

УДК 681.3.06

Медвідь С. О.

ВИКОРИСТАННЯ ЕВОЛЮЦІЙНИХ МЕТОДІВ ПРИ ПОБУДОВІ НЕЙРОННИХ МЕРЕЖ ДЛЯ ОБРОБКИ ЗНАНЬ

У статті досліджено особливості побудови еволюційного алгоритму для знаходження нейронної мережі, призначеної для однорівневої кластеризації текстів. Першим аспектом такого еволюційного алгоритму є визначення функції оцінки його проміжних розв'язків (нейронних мереж). Така функція виведена для двох випадків — еволюції за наявності контрольних текстів з відомим розбиттям на кластери й еволюції за наявності навчальних текстів, для яких таке розбиття невідоме. Іншим аспектом нейроеволюційного алгоритму є особливість знаходження оптимальної топології нейронної мережі, яка повинна бути одночасно і найпростішою, і здатною ефективно виконувати кластеризацію текстів. Ця проблема ефективно долається при застосуванні методу комплексифікації, використання якого доцільне при вирішенні поставленої задачі.

Вступ

У сучасному світі знання набувають все більшого пріоритету й поступово стають основ-

ною цінністю як окремих організацій, так і суспільства в цілому. Все частіше згадується поняття «інформаційне суспільство», знання стає одним із головних ресурсів людства і за важли-

віспо починає наближатися до інших ресурсів, таких як, скажімо, енергетичні.

Ці тенденції стали особливо актуальними останніми роками, і перш за все вони стосуються бізнес-організацій, де завжди є потреба в оцінюванні ресурсів компанії, наприклад визначення конкурентних переваг порівняно з іншими компаніями. Почали з'являтися компанії, половину або більше вартості яких становило дещо недотичне й абстрактне — те, що при детальному вивченні виявлялося знаннями. Саме тому питання управління таким важливим ресурсом, як знання, постало гостро і невідкладно. Адже знання, основними носіями яких є люди (що не належать організаціям), важко піддається управлінню.

Проблема управління знаннями

Сучасні системи управління знаннями застосовують технології, що оперують структурованими даними. За такого підходу зв'язки в інформаційних структурах визначені заздалегідь. Такий спосіб роботи зі знаннями дає змогу використовувати реляційні бази даних—технологію, що має під собою значний теоретичний фундамент. Реляції застосовуються для зберігання будь-яких даних, що є заздалегідь структурованими.

Однак існує безліч бізнес-процесів, що генерують та використовують неструктуровані знання. Організації накопичують таку інформацію у вигляді документів, текстових файлів, вихідних текстів програм, діаграм, поштової кореспонденції тощо. Згідно з результатами деяких досліджень, 80 відсотків корпоративної інформації є неструктурованою [1].

Отже, ефективне управління неструктурованими знаннями передбачає розробку методик створення знань, їх організації, зв'язування, координування потоків знань, ефективної роботи зі збереженими знаннями.

Одним із найбільш досліджених підходів до автоматизованої обробки неструктурованих знань є Data Mining, найголовнішими задачами якого є структуризація цих знань, визначення прихованих закономірностей, трендів і патернів у структурі знань, класифікація та кластеризація елементів знань.

Зосередимося на кластеризації текстів як на одному з аспектів структуризації знань, що передбачає ефективне управління ними. Розглянемо однорівневу кластеризацію (класи рівноправні між собою) як спрощений приклад структуризації текстової інформації.

Нейронні мережі та кластеризація текстів

Використання нейронних мереж для автоматичної обробки неструктурованої текстової інформації є досить ефективним методом. Перш за все дослідження продемонстрували більшу ефективність використання нейронних мереж порівняно зі статистичними методами для задач керування [2], оскільки замість методів визначення статистичних закономірностей нейронна мережа представляє модель поведінки.

Розглянемо типовий приклад використання нейронної мережі для задачі кластеризації текстів (рис. 1).

