Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

ЗОЛОТУХІН ОЛЕГ ВІКТОРОВИЧ

УДК 004.912:004.032.26

МЕТОДИ КЛАСИФІКАЦІЇ ПОЛІТЕМАТИЧНИХ ТЕКСТОВИХ ДОКУМЕНТІВ ІЗ ЗАСТОСУВАННЯМ НЕЙРО-ФАЗЗІ ТЕХНОЛОГІЙ

05.13.23 – системи та засоби штучного інтелекту

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2015

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник – кандидат технічних наук, доцент
Рябова Наталія Володимирівна,
Харківський національний університет радіоелектроніки, професор кафедри штучного інтелекту.

Офіційні опоненти: доктор технічних наук, професор
Гороховатський Володимир Олексійович,
Харківський інститут банківської справи Університету банківської справи Національного Банку України, завідувач кафедри інформаційних технологій;

доктор технічних наук, професор
Шаронова Наталія Валеріївна,
Національний технічний університет «Харківський політехнічний інститут» Міністерства освіти і науки України, завідувач кафедри інтелектуальних комп'ютерних систем.

Захист відбудеться «____» _____ 2015 р. о _____ годині на засіданні спеціалізованої вченої ради Д 64.052.01 Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий «____» _____ 2015 р.

Учений секретар
спеціалізованої вченої ради О.А. Винокурова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Задача класифікації текстових документів - одна з основних в організації текстових даних і Text Mining. Останнім часом для розв'язання задач класифікації текстових документів все частіше використовуються методи обчислювального інтелекту, а саме методи інтелектуального аналізу текстів. Вагомий внесок у розвиток методів класифікації, в тому числі нечіткої, був зроблений такими вченими як: Вятченін Д.А., Рижов А.П., Bezdek J.C., Kohonen T., Ciarelli P., Krishnapuram R., Sanches J., Hammer B., Kim Y. та іншими. Незважаючи на істотні досягнення в галузі класифікації, ще залишається ряд задач, які потребують свого остаточного вирішення.

Особливістю методів інтелектуального аналізу текстів є встановлення наявності та характеру прихованих закономірностей у аналізованих документах. Задача ускладняється тим, що безліч документів є політематичними, тобто відносяться одночасно до декількох категорій. В той же час більшість відомих методів класифікації не враховують цієї особливості і орієнтовані на знаходження чітких класів, а крім того, не можуть в послідовному режимі класифікувати вхідні дані, що є істотним недоліком відомих підходів.

Проблема класифікації політематичних текстових документів потребує створення нових методів класифікації, що працюють в умовах нечітких класів, і які є ефективними в online режимі послідовної обробки. У зв'язку з цим, розробка методів класифікації політематичних текстових документів з урахуванням класів, що перетинаються, та послідовної подачі документів на обробку на основі методів обчислювального інтелекту є актуальною.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках держбюджетних НДР: № 243 «Методи, моделі та інформаційні технології розбудови соціально-економічної освітньо-наукової мережі з метою інтеграції у європейський простір» (№ ДР 0109U002497); № 265 «Методи та моделі самоорганізації інфраструктури інтелектуального інформаційного середовища, що базується на використанні принципів хмарних обчислень» (№ ДР 0112U000206), в яких здобувач брав участь як виконавець. В рамках зазначених тем здобувачем запропоновано: адаптивну нечітку нейронну мережу навчаємого векторного квантування; метод нечіткої класифікації політематичних текстових документів, які представлені у формі векторів-образів; метод нечіткої класифікації політематичних текстових документів з можливістю віднесення текстового документа з різним ступенем належності до різних класів.

Мета і задачі дослідження. Метою даного дослідження є розробка методів нейро-фаззі класифікації політематичних текстових документів, що дозволить поліпшити якість класифікації документів, які надходять на обробку в послідовному режимі.

Відповідно до поставленої мети в дисертаційній роботі необхідно

вирішити такі основні задачі:

– аналіз існуючих методів класифікації політематичних текстових документів;

– розробка архітектури нечіткої ймовірнісної нейронної мережі і методу її навчання для задач класифікації в online режимі;

– розробка архітектури адаптивної нечіткої нейронної мережі векторного квантування для класифікації політематичних текстових документів та методу навчання для розробленої мережі;

– розробка архітектури нейронної мережі зустрічного поширення з контрольованим навчанням, яка дозволить поліпшити якість класифікації за умов класів, що перетинаються;

– імітаційне моделювання та вирішення практичних задач;

– розробка модуля класифікації пошукових результатів інформаційно-пошукової системи.

Об'єкт дослідження – процес класифікації політематичної текстової інформації в інтелектуальних системах обробки документів.

Предмет дослідження – методи класифікації політематичних текстових документів з використанням нейро-фаззі технологій.

Методи дослідження. Основними методами дослідження є апарат обчислювального інтелекту: теорія нечітких штучних нейронних мереж, за допомогою якої були синтезовані нові методи, що дозволяють виконувати класифікацію текстової інформації; методи обробки природно-мовної інформації, які дозволили представити текстові документи в необхідному для машинної обробки форматі. Імітаційне моделювання – для визначення ефективності застосування розроблених методів.

Наукова новизна отриманих результатів. До нових, одержаних особисто автором, належать такі результати:

1. Вперше запропоновано архітектуру і метод навчання нечіткої ймовірнісної нейронної мережі, в якій перший прихований шар сформований не векторами-образами, а прототипами, що дозволяє суттєво скоротити кількість нейронів в мережі, а відповідно і кількість параметрів, які налаштовуються, та підвищити швидкодію при нечіткій класифікації в online режимі надходження на обробку текстових документів.

2. Вперше запропоновано метод навчання адаптивної нечіткої нейронної мережі навчаємого векторного квантування, який налаштовує синаптичні ваги в режимі навчання з учителем з елементами конкуренції за типом «переможець отримує все», що дозволяє в послідовному режимі надходження інформації проводити класифікацію текстових документів за умов класів, що перетинаються.

3. Вперше запропоновано нейронну мережу зустрічного поширення з контрольованим навчанням, яка характеризується поліпшеними апроксимуючими властивості завдяки тому, що вихідний шар утворений елементарними персептрони Розенблатта, а прихований шар сформований на

основі навчаємого векторного квантування, що дозволяє підвищити швидкодію процесу класифікації.

Практичне значення отриманих результатів. Запропоновані в роботі методи нейро-фаззі класифікації політематичних текстових документів можуть бути використані для покращення якості роботи багатьох сервісів Інтернет, при створенні інформаційно-пошукових систем нового рівня здатних у процесі послідовної обробки політематичних текстових документів відносити їх до визначених класів, причому один і той самий документ може одночасно належати до декількох класів.

Розроблені в роботі методи підтвердили свою ефективність в задачах класифікації політематичних текстових документів, які надходять на обробку в online режимі, впроваджені в науковій бібліотеці Харківського національного університету радіоелектроніки, м. Харків (акт впровадження від 10.09.2014) та у комунальному закладі охорони здоров'я «Первомайська центральна районна лікарня» (акт впровадження від 2.10.2014).

Запропоновані в роботі архітектури нечітких нейронних мереж та методи їх навчання для класифікації політематичних текстових документів, були використані в курсах «Штучні нейронні мережі», «Нейромережеві методи обчислювального інтелекту», «Системи інтелектуальної обробки природно-мовної інформації» (акт впровадження від 19.06.2014), та в держбюджетних науково-дослідних роботах Харківського національного університету радіоелектроніки (акт впровадження від 26.06.2014).

