язання завдань інтелектуального аналізу даних.

Розглянемо декілька розроблених підходів до класифікації, що грунтуються на моделі геометричних перетворень та орієнтовані на розв'язання завдань інтелектуального аналізу даних.

1. Метод штрафних функцій на основі моделі геометричних перетворень

Розглянемо метод класифікації на основі моделі геометричних перетворень із застосуванням методу штрафних функцій, розроблений з метою підвищення точності класифікації даних нейромережею на основі моделі геометричних перетворень.

Особливості побудови матриці заохочень для задач класифікації

Оскільки різні помилки мають різну вагу та для багатьох задач інтелектуального аналізу даних важливо не лише те, чи правильно виконано класифікацію, але й те, як саме помилилася система, то задачу класифікації додатково описуємо матрицею штрафів, яка виконує роль обмежень. Тоді цільовою функцією є мінімізація суми штрафів.

Проведемо експеримент щодо застосування методу штрафних функцій для розв'язання задач інтелектуального аналізу даних.

Нехай дані містяться у великому сховищі правильно класифікованих прикладів типу вхід–вихід $\{(\mathrm{x}_i, y)\}_{i=1}^{N}$, де x_i – вхідний вектор, а d_i – бажаний вихідний сигнал, що відповідає вхідному вектору.

Крім того, задача характеризується різною вагою помилок. Вага кожної помилки визначається з матриці ваг, що формується відповідно до умов задачі. Матрицю ваг наведено у табл.1.

Таблиця 1

Матриця ваг

	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}

Оскільки вектори, що належать до різних класів, як правило, розташовуються в просторі нерівномірно, то для покращання точності класифікації пропонується враховувати частотність кожного з класів.

Для цього у початковій вибірці обчислюємо кількість екземплярів кожного з класів ($NC_1, NC_2, ..., NC_K$ відповідно).

Після цього обраховуємо частотності класів за формулою:

$$CH_i = \frac{NC_i}{m} \tag{1}$$

де CH_i – частотність *i*-го класу ($i = 1,...,K$)

NC_i – кількість екземплярів *i*-го класу ($i = 1,...,K$)

m – загальна кількість векторів в базі даних.

Після цього формуємо нову матрицю ваг із врахуванням обчислених частотностей (табл. 2).

Таблиця 2

Матриця ваг із врахуванням частотностей

	Вектор розпізнано як клас 1	Вектор розпізнано як клас 2		Вектор розпізнано як клас K
Вектор належить до класу 1	$a_{11} \times CH_1$	$a_{12} \times CH_1$	...	$a_{1K} \times CH_1$
Вектор належить до класу 2	$a_{21} \times CH_2$	$a_{22} \times CH_2$	...	$a_{2K} \times CH_2$
…	...	...	...	...
Вектор належить до класу K	$a_{K1} \times CH_K$	$a_{K2} \times CH_K$	...	$a_{KK} \times CH_K$

Розв'язання задачі класифікації за допомогою методу штрафних функцій

Під час розв'язання задачі класифікації за допомогою методу штрафних функцій використовуватимемо сформовану вище матрицю штрафів із врахуванням частотностей.

Розглянемо задачу класифікації, в якій необхідно визначити належність об'єкта до одного з двох наперед відомих класів.

При цьому відома матриця ваг:

	Об'єкт класифіковано як об'єкт класу 1	Об'єкт класифіковано як об'єкт класу 2
Об'єкт, що належить до класу 1	a11	a12
Об'єкт, що належить до класу 2	a21	a22

Цю матрицю ваг можна перетворити в таку матрицю штрафів:

Об’єкт, що належить до класу 1 неправильно класифіковано як об’єкт класу 2	a11–a12
Об’єкт, що належить до класу 2 неправильно класифіковано як об’єкт класу 1	a22–a21

Цільовою функцією є мінімізація отриманих штрафних балів.

Отже, запропонований нами алгоритм поєднання використання моделі геометричних перетворень із оптимізаційним методом штрафних функцій має вигляд:

Алгоритм класифікації із використанням методу штрафних функцій

1. Модифікуємо матрицю ваг відповно до частоти розподілу у навчальній вибірці елементів класу 1 та елементів класу 2.

