УДК 004.912:004.8

О. Шубкіна, І. Плісс, Є. Бодянський
Харківський національний університет радіоелектроніки

ВИКОРИСТАННЯ КОНКУРЕНТНОЇ ЙМОВІРНІСНОЇ МЕРЕЖІ У ЗАДАЧАХ ОБРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Запропоновано конкурентну ймовірнісну нейронну мережу, в якій параметр ширини активаційної функції настроюється автоматично. Запропонована штучна нейронна мережа використовується для обробки текстової інформації з метою створення семантичних анотацій текстових документів.

Ключові слова: конкурентна ймовірнісна нейронна мережа, обробка текстової інформації.

The competitive probabilistic neural network with activation function width automatic tuning is proposed. This neural network is used for text processing notably for semantic annotations of text documents creating.

Keywords: competitive probabilistic neural network, text processing.

Вступ

Сьогодні значно поширилися методи та технології інтелектуальної обробки текстової інформації. Одним з таких підходів є семантичне анотування, що полягає у присвоєнні семантичних тегів текстовим документам. Як відомо, раніше широко використовувалися різноманітні мови розмітки, низка мікроформатів, які задають певний набір тегів. У галузі інформаційного пошуку метадані найчастіше розглядають як набір елементів, що описують основні властивості текстового документа (включаючи тематику), які можуть також містити елементи вже наявних схем.

Проте разом із розвитком онтологій, які використовуються як бази знань предметної галузі, підхід семантичного анотування стає все більш актуальним [1]. Найвідоміші роботи у цій галузі досліджень передбачають отримання семантичних тегів або метаданих як проекцію текстового документа на онтологію предметної галузі [2]. Ряд робіт включає наповнення онтології новими концептами і відповідно виділення нових тегів. Отже, стає можливим видобування знань із текстової інформації для їх подальшого використання як машинно-зрозумілого опису ресурсу – семантичної анотації. Для автоматичного отримання семантичних анотацій залежно від обраної стратегії використовуються підходи, засновані на класифікації або кластеризації текстових даних.

Постановка задачі дослідження

Семантична анотація – розмітка або набір метаданих текстового документа, що розглядається, – визначається на основі заданої онтології предметної галузі *Ont* як $LabelSet = \{l_i | \exists c_j \in ConceptSet \wedge l_i = c_j\}$, де *LabelSet* – унікальна множина для кожного текстового документа, що складається з концептів (класів) онтології предметної галузі, які отримані шляхом проекції множини текстових об'єктів цього документа на задану онтологію з використанням методів теорії штучних нейронних мереж (ШНМ). Варто зазначити, що у межах такого підходу не виключається належність одного текстового об'єкта до кількох класів, а саме, на виході класифікатора повинна визначатися ймовірність належності вхідного образу до кожного заданого класу онтології *Ont*. При цьому для заданої онтології предметної галузі *Ont* набір концептів (класів) визначається як $ConceptSet = (c(1), c(2), \mathbf{K}, c(i), \mathbf{K} c(N_1))$, де ***c(i)*** – *i*-й концепт із *Ont*. Для будь- якого текстового корпуса набір текстових об'єктів, отриманих на етапі передобробки, можна представити як $ObjectSet = (x(1), x(2), \mathbf{K}, x(j), \mathbf{K}, x(N_2))$, де ***x(k)*** – *k*-й текстовий об'єкт, представлений у вигляді деякого набору релевантних ознак у векторній формі, N_1 и N_2 – кількість концептів (класів) онтології та потужність вихідної вибірки текстових об'єктів відповідно.

Такий підхід забезпечує можливість оцінювання віднесення текстового об'єкта $x(k)$ до різних концептів онтології, яка може бути задана ієрархією або таксономією. Отриману розмітку згодом можна подати у вигляді триплетів RDF [3] або OWL [4] для подальшого використання різними програмними засобами.

Конкурентна ймовірнісна нейронна мережа

Задача, що розглядається, може бути вирішена на основі методів байєсівської класифікації за допомогою ймовірнісних нейронних мереж (PNN), введених Д.Ф. Шпехтом [5].

Основою байєсівської класифікації є прийняття рішення про належність кожного вхідного образа-вектора до найімовірнішого класу з тих, яким міг би належати цей образ. Таке рішення вимагає оцінювання функції щільності розподілу ймовірностей для кожного класу, яка відновлюється на основі аналізу даних з навчальної вибірки. Для відновлення таких функцій значне поширення отримали оцінки Парзена (Надарая – Ватсона), що використовують вагові функції (потенційні функції, ядерні функції), які мають центри в точках, що відповідають образам із відомою класифікацією з навчальної вибірки.

