алгоритмів.

Основною перевагою алгоритму LMS є гранична обчислювальна простота – для підстроювання коефіцієнтів фільтра на кожному кроці потрібно виконати N +1 пар операцій «множення-складання». Платою за простоту є повільна збіжність і підвищена дисперсія помилки в сталому режимі, що і збільшує рівень вихідного шуму (рис. 3).

Matlab Simulink – дуже потужний інструмент, який може використовуватися для моделювання в системах обробки сигналів. Шляхом побудови імітаційних моделей вдається зручно оцінювати, моделювати та досліджувати прикладні задачі використання адаптивних систем при вирішенні завдань ідентифікації, подавлення шумів та завад, вирівнювання каналу зв'язку тощо.

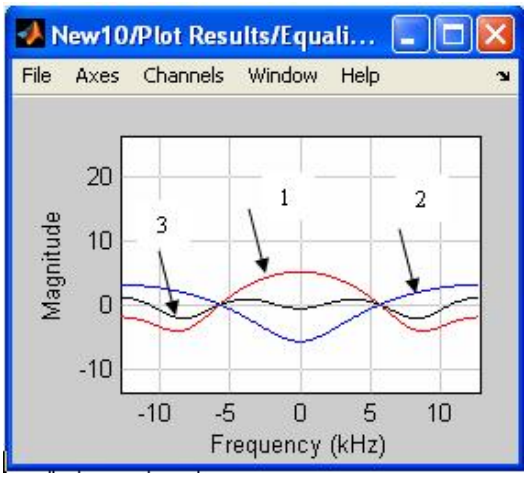

Рис. 6. Частотні відгуки в моделі системи зв'язку із адаптивним вирівнювачем

Література

1. Бойко Ю. М. Ідея адаптивної обробки сигналів /Ю. М. Бойко // Материалы XIII Международного молодежного форума ["Радиоэлектороника и молодежь в XXI веке"]. – Харків, 1.04. 2009. – С. 346.

2. Адаптивные фильтры / [под. ред. К.Ф.Н. Коуэна и П.М. Гранта]; [пер. с англ.]. – М: Мир, 1988. – 392 с.

3. Boyko J., Babiy J., Karpova L. Conceptual Features of Application of Facilities of Adaptive Filtration are in the Tasks of Authentication of Noise of Communication Channels / J. Boyko, J. Babiy, L. Karpova // Proceeding of the Xth International Conference TCSET 2010. – Lviv – 23.02.2010. P. 299.

Надійшла 25.3.2011 р.

УДК 004.9: 355

О.С. АНДРОЩУК

Національна академія Державної прикордонної служби України ім. Богдана Хмельницького, м. Хмельницький

НЕЙРОМЕРЕЖНІ МОДЕЛІ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ

Надано нейромережні моделі на базі топологій багатошарового перцептрону і мережі Кохонена для визначення класів за змістом текстових документів, які застосовуються у діяльності Державної прикордонної служби України. Подання тексту здійснено на підставі моделі терм-документ. Розроблені моделі дають більшу точність та повноту результатів порівняно зі статистичними класифікаторами.

Neurons networks models are represented on the base of topologies of MLP and networks of SOM for determination of classes on maintenance texts documents which are used in activity of Government boundary service. Presentation of text is carried out on the basis of model therm is document. The developed models give greater exactness and plenitude of results as compared to statistical classifiers.

Ключові слова: нейронна мережа, текстовий документ, класифікація.

Вступ

Постановка проблеми. На сьогодні велика увага приділяється підвищенню рівня інтелектуальності різного роду автоматизованих систем, дослідженню і розробці методів і засобів подання знань, отриманню оптимальних рішень на їх основі. Це повною мірою стосується завдання автоматичної класифікації текстів, актуальність якої підвищується по мірі впровадження і розвитку інформаційних технологій.