Особистий внесок здобувача. Основні положення і результати дисертаційної роботи одержані здобувачем самостійно. У роботах, написаних у співавторстві, здобувачеві належать: [1] – запропонована архітектура нечіткої ймовірнісної нейронної мережі; [2] – запропонований метод навчання нейронної мережі зустрічного поширення з контрольованим навчанням; [4] – запропонований метод навчання адаптивного навчаємого векторного квантування; [7] – запропонований метод класифікації із застосуванням нечітких технологій; [17] – запропонований метод навчання нейронної мережі зустрічного поширення;

Апробація результатів дисертації. Основні положення і результати дисертаційної роботи були представлені на: 13-му, 14-му, 16-му, 18-му Міжнародних молодіжних форумах "Радіоелектроніка и молодь в XXI столітті" (Харків, 2009, 2010, 2012, 2014); 1-й Факультетській науково-практичній молодіжній школі-семінарі «Інформаційні інтелектуальні системи» (Харків, 2008); 6-й та 8-й міжнародних науково-практичних конференціях «Математичне та програмне забезпечення інтелектуальних систем» (Дніпропетровськ, 2008, 2010); 1-й Науково-технічній конференції «Сучасні напрямки розвитку інформаційно-комунікаційних технологій та засобів управління» (Харків-Київ, 2010); Міжнародній науково-практичній конференції «Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту» (Євпаторія, 2011); 5-ій Міжнародній конференції молодих вчених «Комп'ютерні науки та інженерія (CSE-2011)» (Львів, 2011);

Міжнародній науковій конференції «Наукова періодика слов'янських країн в умовах глобалізації. Інформаційні технології» (Київ, 2012); 6-ій Міжнародній науково-практичній конференції молодих вчених и студентів «Інформаційні процеси та технології (Інформатика-2013)» (Севастопіль, 2013); 3-й Міжнародній науково-практичній конференції «Інформаційні системи та технології» (Харків, 2014).

Публікації. Основні положення дисертаційної роботи опубліковані в 18 наукових працях: у тому числі 5 статях у наукових фахових виданнях України з технічних наук, що включено до переліку МОН України (3 статті в виданнях, що входять до міжнародних наукометричних баз), 13 публікацій у збірниках праць міжнародних наукових конференцій і форумів.

Структура та обсяг дисертації. Дисертація складається зі вступу, чотирьох розділів, висновків, що містять основні результати, списку використаних джерел і додатку. Загальний обсяг дисертації складає 145 сторінок (з них 122 – основного тексту), 12 рисунків, з них 2 на окремих сторінках, 3 таблиці, список використаних джерел, що включає 142 найменування та займає 16 сторінок, 1 додаток на 6 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету і задачі дослідження, наукову новизну і практичне значення одержаних результатів. Наведено відомості про впровадження результатів роботи, апробацію, особистий внесок здобувача та публікації.

У **першому розділі** наведено аналіз існуючих методів класифікації текстових документів. Розглянуті питання попередньої обробки текстових документів, які є політематичними, для подальшої їх обробки комп'ютерними засобами. Розглянуто процес формування простору ознак та основні методи його скорочення.

Політематичний текстовий документ може бути поданий у вигляді тексту, що складається з менших текстових модулів, кожен з яких відноситься до однієї або декількох тематик документа. Отже, текстовий документ розбивається на неподільні ділянки, які можуть бути розпізнані на різних рівнях в логічній структурі тексту (наприклад, розділ, параграф).

На сьогодні існує небагато методів, які дозволяють обробляти політематичні текстові документи – виконувати їх класифікацію чи кластеризацію, вони мають низку недоліків, пов'язаних з обчислювальною складністю та неможливістю знаходження в процесі своєї роботи нових класів/кластерів. Метою класифікації політематичних текстових документів є віднесення одного й того ж текстового документа до більш ніж однієї тематики. Мета класифікації політематичних текстових документів – класифікувати текстові документи так, щоб кожний документ відносився до більш ніж до одного класу.

На основі проведеного аналізу сформульовано мету та задачі дослідження, які полягають в створенні архітектур нейронних мереж та методів навчання, а також розв'язання за їх допомогою задач класифікації політематичних текстових документів, які поступають на обробку в послідовному режимі.

У **другому розділі** розглянуто задачу створення класифікації політематичних текстових документів на основі нечіткої ймовірнісної нейронної мережі.

Вперше запропоновані архітектура та метод навчання нечіткої ймовірнісної нейронної мережі, яка призначена для вирішення задач класифікації текстових документів, що дозволяє уникнути «прокльону розмірності» при великій кількості і розмірності документів, що

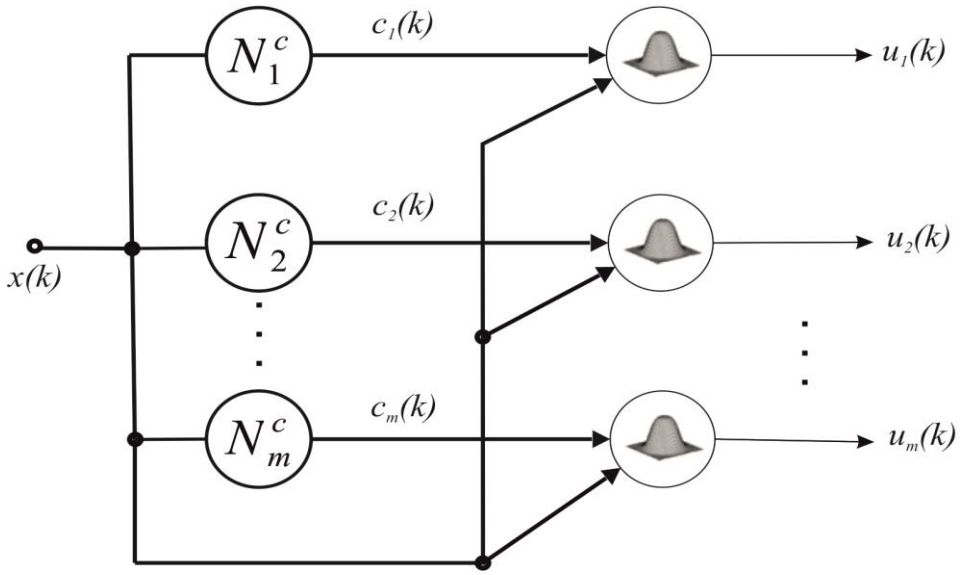

Рисунок 1 – Нечітка ймовірнісна нейронна мережа

класифікуються, та організувати їх обробку в online режимі, а також запропонована адаптивна нечітка нейронна мережа навчаємого векторного квантування, яка дозволяє вирішувати задачу класифікації текстових документів в режимі послідовної обробки в умовах класів, що перетинаються.

Архітектура запропонованої нечіткої ймовірнісної нейронної мережі наведена на рис. 1 Вона містить два шари обробки інформації: перший прихований шар прототипів (замість першого прихованого шару образів у звичайній ймовірнісній нейронній мережі) і вихідний шар обчислення рівнів належності.