Незважаючи на те, що баєсівські методи класифікації відомі давно, їх паралельна нейромережева реалізація дала змогу забезпечити вищу швидкість процесам обробки інформації, пов'язаним із розпізнаванням образів, класифікацією, діагностикою тощо.

Необхідно зазначити, що результат, який отримується за допомогою стандартної ймовірнісної мережі, дає змогу віднести образ $x(k)$ до одного єдиного класу з найщільнішим розподілом в області цього образу. Ключовою властивістю нашого підходу є визначення значення ймовірнісної приналежності $probab\_value$ текстового об'єкта $x(k)$ до кожного класа (концепту) онтології, яке вводиться як одна з властивостей логічного опису триплетів метаданих.

У [6] автори запропонували модифіковану ймовірнісну нейронну мережу (MPNN) для розв'язання поставленої задачі. Як було зазначено, значення параметра ширини активаційний функції для нормованих входів обирається довільно в інтервалі від нуля до одиниці [7]. Разом з тим, слід зазначити, що простого формального рішення, що дає змогу отримати значення цього параметра, сьогодні не існує. У цій роботі знаходження параметра ширини активаційної функції розглянуто докладніше.

Таке вирішення може бути отримано за допомогою запропонованої конкурентної нейронної мережі (CPNN), архітектуру якої наведено на рис.1. CPNN є гібридом стандартної PNN, узагальненої регресійної нейронної мережі (GRNN), також введеної Шпехтом [8], та самоорганізованої мапи Кохонена (SOM) [9]. CPNN містить чотири шари обробки інформації: перший прихований, який іменується шаром образів, другий прихований шар локальних суматорів, третій прихований шар, що містить єдиний загальний суматор, і, нарешті, вихідний шар дільників. Характерною особливістю CPNN є те, що нейрони шару образів об'єднані в групи класів, між якими організовано латеральні зв'язки, які використовуються для настроювання ширини активаційної функції.

Вихідною інформацією для синтезу мережі є навчальна вибірка, сформована «пакетом» n-вимірних векторів $x(1), x(2), \mathbf{K}, x(N)$ з відомою класифікацією. Передбачається також, що всі вхідні вектори пронормовано так, що $\|x(j)\| = 1,\ j = 1,2,\mathbf{K}\, N,$ а самі образи (без втрати загальності) можуть належати, наприклад, одному з трьох класів A, B або C.

Кількість нейронів у шарі образів приймається рівним N (по одному нейрону на кожен навчальний образ), а їх параметри (центри активаційних функцій) визначаються на основі компонент вхідних векторів так, що