Розвиток не тільки глобальних комп'ютерних мереж, але й повнотекстових баз даних призвів до постійного нарощування інформаційних текстових ресурсів. При постійному й інтенсивному зростанні обсягів текстової інформації труднощі пошуку необхідних відомостей серед множини доступних текстів значно зменшують її цінність. Тому особливу значущість автоматична класифікація текстових документів має для інформаційно-пошукових систем глобальних мереж, повнотекстових баз даних. Виходячи з цього, завдання автоматичної класифікації тексту, будучи окремим випадком завдання розпізнавання змісту, є на сьогодні актуальною проблемою, що стосується різних сфер людської діяльності, оскільки її вирішення надасть можливість повністю автоматизувати процес обробки, класифікації і пошуку інформації.

Аналіз останніх досліджень та публікацій. Існує безліч підходів до вирішення завдання автоматичної обробки, розпізнавання і класифікації текстової інформації, проте увага, що приділяється цій проблемі, однозначно свідчить, що жоден з них не є вичерпним. Найбільш часто для розпізнавання і класифікації (або генерації) мови використовуються статистичні методи. До них належать статистичні класифікатори на основі ймовірнісних методів, методи багатовимірного статистичного аналізу, зокрема, факторного аналізу, кластерного аналізу, таксономії, розпізнавання образів без вчителя, частотний аналіз

тексту тощо [1– 3].

У деяких сферах (наприклад, лексико-граматичному аналізі речень, синтаксичному аналізі речень, автоматичному реферуванні) більшою мірою застосовуються лінгвістичні методи. Часто в системах автоматичного реферування застосовуються як лінгвістичні, так і статистичні методи.

Однак, недоліками зазначених вище підходів є: відсутність у моделі відомостей щодо структури й системи зв'язків реального об'єкта, що вносить суб'єктивізм у вибір як самої моделі, так й її структури; недостатня точність класифікації; значна чутливість отриманих результатів до недостатньої інформації та (або) її зашумленість; залежність результату класифікації від кваліфікації аналітика в конкретній предметній сфері.

Метою статті є розробка та дослідження нейромережних моделей визначення класів текстових документів, які застосовуються у діяльності Державної прикордонної служби України (ДПСУ).

Основний розділ

Для застосування математичних методів з подальшою автоматизацією це завдання можна подати як класифікацію описів службових ситуацій (СС), які виникають у службовій діяльності ДПСУ. Тобто необхідно визначити, до якого з відомих класів СС належать об'єкти (неформалізовані текстові описи на природній мові), що досліджуються. Тоді формалізація завдання класифікації СС здійснюється таким чином. Є множина об'єктів T – текстових документів та множина $C=\{c_i\}\ i=1..N_c$ – СС, яка складається з N_c класів об'єктів. Кожний клас c_i подано деяким описом F_i, що має деяку внутрішню структуру. Процедура класифікації f об'єктів $t \hat{I} T$ полягає у виконанні перетворень над ними, після яких робиться або висновок про відповідність t одній зі структур F_i, що означає віднесення t до класу c_i, або висновок щодо неможливості класифікації t. Стосовно текстових описів елементами множини T є електронні версії текстових документів.

На підставі наведеного вище загальну модель класифікатора можна подати у такому вигляді [1]:

$$R=<T, C, F, R_c, f>, \quad (1)$$

де T – множина текстів, які підлягають класифікації;

C – множина класів ОС;

F – множина описів;

R_c – відношення на $C\times F$;

f – операція класифікації виду $T\rightarrow C$.

Нейронні мережі можуть застосовуватися при вирішенні багатьох завдань обробки інформації, зокрема в завданнях класифікації. Як відомо, штучний нейрон виконує такі перетворення вхідного вектора $X=\{x_i\}$: $y=f_a\ (NET)$; $NET=Sw_ix_i$, де w_i – ваговий вектор нейрона (ваги синаптичних зв'язків), NET – результат зваженого підсумовування, f_a – нелінійна функція активації нейрона. У термінах класифікатора (1) X відповідає внутрішнім описам $\{F_i\}$, а функції NET і f_a – компоненти процедури класифікації f. Функціональність нейрона є простою, тому для вирішення конкретних завдань нейрони об'єднуються в мережі. Навчання класифікатора за умови, що є вибраними топологія мережі і функція активації f_a, зводиться до підбору вагових коефіцієнтів кожного нейрона. У цій роботі розглядається застосування двох топологій: багатошарового перцептрону і мережі Кохонена.