Тобто, якщо елементів класу 1 у навчальній вибірці CH1 %, а елементів класу 1 у навчальній вибірці CH2 %, то матриця ваг набуває вигляду:

	Об’єкт класифіковано як об’єкт класу 1	Об’єкт класифіковано як об’єкт класу 2
Об’єкт, що належить до класу 1	$\alpha 1 = k11 * CH1$	$\alpha 2 = k12 * CH1$
Об’єкт, що належить до класу 2	$\alpha 3 = k21 * CH2$	$\alpha 4 = k22 * CH2$

2. У навчальній вибірці замінюємо ідентифікатори класів відповідними парами коефіцієнтів:

$$\text{Клас } 1 \rightarrow (\alpha 1; \alpha 2) = (a11 * CH1; a12 * CH1)$$

$$\text{Клас } 2 \rightarrow (\alpha 3; \alpha 4) = (a21 * CH2; a22 * CH2)$$

3. На отриманій навчальній вибірці вчимо нейронну мережу прямого поширення типу ФМТФ („функціонал на множині табличних функцій”) такої структури [3,4]:

Рис. 1. Структура нейронної мережі ФМТФ

На рис 1. (X_1, ..., X_N) – вхідні дані, (a1, a2) – спрогнозовані коефіцієнти – вихідні дані нейронної мережі.

Нейронну мережу ФМТФ було обрано через такі її основні переваги, як:

- швидке навчання нейронної мережі прямого поширення за наперед визначену кількість кроків, що дозволяє розв’язувати великорозмірні задачі;
- можливість отримання задовільних результатів при невеликій кількості тренувальних даних;
- на відміну від нейромереж типу „чорної скриньки” є можливість аналізувати внутрішню структуру даних.

4. Через навчену нейронну мережу пропускаємо тестові дані.

5. Аналізуємо пари коефіцієнтів (a1, a2), отримані на виходах нейронної мережі для кожного вектора вхідних даних з тестового файла.

Якщо a1 > a2 – присвоюємо цьому вектору ідентифікатор ”клас 1”, якщо a1 < a2 – ідентифікатор ”клас 2”.

Це правило випливає із матриці ваг, оскільки для ”класу 1” більшим завжди є перший коефіцієнт (a11 > a12), а для ”класу 2” – другий (a21 < a22).

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

Тобто: якщо елемент класу 1 розпізнано як елемент класу 2 – додаємо (a11–a12) штрафних балів, якщо елемент класу 2 розпізнано як елемент класу 1 – додаємо (a22–a21) штрафних балів, якщо класифікація виконана правильно – додається нуль.

7. Основною метою алгоритму є мінімізація суми штрафів.

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

Якщо жодна з умов припинення виконання алгоритму не виконується – змінюємо коефіцієнти a11, a12, a21, a22 та повторюємо кроки 2-7.

2. Метод кусково-лінійної класифікації на основі моделі геометричних перетворень

Використання лінійних методів класифікації для задач інтелектуального аналізу даних є оптимальними з точки зору швидкодії розв'язання поставлених задач, однак, вони не дають достатньої точності класифікації.

Для вирішення цієї проблеми ми пропонуємо застосувати кусково-лінійний підхід до класифікації, який, з одного боку, дає змогу врахувати нелінійність задач видобутку даних, але при цьому не вимагає великої кількості часу для виконання завдяки розділенню загальної вибірки на кластери.

Розглянемо застосування кусково-лінійного підходу у поєднанні із методом штрафних функцій для задачі із двома класами. Необхідно зазначити, що для оцінки ефективності методу необхідні дві вибірки з даними: тренувальна – яка використовується для побудови моделі геометричних перетворень та для якої обчислюється частотність і тестова – вибірка, дані якої не використовувались під час побудови моделі геометричних перетворень.

Матриця штрафних функцій в цьому випадку матиме вигляд (табл.3).

Таблиця 3

Матриця штрафних функцій

	Вектор розпізнано як клас 1	Вектор розпізнано як клас 2
Вектор належить до класу 1	$k_{11} = a_{11} \times CH_1$	$k_{12} = a_{12} \times CH_1$
Вектор належить до класу 2	$k_{21} = a_{21} \times CH_2$	$k_{22} = a_{22} \times CH_2$

Для збільшення точності розв’язання задачі класифікації пропонується поєднати використання методу штрафних функцій та дерево поділу на класи.