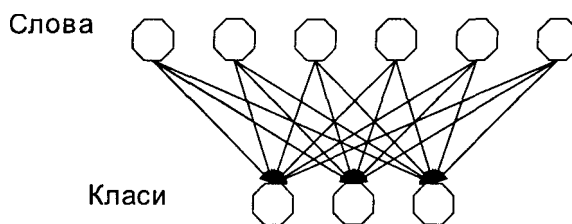


Рис. 1. Найпростіша нейронна мережа для задачі кластеризації текстів

Мережа, що використовується для цього, є одношаровою — у ній вхідні вузли відповідають словам (термінам) тексту, а вихідні — класам (категоріям) текстів. Кожен вхідний вузол має зв'язки з усіма вихідними вузлами.

Здатність такої структури кластеризувати тексти забезпечується алгоритмом зміни вагових коефіцієнтів зв'язків між вузлами під час навчання мережі. Після отримання вхідного сигналу він перетворюється мережею і передається на вузли класів, з яких спрацьовує той, у якого порогове значення менше за отриманий сигнал. Таким чином нейронна мережа задає стратегію кластеризації, а також запам'ятовує інформацію про зв'язок між термінами з тексту та категорією, до якої цей текст належить.

Отже, описана вище структура нейронної мережі здатна навчатися кореляції між інформацією про терміни з тексту та інформацією про те, до якого класу належить такий текст.

Процес побудови нейронної мережі для ефективного вирішення задачі кластеризації текстів зазвичай передбачає визначення топології мережі (наприклад такої, яка описана вище), виставлення початкових коефіцієнтів зв'язків між вузлами і подальше навчання мережі на наборі текстів-прикладів для формування прийнятних коефіцієнтів, з використанням яких проводитиметься кластеризація реальних текстів.

Алгоритм, що формує коефіцієнти, може продовжувати працювати і під час реальної роботи мережі для того, щоб була можливість корекції знань про відповідність між термінами з текстів та належністю текстів до певних кластерів або класів.

Отже, описаний вище процес дає змогу з часом отримати якіснішу класифікацію (кластеризацію) текстів. На початку ж якість класифікації є низькою, оскільки коефіцієнти підбираються емпірично і змінюються поступово.

Еволюційні методи побудови нейронної мережі

Оскільки описана вище задача побудови нейронної мережі є оптимізаційною, то для її вирішення можна застосувати еволюційні методи, зокрема генетичні алгоритми.

Для побудови генетичного алгоритму необхідно запрограмувати формальну функцію оцінки розв'язку, а також задати його кодування у набір даних (хромосому), над якими можна виконувати генетичні оператори.

Проблеми програмування генетичного алгоритму для побудови нейронних мереж досліджуються у рамках одного з напрямів еволюційного числення — нейроеволюції. У межах традиційних підходів до вирішення задачі нейроеволюції [3] топологія мережі обирається на початку процесу еволюції, і в кодуванні нейронної мережі як параметри наявні тільки значення вагових коефіцієнтів сполучень між її вузлами. Такі нейроеволюційні алгоритми з усталеною топологією оптимізують зв'язки між вузлами мережі, які й визначають її функціональність.

Проте зв'язки між вузлами — це лише один з аспектів структури нейронної мережі. Іншим аспектом є її топологія, яка також впливає на функціональність мережі. Останнім часом проводилися дослідження переваг та недоліків еволюціонування топологій нейронних мереж разом з еволюціонуванням вагових коефіцієнтів зв'язків [4]. Переконливі аргументи за використання такого підходу наведено у [5], серед яких — економія часу людей, що намагаються підібрати оптимальну топологію нейронної мережі.

Можна також зробити припущення, що зміна топології нейронної мережі, представленої на рис. 1, шляхом додавання нових її шарів, використання рекурентних зв'язків підвищить ефективність процесу кластеризації текстів. Адже рекурентні зв'язки дають змогу організувати пам'ять у нейронній мережі, а нові шари з'єднань ускладнюють структуру можливих

патернів даних, що спричинює до підвищення якості кластеризації.

Кодування. Проблема кодування нейронних мереж у вигляді хромосом для використання у генетичних алгоритмах є предметом активних досліджень багатьох робочих груп. Підходи до кодування поділяються на бінарне кодування, кодування графів, незхрещуване (англ. non-mating) та непряме кодування. Усі ці підходи проаналізовано у [6], а також запропоновано власний підхід (NEAT, від англ. "Neuro-Evolution through Augmenting Topologies"), що долає недоліки вищезгаданих методів і пропонує ефективний спосіб еволюціонування топологій мереж, який дає можливість позбутися проблеми відсікання надто складних та неефективних топологій.