Способи подання тексту. Нейронні мережі пристосовані обробляти лише інформацію, подану числовими векторами, тому для їх застосування в обробці текстів на природній мові (ТПМ) тексти необхідно подавати у векторному вигляді. У роботі подання тексту здійснено на підставі моделі терм-документ [4]. У цій моделі текст описується лексичним вектором $\{w_i\}\ i=1..N_w$, де w_i – важливість (інформативна вага) терміну в документі, N_w – повна кількість термінів у документальній базі (словнику). Вага терміну, відсутнього в документі, приймається рівною 0. Для зручності ваги нормуються, так що $w_i\in$ [0,1]. У роботі використовувалися дискретні значення: присутній термін у тексті має вагу 1, а відсутній – вагу 0. Перевагами цієї моделі є:

- можливість обліку морфології, коли всі форми одного слова відповідають одному терміну;
- можливість обліку синонімії: слова-синоніми оголошуються одним терміном словника;
- можливість обліку стійких словосполучень: як термін може виступати не окреме слово, а декілька зв'язаних слів, які утворюють єдине поняття.

Недоліки виокремимо таке:

- за відсутності простої додаткової обробки, такої як морфологічний аналіз, суттєво знижується якість класифікатора, оскільки різні форми одного слова вважаються різними термінами; разом із тим, морфологічний аналіз – дуже нетривіальне завдання, що вимагає для вирішення залучення лінгвістів;
- розмірність векторів $\{w_i\}$ залежить від загальної кількості термінів у навчальній вибірці текстів, що в реальних завданнях призводить до необхідності розробляти альтернативні структури даних, відмінні від векторів;
- словник термінів може не охоплювати всіх документів, які підлягають класифікації. Аналізовані документи можуть містити значущі терміни, які не увійшли до навчальної вибірки, що негативно позначається на адекватності моделі.

Багатошаровий перцептрон БШП (англомовний термін – MLP). Базова архітектура БШП включала три шари: N-вхідний шар нейронів, кількість входів дорівнює кількості ознак-термінів (дескрипторів); вихідний шар, що складається з M нейронів згідно з кількістю класів розбиття; проміжний шар – для визначення кількості нейронів у ньому використовувалося евристичне правило, виходячи з якого було визначено число, що дорівнює $(N+M)/2$ нейронів. База даних була розбита на навчальну контрольну і

тестову вибірки; співвідношення між навчальною і контрольною вибіркою складало 3: 1, приклади, що залишилися, були використані як тестові. Навчання мережі проводилося з використанням нейромережного пакету STATISTIKA Neural Networks за алгоритмом Back Propagation (зворотне розповсюдження помилки). Як функція помилки використовувалася середньоквадратична помилка. Як пороги прийняття/відкидання були прийняті значення 0,95/0,5. Навчання мережі потребувало близько 500 ітерацій. У результаті роботи такої мережі було одержано такі характеристики якості її роботи: помилка на навчальній, контрольній і тестовій множинах склала 0,00988, 0,10550 і 0,06545 відповідно. Необхідно відзначити, що використання деяких прийомів, які надають можливість покращити ефективність ШНМ, таких як включення додаткових нейронів у проміжний шар, ослаблення порогів прийняття/відкидання, не принесло бажаних результатів. Отже, можна зробити висновок про те, що для отримання задовільних результатів автоматичної класифікації є необхідною більша кількість даних для навчання.

Мережа Кохонена (англомовний термін – SOM). Призначення мережі Кохонена [6] – розділення векторів вхідних сигналів на групи, тому можливість подання текстів у вигляді векторів дійсних чисел надасть можливість застосовувати цю мережу для їх класифікації. Мережа складається з одного шару, що має форму прямокутних грат для *g*-х зв'язаних нейронів і форму соти для *h*-и зв'язаних.