Використання дерева поділу на класи дає змогу об’єднувати в окремі кластери вектори даних, що мають схожі вхідні показники та аналізувати їх незалежно один від одного. Після того, як отримані значення штрафних балів по кожному з кластерів, вони підсумовуються. Такий підхід дозволяє суттєво підвищити загальну точність класифікації.

Розглянемо алгоритм поєднання методу штрафних функцій та дихотомії (дерева поділу на класи) для розв’язання задач класифікації детальніше.

Метод кусково-лінійної класифікації на основі моделі геометричних перетворень із використанням методу штрафних функцій:

1. Початково кількість штрафних балів мінімізуємо за алгоритмом, наведеним у пункті 1 (алгоритм класифікації із використанням методу штрафних функцій).
2. Після цього вибірки із спрогнозованими виходами, заміненими на ідентифікатори класів (”клас1”, ”клас2”), як тренувальну, так і тестову – ділимо на 2 кластери: вектори, розпізнані нейромережею як ”клас1” та вектори, розпізнані як “клас 2”.

3. Після кластеризації в тренувальні вибірки, отримані для кожного кластера, підставляємо реальні значення виходів.
4. Після цього окремо для кожного кластера підбираємо значення коефіцієнтів k11, k12, k21, k22, для яких сума штрафів при виконанні алгоритму класифікації із використанням методу штрафних функцій в цьому кластері буде мінімальною.
5. Якщо для певного кластера оптимальним є значення k11, k12, k21, k22, при яких нейромережа передбачає для тренувальної вибірки лише один клас (”клас 1” чи “клас 2”) – алгоритм кластеризації для цього кластера зупиняється, інакше – продовжуємо виконувати кластеризацію.
6. Після зупинки обробки кожного з кластерів аналізуємо тестові дані та обчислюємо суму штрафних балів по кожному з кластерів. Штрафні бали, отримані для кожного з кластерів, підсумовуються.

Отже, завдяки розбиттю даних на кластери ми пришвидшуємо виконання процесу класифікації та маємо змогу враховувати частотність елементів кожного з двох класів окремо в кожному з кластерів, збільшуючи таким чином точність класифікації.

3. Вдосконалення методу штрафних функцій шляхом застосування модифікованого алгоритму імітації відпалу металу

Пропонуємо розглянути поєднання методу кусково-лінійної класифікації на основі моделі геометричних перетворень та методу глобальної оптимізації – алгоритму імітації відпалу металу.

Пропонується поєднати метод штрафних функцій на базі 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, y – вихід, що задає належність до визначених класів.

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 таким чином, щоб результуюча сума штрафних балів була мінімальною.

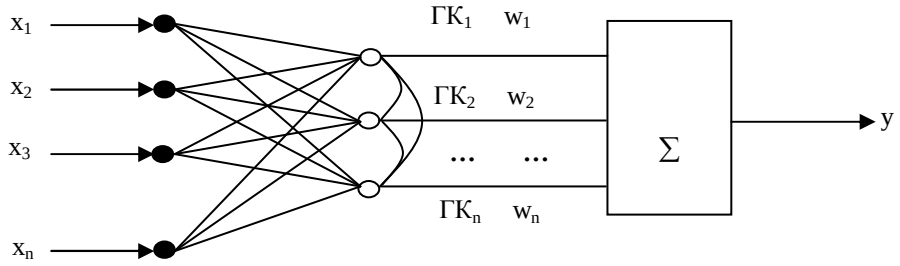

Рис.2. Структурна схема нейромережі на основі ФМТФ

Метод імітації відпалу є алгоритмічним аналогом фізичного процесу керованого охолодження і дозволяє 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 $T = T_{\max}$, 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 $w = w'$; в 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

Висновки

Розглянуті нейромережні методи класифікації на основі машини геометричних перетворень, орієнтовані на розв'язання задач інтелектуального аналізу даних. Описані методи враховують основні особливості завдань видобутку даних та дозволяють опрацьовувати великі обсяги даних за невеликий час. Крім того, вдосконалення нейромережних методів класифікації в завданнях видобутку даних шляхом застосування методу кусково-лінійної класифікації та методу імітації відпалу металу 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.: "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. – Вип.3. –Київ, 2006. – 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 польского И.Д. Рудинского. – М.: Финансы и статистика, 2004. – 344с.