Вихідною інформацією для навчання є послідовність векторів-образів $x(1),x(2),\ldots,x(k),\ldots,x(N)$, $x(k)=(x_1(k),x_2(k),\ldots,x_n(k))^T\in R^n$, які розмічені, при цьому передбачається, що N може змінюватися в часі, кількість класів m також може змінюватися, прототипи (центроїди) класів описуються векторами $c_j=(c_{j1},c_{j2},\ldots,c_{jn})^T$, що підлягають визначенню, позначення $x(k,j)$ означає, що вектор-образ $x(k)$ віднесений до j-ого класу, $j=1,2,\ldots,m$, при чому

кожен клас містить N_j класифікованих образів, $\sum_{j=1}^{m} N_j = N$. Для обчислення прототипів використовується звичайна оцінка середнього арифметичного $c_j = \frac{1}{N_j}\sum_{k=1}^{N_j} x(k,j)$, яку нескладно переписати в рекурентній формі

$$c_j(k) = c_j(k-1) + k^{-1}\left(x(k,j) - c_j(k-1)\right), \qquad (1)$$

що відповідає правилу навчання Т. Кохонена з параметром кроку $\eta(k) = k^{-1}$, відповідним до умов стохастичної апроксимації.

Оскільки в реальних задачах прототипи класів можуть дрейфувати в часі, замість (1) можна використовувати або експоненційне зважене середнє

$$c_j(k) = \alpha c_j(k-1) + (1-\alpha)\left(x(k,j) - c_j(k-1)\right),\ 0 < \alpha < 1,$$

або адаптивну процедуру

$$\begin{cases} c_j(k) = c_j(k-1) + \eta(k)\left(x(k,j) - c_j(k-1)\right), \\ \eta(k) = r^{-1}(k),\ r(k) = \alpha r(k-1) + \left\|x(k,j)\right\|^2,\ 0 \le \alpha \le 1, \end{cases}$$

задовольняючу при $\alpha = 1$ умовам А. Дворецького.

Вихідний шар мережі оцінює рівень належності некласифікованих спостережень $x(k)(k > N)$ до сформованих класів із прототипами $c_j(N)$ за допомогою виразу

$$u_j(k) = \frac{\left\|x(k) - c_j(N)\right\|^{-2}}{\sum_{l=1}^{m}\left\|x(k) - c_l(N)\right\|^{-2}}, \qquad (2)$$

який пов’язаний з ймовірнісною процедурою нечіткої класифікації, відомої як метод нечітких С-середніх. Таким чином, у процесі навчання мережі одночасно використовуються чіткі й нечіткі процедури. Переписавши цей вираз у вигляді $u_j(k) = \left(1 + \left\|x(k) - c_j(N)\right\|^2 \sum_{\substack{l=1 \\ l \ne j}}^{m}\left\|x(k) - c_l(N)\right\|^{-2}\right)^{-1}$, можна помітити, що це є дзвонувата (ядерна) функція активації

$$u_j(k) = \left(1 + \frac{\left\|x(k) - c_j(N)\right\|^2}{\sigma_j^2}\right)^{-1},$$

з параметром ширини рецепторного поля

$$0 \le \sigma_j^2 = \left(\sum_{\substack{l=1 \\ l \ne j}}^{m}\left\|x(k) - c_l(N)\right\|^{-2}\right)^{-1} \le \frac{4}{m-1},$$

що установлюється автоматично в процесі класифікації.

Оскільки вираз, який оцінює рівні належності, відносяться до ймовірнісної нечіткої класифікації, тобто виконується умова

$$\sum_{j=1}^{m} u_j(k) = 1,$$

ситуація, при якій $u_j(k) = m^{-1} \forall j,$ означає, що спостереження $x(k)$ відноситься або до всіх класів однаково, що є малоймовірним, або до жодного з них. У цій ситуації можна збільшити число класів до $m+1,$ поклавши $x(k)$ в якості початкового прототипу нового класу. Якщо ж буде виявлено, що для p класів $p < m$ рівень належності $u_j(k)$ виявиться менше m^{-1}, це означає, що $x(k)$ не може належати цим класам, і рівні належності слід перерахувати за допомогою виразу (2), змінивши верхній індекс підсумовування в знаменнику на $m - p$.

Для виключення можливих p класів, що не включають у себе $x(k),$ може бути також використана процедура, заснована на V-критерії (Vigilance criterion) і перевірці умови

$$e^{u_j(k)} \left\| x(k) - c_j(N) \right\| \le \varepsilon,$$

де граничне значення ε встановлюється емпірично. Зрозуміло, що при $p = m - 1$, одержуємо чіткий результат класифікації.

Для розв'язку задач нечіткої класифікації в умовах класів, що перетинаються, був введений цілий ряд модифікацій LVQ-систем.

Розроблена архітектура нейро-фаззі мережі адаптивного нечіткого навчаємого векторного квантування наведена на рис.2. Архітектура містить два шари обробки інформації, при цьому нейрони першого прихованого шару зв'язані між собою латеральними зв'язками, за допомогою яких реалізуються процеси конкуренції. Вихідною інформацією для навчання є послідовність

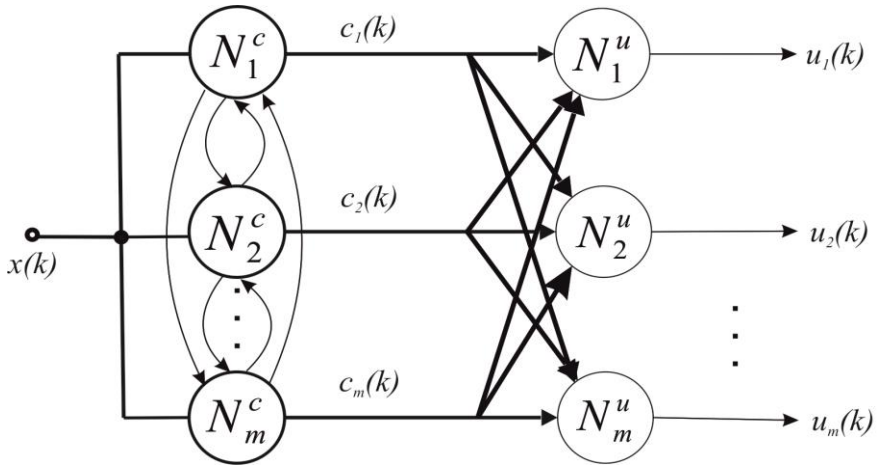

Рисунок 2 – Нейронна мережа адаптивного нечіткого навчаємого векторного квантування

векторів-образів $x(1), x(2), \ldots, x(k), \ldots, x(N)$, $x(k) = (x_1(k), x_2(k), \ldots, x_n(k))^T \in R^n$ з відомою класифікацією, при цьому вхідні сигнали попередньо нормуються так, що $\|x(k)\| = 1$. Нейрони першого прихованого шару N_j^c $(j = 1,2,\ldots,m;\ m -$ апріорі задана кількість можливих класів) призначені для знаходження прототипів (центроїдів) класів $c_j(k) = (c_{j1}(k), c_{j2}(k), \ldots c_{ji}(k) \ldots c_{jn}(k))^T$, при цьому компоненти $c_{j1}(k)$, за суттю, є синаптичними вагами нейрона N_j^c. Нейрони вихідного шару N_j^u обчислюють рівні належності $u_j(k)$ пред'явленого образа $x(k)$ до $j -$ого класу.

