УДК 004.8

ОБРОБКА ТЕКСТОВОЇ ІНФОРМАЦІЇ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ З ВИКОРИСТАННЯМ МЕТОДІВ ОБЧИСЛЮВАЛЬНОГО ІНТЕЛЕКТУ

В.В. Волкова, О.В. Шубкіна

Харківський національний університет радіоелектроніки,
val_volkova@ukr.net, olga.shubkina@gmail.com

В роботі введено спеціалізовану нейронну мережу, що є розвитком радіально-базисних структур, та забезпечує вирішення завдання класифікації іменованих сутностей для побудови семантичних анотацій текстових документів заданої предметної області. Запропоновано модель нейро-фаззі системи кластеризації політематичних текстових документів на основі адаптивної нечіткої нейронної мережі, що самоорганізується.
Ключові слова: класифікація, нечітка кластеризація, іменовані сутності, семантична анотація

The dedicated neural network that is a development of radial-basis structures was proposed. This neural network provides the issue decision of a named entities classification for building of semantic annotations for text documents of specified data domain. The model of neuro-fuzzy system for multi-topic text clustering was proposed. This model is based on an adaptive fuzzy self-organizing neural network.
Keywords: classification, fuzzy clustering, named entities, semantic annotation

В работе введена специализированная нейронная сеть, являющаяся развитием радиально-базисных структур, и обеспечивающая решение задачи классификации именованных сущностей для построения семантических аннотаций текстовых документов заданной предметной области. Предложена модель нейро-фаззи системы кластеризации политематических текстовых документов на основе адаптивной нечеткой самоорганизующейся нейронной сети.
Ключевые слова: классификация, нечеткая кластеризация, именованные сущности, семантическая аннотация

Вступ. В сучасних інформаційних технологіях все більше зростає роль процедур видобування знань. Це викликано постійним збільшенням кількості структурованої чи неструктурованої інформації в різногалузевих організаціях чи мережі Інтернет. Ресурси знань відрізняються залежно від галузей індустрії та застосування, але, як правило, представлені в текстовому вигляді. Текстові документи (ТД) є ресурсом знань, у якому міститься близько 85% всієї інформації. Таким чином, зростає роль видобування знань із текстових джерел, накопичених на різних етапах розвитку організації. Цей процес є основним завданням інтелектуального аналізу текстової інформації.

На сучасному етапі у системах інтелектуального аналізу текстової інформації (text mining) широкий розвиток отримала розробка систем і методів семантичного анотування текстових документів. Основна ідея полягає в створенні опису ТД у машинно-зрозумілій формі на основі онтології предметної галузі, який далі будуть використовувати інтелектуальні агенти [1]. Існує набір стандартних рішень, які розроблені для опису метаданих і формування семантичних анотацій (СА), як наприклад, стандарт Dublin Core, проекти FOAF, SKOS. Однак набір заданих тегів для опису ТД не відображає інформацію, що є актуальною для поточної онтології предметної області, а

найчастіше несе лише загальні відомості (наприклад, creator – розробник, language – мова ресурсу). Варто відзначити також, що створення СА вручну займає досить багато часу та грошових витрат. Це привело до розробки методів напівавтоматичної побудови СА, які у свою чергу мають низку недоліків, наприклад, використання шаблонів заповнення або апріорі заданих правил. Тому актуальним є завдання розробки методів автоматичної побудови СА.

Одним з важливих завдань при видобуванні знань з різних джерел інформації є їх впорядкування та структуризація. Вирішення таких завдань можливе за допомогою методів класифікації та кластеризації, що дозволяють розбивати множину ресурсів знань (текстових документів) на категорії залежно від їх характеристик. Слід зазначити, що більшість текстових документів одночасно відносяться до декількох тематик, що ускладнює процес їх класифікації чи кластеризації.

На сьогоднішній день існує досить велика кількість методів кластеризації текстових документів, але більшість з них не враховують повною мірою наявності кластерів, що перетинаються при класифікації без учителя, тобто такі ситуації, коли один і той же документ може одночасно належати до декількох категорій. Також такі методи обмежені по засобах обробки вхідних даних, тобто не можуть послідовно кластеризувати вхідні дані. Це є істотним їх недоліком. Таким чином, розробка засобів, що дозволяють вирішити ці проблеми, є важливим завданням.