Генетичні оператори. Як відомо, основними генетичними операторами є оператор схрещування та оператор мутації.

Слова

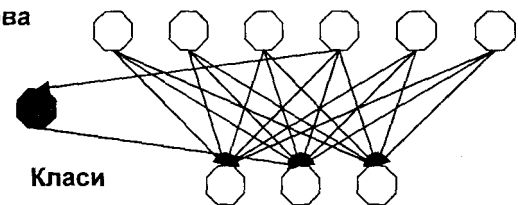


Рис. 2. Результат виконання оператора мутації, який додає вузол до нейронної мережі

У випадку еволюціонування нейронної мережі *оператор схрещування* дає змогу наслідувати у нащадків топології і значення вагових коефіцієнтів нейронних мереж, що схрещуються.

Оператор мутації для нейронної мережі означатиме додавання у топологію вузла, його видалення, додавання або видалення зв'язку між вузлами, а також зміну вагового коефіцієнта з'єднання між вузлами мережі.

Ймовірність додавання або видалення вузла мережі має бути значно меншою за ймовірність змін параметрів зв'язків між вузлами нейронної мережі, для того, щоб алгоритм зміг еволюціонувати розв'язки перш за все у напрямі оптимізації коефіцієнтів з'єднань, а потім у напрямі ускладнення структури мережі.

Метод NEAT. Основною проблемою схрещування топологій є визначення комплементарних вузлів двох нейронних мереж. Оскільки в процесі еволюції результат вдалої мутації може поширитися серед кількох індивідів, має бути спосіб коректного схрещування двох

батьківських хромосом таким чином, щоб однакові вузли не повторювалися у нащадках кілька разів.

Використання методу NEAT [6] гарантує коректне наслідування інформації про топологію під час використання оператора кросоверу. Основна ідея цього методу полягає у присвоєнні кожному новому гену в хромосомі так званої історичної мітки, яка копіюється без змін у хромосоми-нащадки після схрещування. Таким чином, якщо при подальших схрещуваннях зустрінуться дві хромосоми, частина генів яких матиме однакові історичні мітки, оператор кросоверу матиме можливість відстежити це і скопіювати лише один із генів у геном нащадка. Інший ген, відповідно, потрапить до іншого нащадка.

Особливості еволюціонування нейронних мереж для кластеризації текстів

Нейроеволюційний алгоритм для побудови нейронної мережі, що виконуватиме кластеризацію текстів, повинен мати кілька особливостей. Ці особливості стосуються функції оцінки, а також процесу побудови топології мережі.

Задати функцію оцінки можна двома шляхами: маючи набір контрольних текстів і для випадку кластеризації.

Щодо особливостей процесу побудови топології, то генетичний алгоритм має бути налаштований таким чином, щоб ускладнення топології порівняно з найпростішою (наведена на рис. 1) мало місце лише в тих випадках, коли це спричинює до покращення якості процесу кластеризації або класифікації текстів.

Функція оцінки за наявності контрольних текстів. За наявності набору контрольних текстів з інформацією про те, до яких класів вони належать, для розрахунку функції оцінки нейронна мережа визначає класи всіх контрольних текстів.

Оцінку ефективності можна визначити таким чином:

$$E_e = \frac{N_c}{N}, \quad (1)$$

де N_c – кількість правильно визначених класів, N – загальна кількість контрольних текстів.

До функції оцінки введемо також критерій, що оцінюватиме час, необхідний для класифі-

кації всіх текстів. Позначимо максимально прийнятний час як $t_{\max}$ і визначимо критерій оцінки за часом як функцію, що прийматиме значення від 0 до 1. Вона матиме вигляд:

$$E_t = \frac{|t - t_{\max}| - (t - t_{\max})}{2 \cdot t_{\max}}. \quad (2)$$

Функція оцінки буде сумою оцінки ефективності й оцінки за часом. Введемо також вагові коефіцієнти C_e і C_t , де $C_t = 1 - C_e$. Виходячи з (1) та (2), остаточна функція оцінки матиме вигляд:

$$E = C_e \frac{N_c}{N} + C_t \frac{|t - t_{\max}| - (t - t_{\max})}{2t_{\max}}. \quad (3)$$

Для коректної роботи алгоритму кількість контрольних текстів має бути достатньо великою, проте зі зростанням кількості таких текстів зменшуватиметься його ефективність. У подальших роботах планується визначити критерії раціоналізації кількості контрольних текстів.