$$w_{ji} = x_i(j),\ j = 1,2,\mathbf{K}\, N;\ i = 1,2,\mathbf{K}\, n, \tag{1}$$

або у векторному вигляді

$$w_j = x(j),\ j = (x_1(j), x_2(j), \mathbf{K}, x_n(j))^T. \tag{2}$$

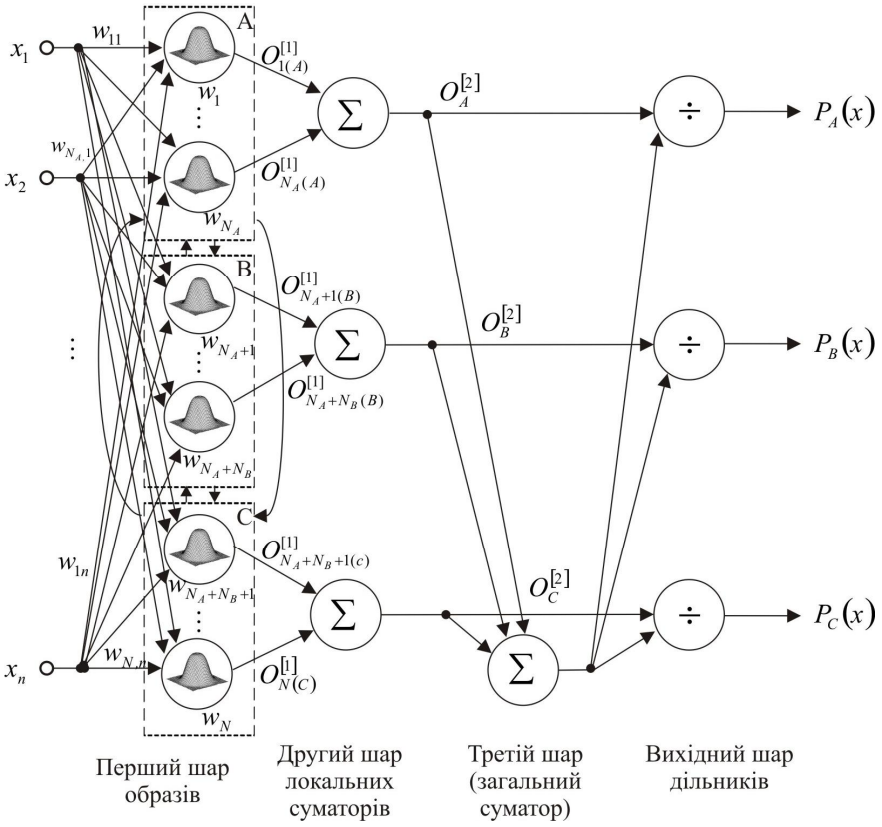

Рис. 1. Конкурентна ймовірнісна нейронна мережа

Отже, у цій мережі реалізується навчання, засноване на пам'яті [9], за принципом «нейрони в точках даних» [11], що робить його вкрай простим та практично миттєвим.

Кожен з нейронів шару образів обчислює зважену суму компонент вхідних сигналів і перетворює її за допомогою нелінійної активаційної функції виду

$$\varphi_j(x) = \exp\left(-\frac{\left\|x - w_j\right\|^2}{2\sigma^2}\right), \tag{3}$$

при цьому параметр ширини σ вибирається зазвичай із емпіричних міркувань [12]. Необхідно зазначити, що занадто мале значення σ призводить до виникнення «дір» у просторі параметрів та погіршення узагальнюючих властивостей мережі, а занадто велике значення σ – до розмивання та перекриття класів, що збільшує ймовірність виникнення помилок класифікації. У зв'язку з цим завдання обґрунтованого вибору параметра ширини σ є актуальним.

Розглянемо спочатку задачу класифікації на два класи A і B. Припустимо, що мережа вже навчена, класи сформовано, але σ не обчислено.

Позначимо w_A^* та w_B^* образи з різних класів такі, що є найближчими один до одного (рис. 2).

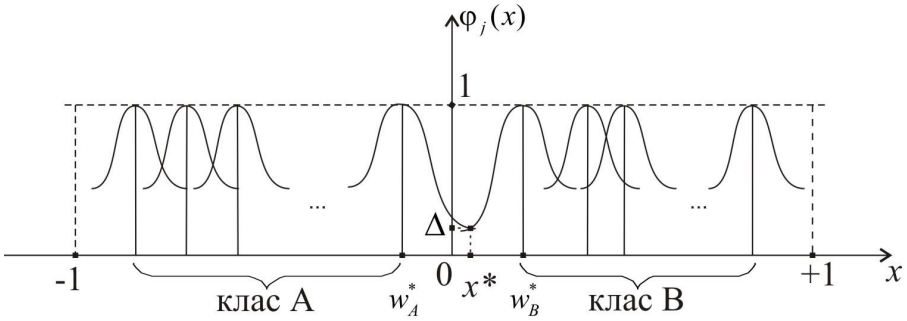

Рис. 2. Приклад бінарної класифікації та обчислення ширини σ

При цьому з кожним із них пов’язана своя активаційна функція:

$$\varphi_A^*(x) = \exp\left(-\frac{\left\|x - w_A^*\right\|^2}{2\sigma^2}\right), \tag{4}$$

$$\varphi_B^*(x) = \exp\left(-\frac{\left\|x - w_B^*\right\|^2}{2\sigma^2}\right). \tag{5}$$

Нескладно бачити, що ці функції перетинаються в точці

$$x^* = w_A^* + 0{,}5\left\|w_B^* - w_A^*\right\| = 0{,}5\left\|w_B^* + w_A^*\right\| \tag{6}$$

та набувають у ній значення

$$\varphi_A^*\left(x^*\right) = \varphi_B^*\left(x^*\right) = \exp\left(\frac{-0{,}125\left\|w_B^* - w_A^*\right\|^2}{\sigma^2}\right). \tag{7}$$

Якщо задати далі деякий поріг класифікації Δ, можна записати:

$$\exp\left(\frac{-0{,}125\left\|w_B^* - w_A^*\right\|^2}{\sigma^2}\right) = \Delta, \tag{8}$$

$$\frac{-0{,}125\left\|w_B^* - w_A^*\right\|^2}{\sigma^2} = \ln\Delta, \tag{9}$$

звідки

$$\sigma^2 = \frac{-0{,}125\left\|w_B^* - w_A^*\right\|^2}{\ln\Delta} = -\frac{\left\|w_B^* - w_A^*\right\|^2}{8\ln\Delta}. \tag{10}$$

Оскільки $\Delta < 1$, то $\ln\Delta < 0$ та $\sigma^2 > 0$.