1. Обробка текстової інформації за допомогою радіально-базисної нейронної мережі

Процес семантичного анотування можна розглядати як проблему класифікації, тому його автоматизація може бути досягнута шляхом застосування різних методів інтелектуальної обробки даних і відповідно різних підходів до навчання. Одним зі способів формування семантичних анотацій для ТД є ідентифікація іменованих сутностей (ІмС) і віднесення їх до заданого класу онтології предметної області [2]. Прикладом ІмС є персони, організації, географічні об’єкти та інші об’єкти, що позначаються в тексті з використанням власних імен.

Формально СА можна представити таким чином. Для заданої онтології предметної області Ont набір концептів визначається як $ConceptSet = \{c_i | c_i \in Ont\}$, де c_i – i-й концепт. Тоді СА (тобто розмітка речення, яке містить ІмС) для даної онтології буде визначена як

$$LabelSet = \{l_i | \exists c_j \in ConceptSet \wedge (l_i = "B\_" + c_j \vee l_i = "I\_" + c_j)\} \cup \{"O"\}, \quad (1)$$

у якому LabelSet складається з «*B_*» чи «*I_*» – префіксів концептів.

Таким чином, метод для побудови СА, що базується на використанні такої послідовності, можна записати, знаючи простір входів $x = \{x_i | x_i = w_1, w_2 \ldots w_t \ldots w_T, w_t \in WordSet\}$, в якому x_i – речення, що складається з

деяких слів, та простір виходів $y = \{y_i | y_i = l_1, l_2 \ldots l_t \ldots l_T, l_t \in LabelSet\}$, де y_i – розмічене на основі класифікації ІмС речення.

Нехай задана деяка множина навчальних прикладів $\{(x_i, y_i)\}_{i=1}^n$, де $x_i \in x, y_i \in y$, в такому випадку функція відображення $f : x \to y$ настроюється, використовуючи набір сформованих прикладів. Для нерозміченого речення *x*, $f(x)$ – деяке нелінійне відображення, яке й буде СА для цього речення. Відзначимо, що в загальному випадку вектор ознак для подання вхідних навчальних даних може бути отриманий на основі контекстної інформації, що включає як лінгвістичні, так і статистичні змінні.

В наш час у сучасних розробках з використанням інтелектуальних технологій добре зарекомендували себе штучні нейронні мережі (ШНМ) [3]. Тому далі розглянемо метод побудови класифікатора на основі радіально-базисної нейронної мережі [4]. Вхідний шар такої мережі – це сенсори, які зв'язують ШНМ із навколишнім середовищем. Прихований шар, утворений нейронами *Ф*, здійснює нелінійне перетворення вхідного простору R^n в прихований простір R^h, як правило, високої розмірності ($h >> n$). І, нарешті, вихідний шар, утворений адаптивними лінійними асоціаторами формує відгук мережі $y = (y_1, y_2, \ldots, y_m)^T$ на вхідний сигнал мережі $x = (x_1, x_2, \ldots, x_n)^T$.

Введемо до розгляду *m* помилок навчання

$$e_j(k) = d_j(k) - y_j(k) = d_j(k) - sign\ u_j(k) \quad (2)$$

та *m* локальних критеріїв [5]

$$E_j(k) = e_j(k)u_j(k) = d_j(k)u_j(k) - |u_j(k)| = (d_j(k) - sign\ w_j^T\varphi(k))w_j^T\varphi(k), \quad (3)$$

де $d_j(k)$ – навчальний сигнал, який приймає значення 1, якщо сигнал знаходиться в *j*-му стані, та -1 – в протилежному випадку, кодує множину заданих класів предметної області; $w_j = (w_{j1}, w_{j2}, \ldots w_{jh})^T$ – вектор синаптичних ваг, що підлягає визначенню; $\varphi(k) = (\varphi_1(k), \varphi_2(k), \ldots, \varphi_h(k))^T$ – вектор вихідних сигналів радіально-базисного шару. У навчальній вибірці обов'язково повинні бути присутні всі можливі стани з аналізованого масиву даних $\{x(k)\}$, в протилежному випадку деякі з «не пред'явлених» образів мережа не розпізнає.