Функція оцінки для процесу кластеризації текстів. Якщо контрольних текстів немає, функцію оцінки можна визначити, оцінюючи ефективність самого процесу кластеризації. Існують критерії визначення ефективності цього процесу для кластеризації текстів, використовуючи які, можна кількісно визначити ефективність застосування тієї чи іншої нейронної мережі для заданої задачі. Один із таких критеріїв визначений у [7].

Цей критерій базується на побудові векторного простору, в якому кожен з документів представлений як $x_1, x_2, \dots, x_d$, де кожен $x_i \in R^w$. Тут d – кількість документів, а w – кількість унікальних слів у них. Тобто такий векторний простір матиме стільки ж вимірів, скільки є унікальних термінів у словнику. Кластеризація колекції документів на підмножини $\pi_1, \pi_2, \dots, \pi_k$ формально записується як

$$\bigcup_{j=1}^k \pi_j = \{x_1, x_2, \dots, x_d\},$$

де $\pi_j \cap \pi_i = \emptyset, j \neq i$.

Оцінка якості (когерентності) кластерів основана на математичному визначенні подібності двох текстів. Ця схожість дорівнює косинусу кута між векторами, що представляють

тексти у просторі R^n . Тому когерентність кластера π_j визначається формулою:

$$\sum_{x_i \in \pi_j} x_i^T c_j, \quad (4)$$

де вектори x_i нормалізовані таким чином, що $\|x_i\|=1$, а c_j — це концептуальний вектор кластера π_j , що виражається як

$$c_j = \frac{\sum_{x_i \in \pi_j} x_i}{\left\| \sum_{x_i \in \pi_j} x_i \right\|}.$$

Якість кластеризації для фіксованої кількості кластерів k (тобто у випадку виконання класифікації текстів) дорівнює сумі когерентностей кластерів (4):

$$Q(\{\pi_j\}_{j=1}^k) = \sum_{j=1}^k \sum_{x_i \in \pi_j} x_i^T c_j. \quad (5)$$

Природно, що за невідомої кількості кластерів (типова задача кластеризації) це значення буде максимальним тоді, коли кожен текст буде віднесений до окремого кластера, тобто збігатиметься із концептуальним вектором цього кластера.

Тому у випадку задачі кластеризації слід задати функцію, що керуватиме двома критеріями: кількістю кластерів і відмінністю між їх концептуальними векторами:

$$\Omega(k, c), \quad (6)$$

де k — кількість кластерів, отриманих нейронною мережею, а c — множина їх концептуальних векторів.

Із (5) і (6) отримаємо функцію, що задає якість кластеризації як:

$$Q = \Omega(k, c) \cdot \sum_{j=1}^k \sum_{x_i \in \pi_j} x_i^T c_j. \quad (7)$$

Виходячи з (7), функцію оцінки для генетичного алгоритму, що будує нейронну мережу, яка виконує кластеризацію N текстів, можна записати як:

$$E_c = \Omega(k, c) \cdot \frac{\sum_{j=1}^k \sum_{x_i \in \pi_j} x_i^T c_j}{N}. \quad (8)$$

Додавши до (8) оцінку кластеризації за часом (2), а також використавши вагові коефіцієнти C_c і C_t , де $C_t = 1 - C_c$, отримаємо кінцеву функцію оцінки для використання у генетичному алгоритмі:

$$E = C_c \cdot \Omega(k, c) \cdot \frac{\sum_{j=1}^k \sum_{x_i \in \pi_j} x_i^T c_j}{N} + C_t \frac{|t - t_{\max}| - (t - t_{\max})}{2t_{\max}}.$$

Комплексифікація топологій нейронних мереж. Інша особливість нейроеволюції для побудови нейронних мереж, що виконуватимуть кластеризацію текстів, полягає у використанні такого процесу еволюції, в якому результат буде оптимізований ще й за складністю топології, тобто менш складні нейронні мережі повинні мати більшу перевагу.