При надходженні образа $x(k)$ в процесі конкуренції визначається нейрон-переможець j^*, синаптичні ваги котрого $c_{j^*}(k-1)$ є найбільш близькими до вхідного сигналу, тобто

$$j^* = \arg\min_j D(x(k), c_j(k-1)) = \arg\min_j \|x(k) - c_j(k-1)\|^2 =$$

$$= \arg\max_j x^T(k), c_j(k-1) = \arg\max_j \cos(x(k), c_j(k-1)),$$

при цьому

$$-1 \le \cos(x(k), c_j(k-1)) = x^T(k), c_j(k-1) \le 1, \qquad (3)$$

а

$$0 \le \|x(k) - c_j(k-1)\|^2 \le 4.$$

Оскільки навчання є контрольованим, то належність вектору $x(k)$ до конкретного класу відома, що дозволяє розглянути дві можливі ситуації, які виникають в навчаємому векторному квантуванні:

- вхідний вектор $x(k)$ та нейрон-переможець $N_{j^*}^c$ належить одному класу;
- вхідний вектор $x(k)$ та нейрон-переможець $N_{j^*}^c$ належить різним класам.

Тоді стандартне LVQ-правило навчання може бути записано в вигляді

$$c_j(k) = \begin{cases} c_{j^*}(k-1) + \eta(k)\left(x(k) - c_{j^*}(k-1)\right) & - \ x(k) \ та \ c_{j^*}(k-1) \in одному\ класу, \\ c_{j^*}(k-1) - \eta(k)\left(x(k) - c_{j^*}(k-1)\right) & - \ x(k) \ та \ c_{j^*}(k-1) \in різним\ класам, \\ c_j(k-1) & - \ j-ий\ нейрон\ не\ переміг. \end{cases}$$

Що стосується вибору величини кроку навчання $\eta(k)$, то загальна рекомендація зводиться до того, що він повинен монотонно зменшуватися у

процесі настройки. Для обчислення кроку пошуку була запропонована процедура

$$\eta(k)=r^{-1}(k),\ r(k)=\alpha r(k-1)+\left\|x(k,j)\right\|^2,\ 0\le\alpha\le 1,$$

при цьому при $\alpha=1$ параметр шага $\eta(k)=k^{-1}$, тобто задовольняє умовам Дворецького. Нормування вхідних сигналів $x(k)$ не гарантує, що прототипи класів будуть відповідати умові $c_j(k)=1$, а його невиконання унеможливить в якості оцінки відстані використовувати скалярні процедури (3). Обійти це утруднення нескладно, ввівши додаткове нормування синаптичних ваг в процесі навчання. В результаті приходимо до адаптивної процедури вигляду

$$c_j(k)=\begin{cases}\dfrac{c_{j^*}(k-1)+\eta(k)\left(x(k)-c_{j^*}(k-1)\right)}{\left\|c_{j^*}(k-1)+\eta(k)\left(x(k)-c_{j^*}(k-1)\right)\right\|} - x(k)\ та\ c_{j^*}(k-1)\in одному\ класу,\\ \dfrac{c_{j^*}(k-1)-\eta(k)\left(x(k)-c_{j^*}(k-1)\right)}{\left\|c_{j^*}(k-1)-\eta(k)\left(x(k)-c_{j^*}(k-1)\right)\right\|} - x(k)\ та\ c_{j^*}(k-1)\in різним\ класам,\\ \eta(k)=r^{-1}(k), r(k)=\alpha r(k-1)+1, 0<\alpha\le 1,\\ c_j(k-1)\ -\ якщо\ j-ий\ нейрон\ не\ переміг.\end{cases} \quad (4)$$

Розраховані за допомогою правила навчання (4) прототипи $c_j(k)\,(c_j(N)$ у випадку, якщо навчальна вибірка має фіксований обсяг), подаються на вхідний шар, де обчислюються рівні належності

$$u_j(k)=\frac{\left\|x(k)-c_j(N)\right\|^{-2}}{\sum_{l=1}^{m}\left\|x(k)-c_l(N)\right\|^{-2}}, \quad (5)$$

обумовлені координатами седлової точки лагранжіана

$$\nabla_{u_j}Z(u_j,c_j,\lambda)=\nabla_{u_j}\left(\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^2(k)\left\|x(k)-c_j\right\|^2+\lambda\left(\sum_{j=1}^{m}u_j(k)-1\right)\right)=\vec{0},$$

що полягає в основі широко розповсюдженого методу нечітких С-середніх (FCM) ймовірнісної нечіткої кластеризації.

Переписавши (5) у вигляді

$$u_j(k)=\frac{1}{1+\left\|x(k)-c_j(k)\right\|^2\sum\limits_{\substack{l=1\\ l\ne j}}^{m}\left\|x(k)-c_l(k)\right\|^{-2}}=\frac{1}{1+\dfrac{\left\|x(k)-c_j(k)\right\|^2}{\sigma_j^2(k)}}, \quad (6)$$

нескладно помітити, що вираз (6) задає дзвонувату функцію належності з центром в точці $c_j(k)$ і параметром ширини

$$0 \le \sigma_j^2 = \left(\sum_{\substack{l=1 \\ l \ne j}}^{m} \left\| x(k) - c_l(N) \right\|^{-2} \right)^{-1} \le \frac{4}{m-1},$$

тобто питання про конкретну форму функції належності тут вирішується автоматично. Таким чином, співвідношення (4), (5) задають online алгоритм навчання нечіткої нейронної мережі навчаємого векторного квантування.

Третій розділ присвячений розробці нейронної мережі зустрічного поширення з контрольованим навчанням для класифікації політематичних текстових документів. Архітектура розробленої мережі зустрічного поширення з контрольованим навчанням наведена на рис. 3.

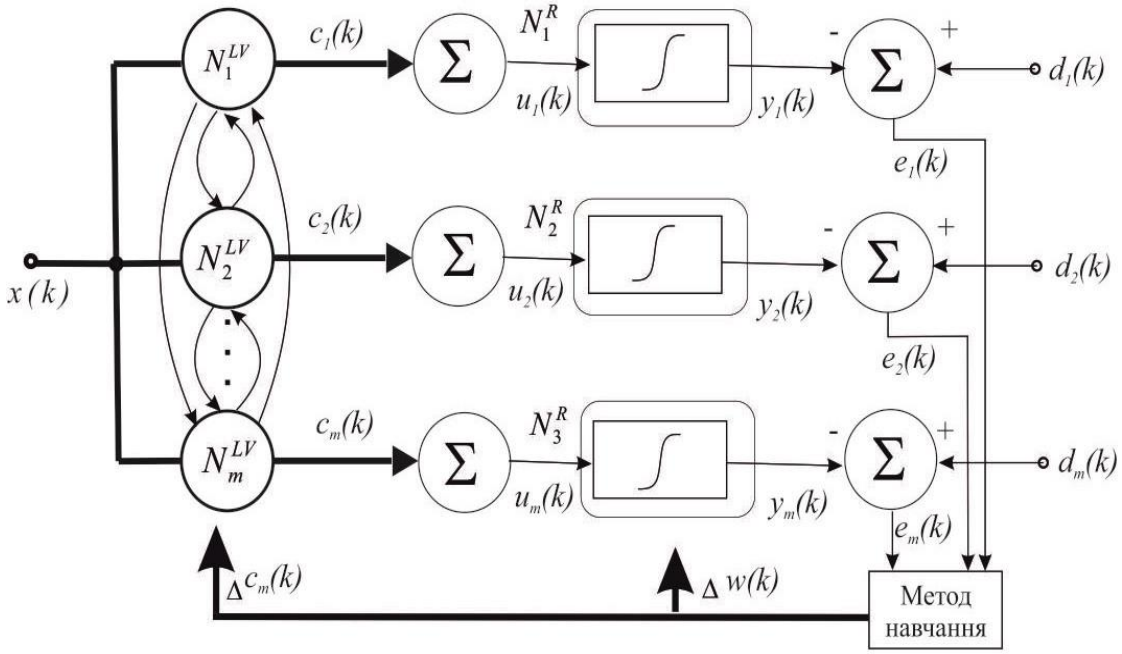

Рисунок 3 –Нейронна мережа зустрічного поширення з контрольованим навчанням

Найбільш відомі мережі зустрічного поширення є гібридом самоорганизовної мапи Т. Кохонена (перший прихований шар) та набору зірок С. Гроссберга (вихідний шар) і відповідно поєднує в собі конкурентне самонавчання з контрольованим навчанням з вчителем. При цьому можна відзначити, що вузлами цієї мережі, як правило, є адаптивні лінійні асоціатори, чий вхід лінійно залежить від синаптичних ваг. Цей факт визначає високу швидкість їх налаштування.