Для настроювання синаптичних ваг будемо використовувати стандартну градієнтну процедуру

$$w_{ji}(k+1) = w_{ji}(k) - \eta(k)\partial E_j(k)/\partial w_{ji}(k),\ j = 0,1,\ldots,m;\ i = 1,2,\ldots,h\ , \quad (4)$$

де $\eta(k)$ – параметр кроку пошуку, яка з урахуванням (3) приймає значення

$$w_{ji}(k+1) = w_j(k) + \eta(k)e_j(k)\varphi(k) = w_j(k) + \eta(k)(d_j(k) - sign\, w_j^T(k)\varphi(k))\varphi(k),\ j = 0,1,\ldots,m. \quad (5)$$

Якщо ввести загальний критерій класифікації

$$E(k) = \sum_{j=1}^m E_j(k) = -\sum_{j=1}^m e_j(k)u_j(k), \quad (6)$$

можна записати алгоритм одночасного навчання всіх синаптичних ваг у формі

$$W(k+1) = W(k) + \eta(k)(d(k) - sign\ W(k)\varphi(k))\varphi^T(k), \quad (7)$$

де $sign(u_1(k),u_2(k),\ldots,u_m(k))^T = (sign\ u_1(k), sign\ u_2(k),\ldots, sign\ u_m(k))^T$,

$d(k)=(d_1(k),d_2(k),\ldots,d_m(k))^T$, $W(k+1)=(w_1^T(k),w_2^T(k)\ldots w_m^T(k))^T$ – $(m\times h)$ – матриця синаптичних ваг, що підлягають настроюванню та визначенню.

З метою підвищення швидкодії алгоритм навчання (7) можна модифікувати до такого виду

$$\begin{cases} w_j(k+1)=w_j(k)+\dfrac{d_j(k)-sign\ w_j^T(k)\varphi(k)}{\eta(k)}\varphi(k),\ j=1,2,\ldots,m, \\ \eta(k)=\eta(k-1)+\|\varphi(k)\|^2,\ \eta(0)=1, \end{cases} \tag{8}$$

який є розширенням процедури ідентифікації Гудвіна-Ремеджа-Кейнеса на завдання навчання ШНМ [4].

2. Обробка політематичної текстової інформації на основі нейро-фаззі підходу

Більшість джерел текстової інформації за своєю природою є політематичними, тобто одночасно відносяться до декількох тематик. Це ускладнює процес їх обробки та видобування з них знань, зокрема виконання процедур кластеризації та класифікації. Метою класифікації політематичних ТД є віднесення одного й того ж ТД до більш, ніж одної тематики. Мета кластеризації політематичних ТД – кластеризувати ТД таким чином, щоб кожний документ відносився до більш, ніж одного кластеру [6]. Процедура кластеризації є методом навчання без вчителя, тобто категорії текстових документів апріорі не задані.

Основні методи, що дозволяють виконувати процедури кластеризації політематичних документів, мають ряд недоліків, пов'язаних з обчислювальною складністю та неможливістю знаходження в процесі своєї роботи нових класів/кластерів. Тому пропонується модель системи кластеризації текстових документів, в основі якої лежать нейро-фаззі принципи.

Система складається з двох паралельно працюючих адаптивних нечітких нейронних мереж, що самоорганізуються, при цьому мережі обмінюються між собою інформацією та настроюються за допомогою рекурентних ймовірнісного та можливісного алгоритмів самонавчання, що є узагальненнями правил навчання Т.Кохонена [7-9].

Використовуючи позначення, введені в [10], опишемо процес навчання та функціонування запропонованої моделі.

Модель адаптивної нечіткої нейронної мережі містить єдиний прихований шар нейронів N_i, $i=1.2.\ldots,p$. На рецепторний шар мережі надходять образи документів, що підлягають кластеризації, у вигляді $(n\times 1)$-векторів ознак $x(t)=(x_1(t),x_2(t),...,x_n(t))^T$, де $t=1,2,...,V$ має зміст номеру образу в навчальній вибірці, або поточного дискретного часу. При цьому самі вектори ознак $x(t)$

формуються на основі усічених гістограм частот зустрічаємості окремих слів у політематичних текстових документах, що піддаються обробці.