Використання простіших топологій має сенс тому, що, по-перше, простіші мережі виконуватимуть кластеризацію швидше. По-друге, в процесі еволюції генетичний алгоритм швидше оперуватиме менш складними структурами даних.

Цю задачу ефективно вирішує метод комплексифікації топологій, описаний у [8], що полягає у генерації початкової популяції розв'язків таким чином, що вона представлятиме найпростіші за топологією нейронні мережі. Під час подальшої роботи алгоритму ймовірність виникнення мутації, що ускладнить топологію мережі, значно менша, ніж ймовірність зміни вагових коефіцієнтів зв'язків між вузлами. До таких мутацій належить додавання нових вузлів до мережі, а також додавання нових з'єднань між вузлами (див. рис. 2).

При побудові нейронних мереж, що виконують кластеризацію текстів, на першому етапі генетичного алгоритму генеруватимемо мережі з єдиною топологією, що являє собою найпростіший випадок, зображений на рис. 1. Відмінності між цими мережами полягатимуть у різних вагових коефіцієнтах зв'язків між вузлами.

Висновки та подальша робота

У статті досліджено особливості побудови еволюційного алгоритму для знаходження нейронної мережі, призначеної для однорівневої кластеризації текстів. Першим аспектом такого еволюційного алгоритму є визначення функції оцінки його проміжних розв'язків (нейронних мереж). Така функція виведена для двох випадків - еволюції за наявності контрольних текстів з відомим розбиттям на кластери і еволюції за наявності навчальних текстів, для яких таке розбиття невідоме. Іншим аспектом нейроеволюційного алгоритму є особливість зна-

ходження оптимальної топології нейронної мережі, яка має бути одночасно і найпростішою, і здатною ефективно виконувати кластеризацію текстів. Ця проблема ефективно долається при застосуванні методу комплексифікації, використання якого доцільне при вирішенні поставленої задачі.

Подальші дослідження планується провести у напрямі аналізу функції керування оптимальною кількістю кластерів, критеріїв оптимізації кількості контрольних текстів, способів ієрархічної кластеризації текстів, а також перспектив використання рекурентних нейронних мереж для задачі кластеризації.

1. Satydas A., Harigopal, Cassaigne (2001). Knowledge Management Tutorial // IEEE Transactions on systems, Man and Cybernetics.- Part C: Applications and Reviews.— Vol. 31.- № 4 (November).
2. Kaelbling L. P., Littman M., Moore A. W. Reinforcement learning: A survey // Journal Of Artificial Intelligence.— 1996.- 4:237-285.
3. Gomez F., Miikkulainen R. (1999). Solving non-Markovian control tasks with neuroevolution. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1356—1361, Morgan Kaufmann, San Francisco, California.
4. Gruau F., Whitley D., Pyeatt L. (1996). A comparison between cellular encoding and direct encoding for genetic neural networks. Genetic Programming 1996: Proceedings of the First Annual Conference, 81-89, MIT Press, Cambridge, Massachusetts.
5. Yao X. (1999). Evolving artificial neural networks. Proceedings of the IEEE, 87(9):1423-1447.
6. Stanley K. O., Miikkulainen R. (2002). Evolving Neural Networks through Augmenting Topologies // Evolutionary Computation.- Vol. 10, 2, MIT Press, Cambridge, Massachusetts.
7. Dhillon I. S., Fan J., Guan Y. Efficient clustering of very large document collections. Data Mining for Scientific and Engineering Applications, 357—381. Kluwer Academic Publishers, 2001.
8. Stanley K. O., Miikkulainen R. (2004). Competitive Coevolution through Evolutionary Complexification // Journal of Artificial Intelligence Research 21, 63—100.

S. O. Medvid

USING EVOLUTIONARY METHODS FOR BUILDING KNOWLEDGE-PROCESSING NEURAL NETWORKS

The article researches peculiarities of evolutionary algorithm used for building neural network that will do one-level text clustering. The first aspect of such an algorithm is defining fitness function for its intermediate solutions (neural networks). This function is created for two cases: evolution using control texts with known information about their classes and evolution using training texts with no information about optimal clustering. Another aspect of neuroevolution algorithm is a way of finding optimal neural network topology: it should be simple on one hand and perform effective text clustering on another. This problem can be solved effectively using method of complexification, which is reasonable for solving the task set.