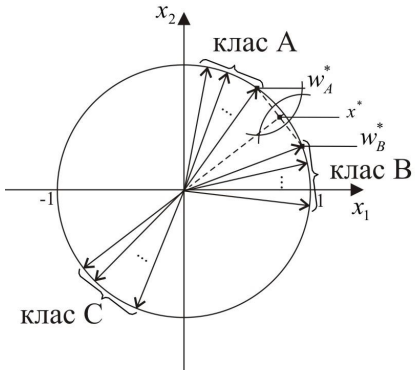

Рис. 3. Приклад двовимірного простору з трьома класами

Багатовимірну ситуацію розглянемо в двовимірному просторі з трьома класами A, B, C (рис. 3) так, що $\|x\| = \|w\| = 1$. У цьому випадку потрібно знайти мінімальну відстань між образами з різних класів. Зрозуміло, що на рис. 3 це $\left\|w_B^* - w_A^*\right\|$. Тут виникає процес, аналогічний конкуренції в мапах Кохонена (SOM), але як «переможців»приймають найближчі образи з різних класів. Якщо *j*-й образ $x_R(j)$ належить до класу *R* (класу A або B або C), а *k*-й $x_{\overline{R}}(k)$ не належить, то для усіх $j = 1,2,\mathbf{K},N; k = 1,2,\mathbf{K},N,$ необхідно знайти пару, для якої $x_R^T(j)x_{\overline{R}}(k)$ є максимальним.

Для цього в MPNN вводять латеральні зв’язки як в SOM, формуючи конкурентну ймовірнісну нейронну мережу (Competitive Probabilistic Neural Network – CPNN). Оскільки

$$\left\|w_B^* - w_A^*\right\|^2 = \left\|w_B^*\right\|^2 - 2{w_B^*}^T w_A^* + \left\|w_A^*\right\|^2 = 2\left(1 - {w_B^*}^T w_A^*\right), \tag{11}$$

то

$$\varphi_A^*\left(x^*\right) = \varphi_B^*\left(x^*\right) = \exp\frac{{w_B^*}^T w_A^* - 1}{4\sigma^2}, \tag{12}$$

звідки

$$\sigma^2 = \frac{w_B^{*T} w_A^{*} - 1}{4 \ln \Delta}. \qquad (13)$$

Оскільки $-1 \le w_B^{*T} w_A^{*} \le 1$, то чисельник (13) належить інтервалу $[-2,0]$.

Тобто, об'єднавши нейрони в групи класів та ввівши латеральні зв'язки між ними, отримуємо конкурентну імовірнісну нейронну мережу, що дає змогу визначати значення ширини активаційної функції автоматично.

Результати експериментальних досліджень

Запропонована конкурентна ймовірнісна нейронна мережа використовувалася для обробки текстової інформації та отримання вихідних значень метаданих для складання семантичних анотацій. Робота CPNN оцінювалась на вибірці «20 Newsgroups DataSet» (comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware).

Під час експерименту розглядалася якість роботи класифікатора та вплив на нього значень σ. Показано, що в результаті роботи формується набір значень імовірностей належності вхідного текстового об'єкта до кількох класів, які розглядаються як концепти онтології предметної галузі. Для тестування при проведенні експериментів використовували 40% вихідних даних навчальної вибірки. У цьому випадку $\sigma = 0{,}33663$ – значення ширини активаційної функції, що відповідає порогу класифікації $\Delta = 0.5$, визначено автоматично. За результатами експерименту було встановлено також, що для цієї задачі кращим є значення $\Delta \in [0{,}4;0{,}5]$. Для значень у межах *(0; 0,4)* значення щільності ймовірності для класів, до яких об'єкт не належить, визначається занадто малими частками одиниці, а для $\Delta \in [0{,}5;1]$ ширина активаційної функції набуває значень, які призводять до розмиття класів і помилок класифікації.

Приклад роботи програми

№ з/п	Імовірність належності до 1-го класу	Імовірність належності до 2-го класу	Імовірність належності до 3-го класу	Вихідний клас
1	0,99797	0,00062412	0,0014012	1
2	0,65024	0,11225	0,23751	1
3	0,36445	0,55605	0,079505	2
4	0,21421	0,73347	0,052324	2
5	0,17539	0,058827	0,76578	3

У результаті експериментальних досліджень було встановлено, що запропонований метод має високі показники точності і швидкості роботи, що дає можливість підвищити якість видобування знань із текстових джерел при обмеженій вибірці. Крім того, обчислення ширини активаційної функції за допомогою запропонованого підходу дає змогу отримати точніші значення ймовірностей належності текстового документа до кожного класу (концепту онтології).