Вперше розроблено нейронну мережу зустрічного поширення з контрольованим навчанням. Специфіка задачі класифікації текстових документів вимагає істотного удосконалення як архітектури нейронної мережі зустрічного поширення, так і методів її навчання. По-перше, оскільки у

режимі навчання на вхід мережі подаються класифіковані образи, доцільно у першому прихованому шарі використовувати не традиційну самоорганізовну мапу, яка навчається без вчителя, а нейронну мережу векторного квантування, яка навчається з вчителем, що дозволяє підвищити швидкодію, і крім цього, може оцінити центроїди та границі класів.

По-друге, у вихідному шарі замість зірок Гроссберга доцільно використовувати елементарні персептрони Розенблатта з нелінійною функцією активації (релейної), що приймає тільки два значення: 1, якщо пропонований образ належить до даного конкретного класу, і 0 – якщо ні.

Вихідною інформацією для навчання є послідовність векторів-образів $x(1), x(2),..., x(k)...,$ де $x(k) = (x_1(k), x_2(k),..., x_n(k))^T \in R^n$ з відомою класифікацією, яка в online режимі, що характерно для завдань Text Mining, подається на нульовий (рецепторний) шар мережі. Нейрони першого прихованого шару N_j^{LV} ($j = 1,2,...,m$, m – кількість можливих класів), задаються апріорно, будучи по суті адаптивними лінійними асоціаторами, та призначені для знаходження центроїдів і границь класів, при цьому $(n \times 1)$-вектори, що описують ці центроїди $c_j(k) = (c_{j1}(k), c_{j2}(k),..., c_{jn}(k))^T$, є синаптичними вагами, що настроюються, кожного з N_j^{LV} нейронів . При цьому всі вхідні сигнали перед подачею на нульовий шар нормуються так, що $\|x(k)\| = 1.$

Вихідний шар мережі утворений m-елементарними персептронами Розенблатта N_j^R із сигмоїдальною функцією активації, при цьому

$$y_j(k) = \psi\left((\gamma u_j(k)\right) = \psi\left(\sum_{i=1}^{n} \gamma w_{ji} c_{ji}(k)\right) = \psi\left(\gamma w_j^T c_j(k)\right) = \frac{1}{1+e^{-\gamma u_j(k)}} = \frac{1}{1+e^{-\gamma w_j^T c_j(k)}},$$

де γ – параметр крутості активаційної функції, $w_j = (w_{j1}, w_{j2},..., w_{jn})^T$ – вектор синаптичних ваг N_j^R.

На рис. 4 показано залежність сигмоїдальної активаційної функції $\psi(\gamma u_j)$ залежно від параметра крутості γ. При цьому, чим більше значення γ, тим ближче $\psi(\gamma u_j)$ до релейної функції

$$\psi(\gamma u_j) = \begin{cases} 1, & при\ u_j \geq 0, \\ 0, & при\ u_j < 0, \end{cases} \qquad (7)$$

що звичайно використовується в задачах розпізнавання образів.

Зрозуміло, що при $\gamma \to \infty$, $\psi(\gamma u_j)$ співпадає з (7), не перетерплюючи при цьому розриву похідної в точці $u_j = 0$.

Слід зазначити, що вектори $c_j(k)$, будучи синаптичними вагами N_j^{LV}, подаються в якості вхідних сигналів на вихідні нейрони N_j^R.

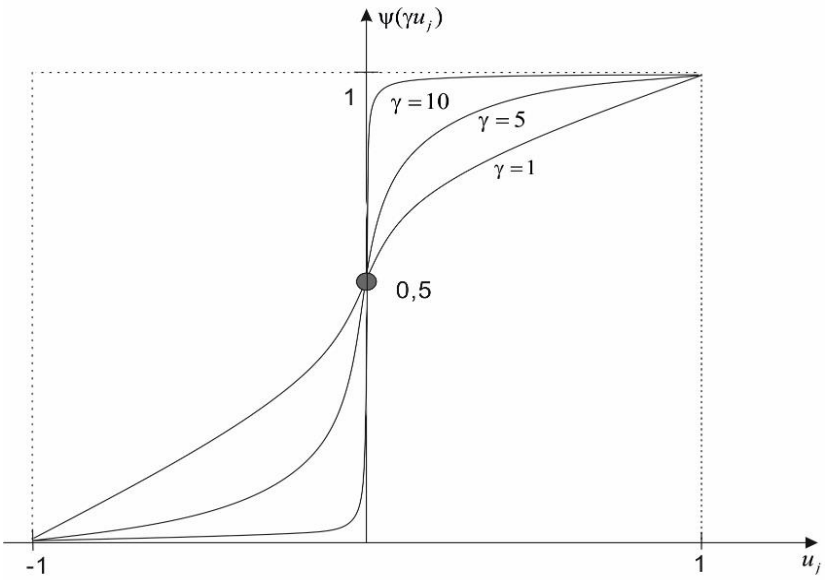

Рисунок 4 – Залежність сигмоїдальної функції від параметра крутості

Вихідні сигнали мережі $y_j(k)$ приймають два значення

$$y_j(k) \approx \begin{cases} 1, & \textit{якщо } x(k) \textit{ відноситься до } j-\textit{ого класу}, \\ 0, & \textit{в інших випадках}, \end{cases}$$

при цьому точні значення 1 або 0 ніколи не досягаються.

Навчання нейронної мережі зустрічного поширення з контрольованим навчанням реалізується таким чином. При подачі на вхід нейронної мережі вектора-образа $x(k)$ ($\|x(k)\| = 1$) в процесі конкуренції, реалізованої по латеральних (поперечних) зв'язкам першого прихованого шару між нейронами N_j^{LV}, визначається нейрон-переможець j^*, вектор синаптичних ваг якого $c_{j^*}(k-1)$ в сенсі прийнятої метрики (як правило, евклідової) є найбільш близьким до вхідного сигналу:

$$j^* = \arg\min_j D(x(k), \mathrm{c}_j(k-1)) = \arg\min_j \left\| x(k) - c_j(k-1) \right\|^2 =$$

$$= \arg\max_j c_j^T(k-1)x(k) = \arg\max_j \cos(\mathrm{c}_j(k-1), x(k)),$$

при цьому $-1 \le \cos(c_j(k-1), x(k)) = c_j^T(k-1)x(k) \le 1$, а

$0 \le \left\| x(k) - c_j(k-1) \right\|^2 \le 4$.