Синаптичні ваги $m_{ij}, i = 1,2,..., p; j = 1,2,..., n$ визначають координати центроїдів p кластерів $m_i(t)$, що взаємно перетинаються, а виходом мережі є $(p \times 1)$-вектор $u(t) = (u_1(t), u_2(t),..., u_p(t))^T$, що визначає рівень належності образу $x(t)$ до кожного з p кластерів, що формуються, та обчислюваний нейронами N_i. По латеральним зв'язкам нейрони обмінюються координатами $m_i(t)$, необхідними для обчислення належностей $u_i(t)$.

В основі самонавчання лежить рекурентний ймовірнісний метод кластеризації, що базується на оптимізації цільової функції виду [11]:

$$E(u_i, m_i) = \sum_{t=1}^{V} \sum_{i=1}^{p} u_i^{\beta}(t) \left\| x(t) - m_i \right\|^2 \qquad (9)$$

при обмеженнях $\sum_{i=1}^{p} u_i(t) = 1, \quad t = 1,2,...,V, 0 \le \sum_{t=1}^{V} u_i(t) \le V, \quad i = 1,2,..., p,$ де $u_i(t) \in [0,1]$; V – кількість образів документів, що обробляються; β – додатній параметр («фаззіфікатор»), який визначає нечітку границю між кластерами, та впливає на рівень нечіткості у кінцевому розбитті даних по кластерах [10].

Адаптивний ймовірнісний метод самонавчання нечіткої нейронної мережі, що самоорганізується, може бути записаний у формі процедури стохастичної апроксимації [9], що є простою в обчисленні та дає можливість послідовної обробки даних і нечітке розбиття на кластери.

Як альтернатива ймовірнісному методу в [8] запропоновано рекурентний можливісний метод навчання адаптивної нечіткої нейронної мережі, що самоорганізується. В основі цього правила самонавчання лежить оптимізація локальної цільової можливісної функції [13]:

$$E_t(u_i(t), m_i(t)) = \sum_{i=1}^{p} u_i^{\beta}(t) d^2(x(t), m_i) + \sum_{i=1}^{p} \mu_i (1 - u_i(t))^{\beta}, \qquad (10)$$

де $u_i(t) \in [0,1], \beta$ – фаззіфікатор, $d^2(x(t), m_i) = \left\| x(t) - m^i \right\|^2$ – квадрат евклідової відстані між образом та центроїдом, $\mu_i > 0$ – скалярний параметр, що визначає відстань, на якій рівень належності приймає значення 0,5, тобто якщо $d^2(x(t), m_i) = \mu_i$, то $u_i(t) = 0{,}5$.

Варто відзначити, що важливою властивістю цього метода самонавчання є умова $\sum_{i=1}^{p} u_i^{POS}(t) \ne 1,$ яка на відміну від процедур ймовірнісної нечіткої кластеризації дозволяє знаходити нові кластери в процесі навчання нейронної мережі, а також коректно оцінювати викиди та накопичення вибірки в реальному часі, по мірі надходження.

Паралельне застосування адаптивних ймовірнісного та можливісного методів веде до об'єднаної процедури (для $\beta = 2$), що є методом самонавчання нейро-нечіткої системи.

$$\begin{cases} m_i^{PR}(t+1) = m_i^{POS}(t) + \alpha(t)\left(u_i^{POS}(t)\right)^2 \times \left(x(t+1) - m_i^{POS}(t)\right), \\ u_i^{PR}(t+1) = \left\|x(t+1) - m_i^{PR}(t+1)\right\|^{-2} \times \sum_{l=1}^{p} \left\|x(t+1) - m_l^{PR}(t+1)\right\|^2, \\ m_i^{POS}(t+1) = m_i^{PR}(t+1) + \alpha(t)\left(u_i^{POS}(t)\right)^2 \times \left(x(t+1) - m_i^{PR}(t+1)\right), \\ u_i^{POS}(t+1) = \mu_i(t) \times \left(\mu_i(t) + \left\|x(t+1) - m_i^{POS}(t+1)\right\|^2 \right)^{-1}, \\ \mu_i(t+1) = \left(\sum_{p=1}^{t+1} \left(u_i^{POS}(p)\right)^2 \left\|x(p) - m_i^{POS}(t+1)\right\|^2 \right) \times \sum_{p=1}^{t+1} \left(u_i^{POS}(p)\right)^{-2}, \end{cases} \quad (11)$$

де $\alpha(t)$ – параметр кроку пошуку, що визначає швидкість навчання.