Висновки

У роботі запропоновано конкурентну ймовірнісну нейронну мережу, яка є гібридом стандартних PNN, GRNN і SOM та дає змогу налаштовувати значення ширини активаційної функції автоматично завдяки введенню латеральних зв'язків між групами класів у шарі образів. Це рішення покладено в основу методу семантичного анотування текстових документів. Отже, стає можливим визначити точніші значення ймовірностей належності вхідного текстового об'єкта до кожного з потенційно можливих класів онтології предметної галузі для формування семантичних анотацій. Запропонований метод передбачає можливість оброблення інформації у міру її надходження, характеризується простотою числової реалізації і високою швидкістю.

1. V. Uren, Ph. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics:

Science, Services and Agents on the World Wide Web 2006, Vol. 4, No. 1. P. 14–28. 2. G. Xu, Ya. Zhang, L. Li. Web Mining and Social Networking: Techniques and Applications. –Springer: Web Information Systems Engineering and Internet Technologies Book Series, 2011. – 210 p. 3. Resource Description Framework: Overview. [Електронний ресурс]. – Режим доступу: http://www.w3.org/RDF/. 4. OWL Web Ontology Language: Overview. [Електронний ресурс]. – Режим доступу: http://www.w3.org/TR/owl-features/. 5. D.F. Specht Probabilistic neural network. Neural Networks 1990, 3. P. 109–118. 6. Бодянский Е.В., Шубкина О.В. Семантическое аннотирование текстовых документов с использованием модифицированной вероятностной нейронной сети // Системные технологи: Региональный межвузовский сборник научных трудов. – Днепропетровск, 2011. – Вып. 4 (75). – С. 48–55. 7. L.H. Tsoukalas, R.E. Uhrig. Fuzzy and Neural Approaches in Engineering. – N.Y.: John Willey and Sons Inc., 1997. – 587 p. 8. D.F. Spech. A general regression neural network. IEEE Trans. on Neural Networks 1991, 2. P. 568–576. 9. O. Nelles. Nonlinear System Identification. – Berlin: Springer, 2001. – 785 p. 10. D.R. Zahirniak, R. Chapman, S.K. Rogers, B.W. Suter, M. Kabriski, V. Pyatti. Pattern recognition using radial basis function network. Proc: 6-th Ann. Aerospace Application of AI Conf. Dayton, OH, 1990, Pp. 249–260. 11. R. Callan. The Essence of Neural Networks. – London: Prentice Hall Europe, 1999. – 248 p.

УДК 004.4’232

О. Овсяк
Київський національний університет культури і мистецтв,
Українська академія друкарства

МОДЕЛЬ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОПРАЦЮВАННЯ ФОРМУЛ АЛГОРИТМІВ

Модель інформаційної технології опрацювання формул алгоритмів описано розширеною алгеброю алгоритмів у вигляді рекурентно-декомпозиційної моделі. За ознакою функціонального призначення модель декомпоновано на субмоделі. Описано візуальну і функціональну підмоделі інформаційної технології. Наведено фрагмент програмної реалізації моделі.

Ключові слова: модель, інформаційна технологія, декомпозиція, підмодель, секвентування, елімінування, паралелення, функційний унітерм, унітерм.

Described the extended algebra algorithms recurrent – decomposition model of information technology for synthesis and processing algorithms formulas. Model by functional appointment of the decomposing at sub models. Described visual and functional sub models information technology. An piece of software implementation model.

Keywords: model, information technology, decomposition, sub models, sequention, elimination, parallelization, function uniterm, uniterm.

Вступ і формулювання задачі

Модель декомпозиції комп’ютерної системи генерування програмного коду з формул алгоритмів [1–3] описана у статті [4]. Але у цій моделі не подано інформації про наслідування підмоделей і декомпозиції підмоделей на графічну та функціональну субмоделі. Графічна субмодель реалізується у вигляді графічних вікон, які використовуються для отримання інформації, вибору і задання параметрів, введення і виведення графічно-текстових даних тощо. Тоді як функціональна підмодель призначена для опису функціонування. Графічна і функціональна субмоделі мають бути поєднані між собою. Наприклад, таке поєднання субмоделей може бути виконано через вибір функційних унітермів, який здійснюється при виконанні певних умов.