Оскільки навчання в першому прихованому шарі є контрольованим (на відміну від традиційної мережі зустрічного поширення), належність кожного вектора $x(k)$ до конкретного класу відома, що дозволяє розглянути дві можливі ситуації, що виникають у навчаємому векторному квантуванні:

– вхідний вектор $x(k)$ і нейрон-переможець $N_{j^*}^{LV}$ належать одному класу;

– вхідний вектор $x(k)$ і нейрон-переможець $N_{j^*}^{LV}$ належать різним класам.

Тоді LVQ-правило навчання може бути записане у вигляді:

$$c_j(k)=\begin{cases}\dfrac{c_{j^*}(k-1)+\eta(k)(x(k)-c_{j^*}(k-1))}{\left\|c_{j^*}(k-1)+\eta(k)(x(k)-c_{j^*}(k-1))\right\|},\\ \textit{якщо } x(k) \textit{ та } c_{j^*}(k-1)\in \textit{одному класу},\\ \dfrac{c_{j^*}(k-1)-\eta(k)(x(k)-c_{j^*}(k-1))}{\left\|c_{j^*}(k-1)-\eta(k)(x(k)-c_{j^*}(k-1))\right\|},\\ \textit{якщо } x(k) \textit{ та } c_{j^*}(k-1)\in \textit{різним класам},\\ c_j(k-1), \textit{ якщо } j-\textit{ий нейрон не переміг},\end{cases} \tag{8}$$

де $0<\eta(k)\le 1$ – параметр кроку навчання.

Правило контрольованого навчання (8) має ясний фізичний зміст: якщо нейрон-переможець і пропонований образ належать до одному класу, то центроїд $c_{j^*}(k-1)$ притягується до $x(k)$; якщо ж $c_{j^*}(k-1)$ і $x(k)$ належать до різних класів, то $c_{j^*}(k-1)$ відштовхується від $x(k)$. Таким чином, у процесі навчання мінімізується або максимізується відстань

$$D(x(k),c_{j^*}(k-1))=\left\|x(k)-c_{j^*}(k-1)\right\|^2,$$

при цьому автоматично проводиться нормування $\left\|c_j(k)\right\|=1.$

Що стосується вибору кроку навчання $\eta(k)$, то загальна рекомендація зводиться до того, що він повинен монотонно зменшуватися в процесі настроювання згідно із правилами стохастичної апроксимації. Вибір $\eta(k)=k^{-1}$ відповідає рекурентній оцінці середнього арифметичного за допомогою першого співвідношення системи (8). При $\eta(k)=1$ замість оцінок $c_j(k)$ на вихідний шар класифікації подаються самі вхідні образи $x(k)$.

Для навчання синаптичних ваг вихідного шару, утвореного елементарними персептронами Розенблатта, використовується стандартний квадратичний критерій вигляду

$$E_j(k)=\frac{1}{2}e_j^2(k)=\frac{1}{2}\left(d_j(k)-y_j(k)\right)^2=$$

$$=\frac{1}{2}\left(d_j(k)-\psi(\gamma u_j(k)\right)^2=\frac{1}{2}\left(d_j(k)-\psi\left(\sum_{i=1}^{n}\gamma w_{ji}c_{ji}(k)\right)\right)^2,$$

де $d_j(k) = \begin{cases} 1, \ якщо\ x(k)\ відноситься\ до\ j-ого\ класу, \\ 0, \ в\ іншому\ випадку. \end{cases}$

Тут слід зазначити, що хоча в завданні розпізнавання образів звичайно використовується, так званий, персептронний критерій, застосування квадратичного критерію з більшим значенням γ дозволяє оптимізувати процес навчання за швидкодією.

Мінімізація критерію навчання може бути забезпечена за допомогою рекурентної процедури (δ -правила навчання) вигляду:

$$w_{ji}(k) = w_{ji}(k-1) - \eta^R(k)\frac{\partial E_j(k)}{\partial e_j(k)} \bullet \frac{\partial e_j(k)}{\partial w_{ji}} = w_{ji}(k-1) - \eta^R(k) e_j(k)\frac{\partial e_j(k)}{\partial w_{ji}} =$$

$$= w_{ji}(k) - \eta^R(k) e_j(k)\frac{\partial e_j(k)}{\partial u_j(k)} \bullet \frac{\partial u_j(k)}{\partial w_{ji}} = w_{ji}(k) + \eta^R(k) e_j(k)\psi'(\gamma u_j(k)) c_{ji}(k) =$$

$$= w_{ji}(k) + \eta^R(k)\delta_j(k) c_{ji}(k),$$

де $\delta_j(k) = e_j(k)\psi'(\gamma u_j(k)) = -\frac{\partial E_j(k)}{\partial u_j}$ – локальна похибка.

У векторній формі алгоритм можна представити як:

$$w_j(k) = w_j(k-1) + \eta^R(k)\delta_j(k) c_j(k),$$

а з урахуванням сигмоїдальної активаційної функції:

$$w_j(k) = w_j(k-1) + \eta^R(k)\gamma e_j(k) y_j(k)(1-y_j(k)) c_j(k) =$$

$$= w_j(k-1) + \eta^R(k)\gamma (d_j(k) - w_j^T(k-1) c_j(k)) y_j(k)(1-y_j(k)) c_j(k) =$$

$$= w_j(k-1) + \eta^R(k) e_j(k) J_j(k),$$

де $J_j(k) = \gamma y_j(k)(1-y_j(k)) c_j(k)$.

Підвищити швидкодію процесу навчання вихідного шару можна, переходячи від градієнтних процедур до псевдон'ютонівських алгоритмів, серед яких можна відзначити популярний у теорії і практиці нейронних мереж алгоритм Левенберга-Марквардта. Вводячи однокрокову модифікацію цього алгоритму

$$w_j(k) = w_j(k-1) + (J_j(k) J_j^T(k) + \eta^R I)^{-1} J_j(k)(d_j(k) - \psi(\gamma w_j^T(k-1) c_j(k))) \qquad (9)$$

(тут $\eta^R > 0-$ параметр регуляризації, $I-(n\times n)-$ одинична матриця) і використовуючи формулу Шермана-Моррісона обернення матриць, можна переписати (9) простому виді

$$w_j(k) = w_j(k-1) + \frac{d_j(k) - \psi(\gamma w_j^T(k-1) c_j(k))}{\eta^R + \left\| J_j(k) \right\|^2} J_j(k),$$

що є поширенням на нелінійній випадок широко використовуваного оптимального алгоритму навчання нейронних мереж Уідроу-Хоффа і адитивного алгоритму Качмажа, прийнятого в задачах адаптивної ідентифікації.

Четвертий розділ присвячено імітаційому моделюванню та вирішенню практичної задачі класифікації політематичних текстових документів в інформаційно-пошуковій системі. Створений модуль класифікації результатів роботи цієї системи та дослідженню ефективності відображення та обробки результатів пошуку в ній.

Аналіз отриманих результатів показав, що розроблений у дисертаційній роботі метод класифікації політематичних текстових документів, що базується на адаптивній нечіткій самоорганізовній нейронній мережі з рекурентними методами навчання, а також модуль класифікації, створений на основі даного методу, забезпечує спрощення обробки результатів пошуку та скорочення часу обробки релевантної інформації шляхом автоматичного розбиття результатів інформаційного пошуку на категорії.

У **висновках** сформульовано теоретичні та практичні результати роботи.