Вектори *X*, що аналізуються, подаються на входи всіх нейронів. За наслідками навчання геометрично близькі нейрони виявляються чутливими до схожих вхідних сигналів, що може бути використано в завданні класифікації таким чином. Для кожного класу визначається центральний нейрон і довірча область навколо нього. Критерієм межі довірчої області є відстань між векторами сусідніх нейронів і відстань до центрального нейрона області. При подачі на вхід навченої мережі вектора тексту активізуються деякі нейрони (можливо з різних областей), текст належить до того класу, у довірчій області якого активізувалась найбільша кількість нейронів і якомога ближче до її центру. Алгоритм навчання мережі полягає в наступному. Усі вектори повинні лежати на гіперсфері одиничного радіусу. Задається міра сусідства нейронів, що надає можливість визначати зони топологічного сусідства в різні моменти часу. На рис. 1 показано зміну цієї величини $NE_j(t)$ для деякого *j*-го нейрона.

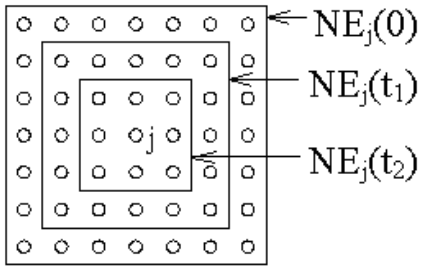

Рис. 1. Зони топологічного сусідства на мапі ознак

Крім того, задається розмір грати і розмірність вхідного вектора, а так само визначається міра подібності векторів *S*. Далі виконуються такі кроки для кожного вектора навчальної вибірки:

1. Початкова ініціалізація площини може бути проведена, наприклад, довільним розподілом вагових векторів на гіперсфері одиничного радіусу.

2. Мережі подається вхідний вектор тексту X_u, обчислюється міра подібності $S(X, W_j)$ для кожного *j*-го нейрона мережі. Нейрон, для якого S_j є максимальною, вважається поточним центром і для нього визначається зона сусідства $NE_j(t)$.

3. Для всіх нейронів, що потрапляють у зону $NE_j(t)$ (див. рис. 1), проводиться корекція ваг за правилом $w_{ij}(t+1) = w_{ij}(t) + l\,(x_i(t) - w_{ij}(t))$, де λ – крок навчання, що зменшується з часом. Величина $NE_j(t)$ зменшується з часом так, що спочатку вона охоплює всю мережу, а в кінці навчання зона звужується до одного-двух нейронів, коли λ також достатньо мале.

Як свідчать експерименти, на навчання мережі Кохонена впливає наступне:

1. Кількість нейронів та їх розміщення. Кількість нейронів слід вибирати не менше, ніж кількість груп, які потрібно одержати. Розташування нейронів на двовимірній площині залежить від завдання, що вирішується. Як правило, вибирається або квадратна матриця нейронів, або прямокутна з відношенням сторін, близьким до одиниці.

2. Початковий стан. У цьому випадку застосовується ініціалізація випадковими значеннями. Це не завжди призводить до бажаних результатів. Один із можливих варіантів покращання цього – обчислення характеристичних векторів репрезентативної вибірки текстів, що визначають межу двовимірної площини проекції. Після цього вагові вектора нейронів рівномірно розподіляються в одержаному діапазоні.

3. Характер зміни топологічної зони сусідства $NE_j(t)$. Визначає область нейронів, які підлягають навчанню. Чим швидше скорочуватиметься ця область, тим більше класів буде утворено, тим більшою є точність і меншою повнота.

4. Тип даних, що подаються на вхід. Для лексичних векторів фактично проводиться обробка по наявних в документі термах, що дає достатньо добрі результати. У цьому випадку можна виокремлювати документи за специфікою словарного набору. Проте без застосування морфологічного аналізу цей метод неможливо застосовувати, оскільки різко збільшується обчислювальна складність.

5. Послідовність подачі на вхід векторів документів із різних груп. Оскільки коефіцієнт швидкості навчання з часом змінюється, результати подачі на вхід різних векторів текстів виявляються різними. При великому початковому значенні λ відбувається інтенсивна модифікація всіх нейронів навколо переможця. При випадковій подачі документів із різних груп області близькості утворюються рівномірно.