Ознакою коректного відновлення прототипів за допомогою методу (11) є виконання нерівності $\sum_{l=1}^{p} d^2\left(m_l^{PR}(t), m_l^{POS}(t)\right) \le \varepsilon,$ де параметр ε визначає точність кластеризації.

Вводячи деяке порогове значення ξ ($\xi \approx 0{,}1 \div 0{,}2$) та контролюючи виконання умови $\sum_{i=1}^{p} u_i^{POS}(t) \le \xi,$ можна говорити про появу нового кластера. Варто відзначити, що якість кластеризації суттєвим чином залежить від вибраного простору ознак та об'єму даних.

3. Експериментальні дослідження

Робота запропонованого алгоритму оцінювалася на вибірці колекції статей Reuters, що є однією з найбільш часто використовуваних тестових вибірок в text mining. Для експерименту було обрано 460 документів з категорій Arts і Business. Експеримент показав, що запропонований алгоритм класифікації дає більш точні результати (у середньому 4-7%) у порівнянні з іншими методами класифікації, використовуваними для завдання класифікації ІмС [2]. Використання запропонованої нейро-фаззі системи забезпечує більшу точність кластеризації, порівняно з чіткими методами кластеризації, та дозволяє виконувати обробку даних в режимі реального часу [7].

Висновки. Введено спеціалізовану нейронну мережу, що є розвитком радіально-базисних структур, та забезпечує вирішення завдання класифікації іменованих сутностей для побудови семантичних анотацій текстових документів заданої предметної області.

Запропоновано модель нейро-фаззі системи кластеризації політематичних текстових документів на основі адаптивної нечіткої нейронної мережі, що самоорганізується, яка відрізняються від існуючих моделей своєю архітектурою та методами навчання. Запропонована модель дозволяє обробляти дані в режимі реального часу та враховувати нечіткі кластери при обробці політематичних текстових документів, що послідовно до неї надходять.

Експериментальні дослідження підтвердили ефективність запропонованих моделей.

Література

1. Шубкина О.В. Методы разметки последовательностей для создания семантических аннотаций информационных ресурсов // Материалы междунар. науч.-практ. конф.: «Информационные технологии и информационная безопасность в науке, технике и образовании «ИНФОТЕХ-2009». – Севастополь: Изд-во СЕВНТУ, 2009. – С. 197–200.
2. Шубкина О.В. Использование радиально-базисной нейронной сети для классификации именованных сущностей // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта: Материалы международной научной конференции. Том 2. – Херсон: ХНТУ, 2010. – с. 506-509.
3. Rojas R. Neural Networks. A Systematic Introduction. – Berlin: Springer – Verlag, 1996. – 502 p.
4. Бодянский Е.В., Кучеренко Е.И., Михалев А.И. Нейро-фаззи сети Петри в задачах моделирования сложных систем. – Днепропетровск: Системные технологии, 2005. – 311 с.
5. Shynk J.J. Performance surfaces of a single-layer perceptron // IEEE Trans. on Neural Networks. – 1990. – 1. – P. 268–274.
6. Бодянский Е.В., Волкова В.В., Егоров А.С. Кластеризация массивов текстовых документов на основе адаптивной нечеткой самоорганизующейся нейронной сети // Радиоэлектроника. Информатика. Управление. – 2009. – Вып. 1(20). – С.113–117.
7. Бодянский Е.В., Волкова В.В., Колчигин Б.В. Самообучающаяся нейро-фаззи система для адаптивной кластеризации текстовых документов. – Бионика интеллекта. – 2008. – Вып. 1(70). – С.34–38.
8. Kohonen T. Self-Organizing Maps. Berlin: Springer-Verlag. – 1995. – 362 p.
9. Kohonen T., Kaski S., Lagus K., Salojarvi J., Honkela J., Paatero V., Saarela A. Self organization of a massive document collection // IEEE Trans. on Neural Networks. – 2000. – 11. – P. 574–585.
10. Hoppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis, and Image Recognition // Chichester: John Willey&Sons Ltd. – 1999. – 289 p.
11. Krishnapuram R., Keller J. A possibilistic approach to clustering // IEEE Trans. on Fuzzy Systems. – 1993. – 1. – P. 98–110.