У **додатках** наведено акти впровадження результатів дослідження в науковій бібліотеці Харківського національного університету радіоелектроніки, у комунальному закладі охорони здоров’я «Первомайська центральна районна лікарня», а також в навчальний процес і науково-дослідні роботи Харківського національного університету радіоелектроніки.

ВИСНОВКИ

У дисертаційній роботі представлені результати класифікації політематичних текстових документів в режимі послідовної обробки з використанням нейро-фаззі технологій, які відповідно до поставленої мети є розв’язанням актуальної науково-практичної задачі. Проведені дослідження дозволили зробити такі висновки.

1. Розроблено архітектуру і метод навчання нечіткої ймовірнісної нейронної мережі, яка характеризується наявністю в першому прихованому шарі прототипів, замість шару образів, що дозволяє уникнути «прокльону розмірності» при великій кількості і розмірності документів, що класифікуються, і організувати їх обробку в online режимі.

2. Розроблено архітектуру адаптивної нечіткої нейронної мережі векторного квантування для класифікації політематичних текстових документів, яка дозволяє в послідовному режимі надходження інформації проводити класифікацію текстових документів за умов класів, що перетинаються.

3. Розроблено метод навчання адаптивної нечіткої нейронної мережі навчаємого векторного квантування, яка характеризується настройкою семантичних ваг в режимі навчання з учителем з елементами конкуренції за

типом «переможець отримує все», що дозволяє вирішувати задачу класифікації текстових документів в режимі послідовної обробки в умовах перетину класів.

4. Розроблено нейронну мережу зустрічного поширення з контрольованим навчанням, яка характеризується поліпшеними апроксимуючими властивості завдяки тому, що вихідний шар утворений елементарними персептронами Розенблатта, а прихований шар сформований на основі навчаємого векторного квантування, що дозволяє підвищити швидкість процесу класифікації.

5. Проведено імітаційне моделювання на тестових даних корпусу текстів. Слід відзначити, що розроблена нейро-фаззі мережа показала високу якість класифікації політематичних текстових документів.

6. Розроблені в дисертаційній роботі методи використано при створенні модуля класифікації політематичних текстових документів в інформаційно-пошукової системи наукової бібліотеки ХНУРЕ, у комунальному закладі охорони здоров'я «Первомайська центральна районна лікарня» для діагностики у ургентних хворих патології внутришньочеревних органів, що підтверджено відповідними актами.

7. Розроблені в дисертаційній роботі методи класифікації політематичних текстових документів використано на кафедрі штучного інтелекту при підготовці дисциплін для освітньо-кваліфікаційного рівня бакалавр: «Штучні нейронні мережі: архітектури, навчання, застосування» «Системи обробки природно-мовної інформації»; для освітньо-кваліфікаційного рівня магістр: «Нейромережеві методи обчислювального інтелекту»; в держбюджетних науково-дослідних роботах ХНУРЕ, що підтверджено відповідними актами впровадження.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Бодянский, Е.В. Классификация текстовых документов с помощью нечеткой вероятностной нейронной сети / Е.В. Бодянский, Н.В. Рябова, О.В. Золотухин // Восточно-европейский журнал передовых технологий. – 2011. – № 6/2 (54). – С. 16-19. (Входить до міжнародних наукометричних баз: Ulrich's Periodicals Directory, DRIVER, Index Copernicus, WorldCat, Bielefeld Academic Search Engine (BASE), DOAJ).

2. Бодянский, Е.В. Классификация текстовых документов с помощью нейронной сети встречного распространения с контролируемым обучением / Е.В. Бодянский, Н.В. Рябова, О.В. Золотухин // Біоніка інтелекту – 2014. – № 1(82). – С. 3-7.

3. Золотухин, О.В. Классификация политематических текстовых документов с использованием нечетких нейро-сетевых технологий / О.В. Золотухин // Системи обробки інформації. – 2012. – № 9(107). – С. 101-105.

4. Бодянский, Е.В. Обработка текстовых документов с помощью адаптивного нечеткого обучаемого векторного квантования / Е.В. Бодянский, Н.В. Рябова, О.В. Золотухин // Вісник національного технічного університету «ХПІ». Нові рішення в сучасних технологіях. – 2011. – №. 53. – С. – 109-115. (Входить до міжнародних наукометричних баз: Ulrich's Periodical Directory, Directory of Open Access Journals, Google Scholar).

5. Золотухин, О.В. Нечеткая кластеризация и нечеткая классификация политематических текстовых документов / О.В. Золотухин // Вісник національного технічного університету «ХПІ». Нові рішення в сучасних технологіях. – 2012. – №. 68(974). – С. – 79-82. (Входить до міжнародних наукометричних баз: Ulrich's Periodical Directory, Directory of Open Access Journals, Google Scholar).

6. Золотухин, О.В. Система кластеризации текстовых ресурсов при помощи нейронечеткого подхода и семантического анализа / О.В. Золотухин // Шоста міжнародна науково-практична конференція «Математичне та програмне забезпечення інтелектуальних систем (MPZIS-2008)», 12-14 листопада 2008р.: тези доп. – Дніпропетровськ, 2008 – С. 134-135.

7. Золотухин, О.В. Нейро-фаззи подход к кластеризации текстовых документов / О.В. Золотухин, В.В. Волкова // Информационные интеллектуальные системы – 2008: сб. научн. трудов по материалам 1-й факультетской науч.-практ. молодежной школы-семинара: 2-4 декабря 2008 г. – Харьков, 2008. – С. – 15-18.

8. Золотухин, О.В. Нейро-нечеткая кластеризация текстовых документов / О.В. Золотухин // 13-й Международный молодежный форум «Радиоэлектроника и молодежь в XXI веке», 30 марта-1 апреля 2009г.: матер. конф. – Харьков, 2009. – С. 120.

9. Золотухин, О.В. Нейро-фаззи классификация политематических документов / О.В. Золотухин // Восьма міжнародна науково-практична конференція «Математичне та програмне забезпечення інтелектуальних систем (MPZIS-2010)», 10-12 листопада 2010р.: тези доп. – Дніпропетровськ, 2010 – С. 85-86.

10. Золотухин, О.В. Нечеткая нейросетевая модель классификации политематических документов / О.В. Золотухин // Перша науково-технічна конференція «Сучасні напрямки розвитку інформаційно-комунікаційних технологій та засобів управління», 13-14 грудня 2010р.: матер.конф. – Харків-Київ, 2010 – С. 43.

11. Золотухин, О.В. Классификация текстов на основе нейро-фаззи подхода / О.В. Золотухин // 14-й Международный молодежный форум «Радиоэлектроника и молодежь в XXI веке», 18-20 марта 2010г.: матер. конф. – Харьков, 2010. – С. 121.

12. Золотухин, О.В. Гибридный метод для классификации текстовых документов / О.В. Золотухин // Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального

інтелекту», 16-20 травня 2011р.: матер. конф. – Євпаторія, 2011. – Том 1. – С. 251-254.

13. Золотухін, О.В. Класифікація політематичних документів із застосуванням нейро-нечіткого підходу / О.В. Золотухін // 5-та міжнародна конференція молодих вчених «Комп`ютерні науки та інженерія (CSE-2011)», 24-26 листопада 2011р.: матер.конф. – Львів, 2011 – С. 184-186.

14. Золотухин, О.В. Интеллектуальная обработка текстовых документов с помощью адаптивного нечеткого векторного квантования / О.В. Золотухин // 16-й Международный молодежный форум «Радиоэлектроника и молодежь в XXI веке», 17-19 апреля 2012г.: матер. конф. – Харьков, 2012. – С. 47-48.

15. Золотухин, О.В. Класифікація політематичних текстових документів із застосуванням нейро-нечіткого підходу / О.В. Золотухин // Міжнародна наукова конференції «Наукова періодика слов`янських стран в умовах глобалізації». 10-12 жовтня 2012р.: матер. конф.: – Київ, 2012 – С. 17-18.

16. Золотухин, О.В. Нечеткие искусственные нейронные сети в задаче классификации политематических текстовых документов / О.В. Золотухин // VI Междунар. научн.-практич. конф. молодых ученых и студентов «Информационные процессы и технологии «Информатика-2013», 22-26 апреля 2013 г.: матер. конф. – Севастополь, 2013 – С. 72-74.

17. Золотухин, О.В. Нейронная сеть встречного распространения с контролируемым обучением для задач классификации политематических текстовых документов / О.В. Золотухин, М.О. Абросимов // 18-й Международный молодежный форум «Радиоэлектроника и молодежь в XXI веке», 14-16 апреля 2014г.: матер. конф. – Харьков, 2014. – Том 6.– С. 32-33.

18. Золотухин, О.В. Архитектура нейронной сети встречного распространения с контролируемым обучением в задачах классификации текстовых документов / О.В. Золотухин // 3-я международная научно-практическая конференция «Информационные системы и технологии (ИСТ-2014)», 15-18 сентября 2014г.: матер. конф.: – Харьков, 2014 – С. 35-36.

АНОТАЦІЯ

Золотухін О.В. Методи класифікації політематичних текстових документів із застосуванням нейро-фаззі технологій. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2015.

Дисертацію присвячено розробці методів класифікації політематичних текстових документів в режимі послідовної online обробки даних та за наявності класів, які перетинаються. Розглянуто задачу класифікації політематичних текстових документів, основні методи обробки документів та існуючі методи їх класифікації, визначено основні недоліки та переваги

розглянутих методів. Вперше запропоновано: архітектуру і метод навчання нечіткої ймовірнісної нейронної мережі, що дозволяє суттєво скоротити кількість нейронів в мережі, а відповідно і кількість параметрів, які налаштовуються, та підвищити швидкодію при нечіткій класифікації в online режимі надходження на обробку текстових документів; адаптивну нечітку нейронну мережу навчаємого векторного квантування та метод її навчання, що дозволяє виконувати нечітку класифікацію політематичних текстових документів, які в послідовному режимі надходять на вхід мережі; нейронну мережу зустрічного поширення з контрольованим навчанням з нелінійним вихідним шаром, яка дозволяє поліпшити якість класифікації за умов класів, що перетинаються.

Ключові слова: політематичні текстові документи, нечітка класифікація, штучні нейронні мережі, нечітке векторне квантування, методи навчання.

АННОТАЦИЯ

Золотухин О.В. Методы классификации политематических текстовых документов с применением нейро-фаззи технологий. – На правах рукописи.

Диссертация на соискание научной степени кандидата технических наук по специальности 05.13.23 - системы и средства искусственного интеллекта. - Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2015.

Диссертационная работа посвящена разработке методов нечеткой классификации политематических текстовых документов в online режиме.

В первом разделе приведен анализ существующих методов классификации текстовых документов, в том числе в последовательном режиме, определены основные преимущества и недостатки. Обоснована целесообразность разработки методов классификации политематических текстовых документов на основе технологий вычислительного интеллекта, которые позволяют выполнять классификацию политематических текстовых документов в последовательном режиме в условиях пересекающихся классов.

Во втором разделе впервые предложены архитектуры нечеткой вероятностной нейронной сети и нечеткой нейронной сети обучаемого векторного квантования и методов их обучения. В нечеткой вероятностной нейронной сети первый скрытый слой сформирован прототипами, что значительно сокращает число нейронов в сети. В нечеткой нейронной сети обучаемого векторного квантования содержится два слоя обработки информации, при этом нейроны первого скрытого слоя связаны между собой латеральными связями, с помощью которых реализуются процессы конкуренции. Исходной информацией для обучения является последовательность векторов-образов с известной классификацией, при этом входные сигналы предварительно нормируются. Нейроны первого скрытого слоя предназначены для нахождения прототипов (центроидов) классов, при этом их компоненты являются настраиваемыми синаптическими весами

нейрона. Нейроны выходного слоя вычисляют уровни принадлежности предъявленного образа к определенному классу.

Предложенная архитектура сети предназначена для решения задачи классификации политематических текстовых документов при последовательном режиме поступления с учетом наличия пересекающихся классов.

В третьем разделе диссертационной работы впервые предложена архитектура нейронной сети встречного распространения с контролируемым обучением для классификации политематических текстовых документов. В первом скрытом слое используется нетрадиционная самоорганизующаяся карта, которая обучается без учителя, а нейронная сеть векторного квантования, которая обучается с учителем, что позволяет повысить быстродействие, а, кроме того, нейронная сеть векторного квантования может оценить центроиды и границы классов. В выходном слое вместо звезд Гроссберга используются элементарные персептроны Розенблатта с нелинейной функцией активации (релейной), принимающей только два значения.

Проведено имитационное моделирование разработанных методов классификации политематических текстовых документов на основе нейро-фаззи технологий: нечеткой вероятностной нейронной сети; нейронной сети адаптивного нечеткого обучаемого векторного квантования; нейронной сети встречного распространения с контролируемым обучением. Показаны их преимущества перед известными архитектурами и методами обучения как по точности, так и по быстродействию в задачах последовательной нечеткой классификации политематических текстовых документов. Решены практические задачи: нечеткой классификации политематических текстовых документов при обработке результатов поиска в информационно-поисковой системе научной библиотеки, которая показала, что методы, которые предложены в работе, значительно улучшают процесс обработки пользователем результатов работы информационно-поисковой системы, а также сокращают время обработки результатов запросов; автоматической классификации полнотекстовых документов для диагностики у ургентных пациентов заболеваний внутрибрюшных органов, что позволяет сократить время постановки диагноза.

Ключевые слова: политематический текстовый документ, нечеткая классификация, искусственные нейронные сети, обучаемое векторное квантование, методы обучения.

ABSTRACT

Zolotukhin O.V. Classification methods of multi-topic text documents using neuro-fuzzy technologies . – Manuscript.

A thesis for the candidate degree in technical sciences in the specialty 05.13.23 – systems and tools of artificial intelligence. – Kharkiv National University

of Radio Electronics, Ministry of Education and Science, Kharkiv, 2015.

The thesis is devoted to methods classification of multi-topic text documents development in a sequential (online) mode under conditions of overlapping classes. The task of the multi-topic text documents classification, basic methods of document processing and existing classification methods, the main advantages and disadvantages of these methods have been discribed. A fuzzy probabilistic neural network's architecture and an adaptive neural network of fuzzy vector quantization and its learning method which provides fuzzy classification of multi-topic text documents in a sequential mode have been developed for the first time. The proposed method of neural network's learning is characterized by high speed and low computational complexity. A counter-propagation neural network's model of controlled studies has been proposed for the first time. A learning method for counter-propagation neural networks which provides a better classification for classes overlapping and increase the information processing speed has been developed.

Keywords: multi-topic text documents, fuzzy classification, artificial neural networks, learning vector quantization, learning methods.