Моделювання проводилося для баз текстових документів двох типів, які стосуються надзвичайних ситуацій природного та техногенного походження (171 документ, 270 термінів) у діяльності ДПСУ. У

найкращому випадку класифікація здійснює абсолютно правильне розбиття документів на дві групи: один клас відповідає природнім, інший – техногенним надзвичайним ситуаціям.

Результати застосування різних класифікаторів надано у табл. 1.

Таблиця 1

Експериментальне дослідження класифікаторів

Вид класифікатора / Показник	Статистичні класифікатори	Класифікатори, які засновані на функціях подібності	Нейромережні класифікатори	
			Багатошаровий перцептрон	Мережа Кохонена
Точність (*V*)	0,25	0,43	0,68	0,89
Повнота (*U*)	0,34	0,45	0,75	0,81

Висновки

Отже, у роботі надано класифікатори текстових документів на підставі нейромережних моделей. Проведено їх дослідження для здійснення класифікації текстових документів, які свідчать, що розроблені моделі дають більшу точність та повноту результатів порівняно зі статистичними класифікаторами.

Наступним кроком дослідження є обґрунтування та вибір алгоритмів класифікації, їх можлива модифікація, що найкращим чином буде відповідати умовам автоматизації текстових документів.

Література

1. Классификация и кластер: [сб. ст. / под ред. Дж. В. Райзина]. – М.: Мир, 1980.

2. Кочетков П. А. Краткий курс теории вероятностей и математической статистики: [учебное пособие] / Кочетков П. А. – М.: МГИУ, 1999. – 51 с.

3. Елисеева И. И. Общая теория статистики: [учебник] / И. И. Елисеева, М. М. Юзбашев; под ред. И. И. Елисеевой. – [5-е изд., перераб. и доп.]. – М.: Финансы и статистика, 2004. – 656 с.

4. Солтон Дж. Динамические библиотечно-поисковые системы / Солтон Дж. – М.: Мир, 1979.

5. Осовский С. Нейронные сети для обработки информации / С. Осовский; [пер. с польского И. Д. Рудинского]. – М.: Финансы и статистика, 2002. – 344 с.

6. Круглов В. В. Искусственные нейронные сети / В. В. Круглов, В. В. Борисов. – М.: Горячая линия – Телеком, 2002. – 382 с.

Надійшла 19.3.2011 р.

УДК 004.413.5; 004.416.6

В.С. ЯКОВИНА, Я.М. ЧАБАНЮК, М.М. СЕНІВ, У.Т. ХІМКА
Національний університет "Львівська політехніка"

ОЦІНЮВАННЯ ТА ПРОГНОЗУВАННЯ НАДІЙНОСТІ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ НА ОСНОВІ МОДЕЛІ З ІНДЕКСОМ СКЛАДНОСТІ ПРОЕКТУ

В роботі описано процес оцінювання та прогнозування надійності програмного забезпечення на основі моделі з індексом складності програмного продукту. За статистичними характеристиками опису експериментальних даних різними моделями встановлено, що модель з індексом складності продукту більш адекватно описує реальні експериментальні дані ніж S-подібна модель та модель Goel– Okumoto. Встановлено, що залежності параметрів моделі для різних тестових профілів є близькими між собою і виявляють однакову поведінку, що підтверджує ефективність використання цих параметрів для критерію достатності процесу тестування та визначення точки переходу вибірки вхідних даних до пуассонового розподілу.

This paper describes the process of evaluation and prediction of software reliability based on a model with software project complexity index. According to the statistical characteristics of experimental data description by the different models it has been determined that the model with product complexity index more adequately describes the real experimental data than the S-shaped model and the Goel-Okumoto model. It has been determined that the dependencies of the model parameters for different test profiles are close to each other and show the same behavior. It confirms the effectiveness of using these parameters for the testing process adequacy criterion and for the determination of the transition point of sample input data to Poisson distribution.

Ключові слова: надійність програмного забезпечення, життєвий цикл програмного забезпечення, моделювання надійності, програмна інженерія.

Вступ

Підвищення рівня складності сучасної техніки та все ширше її використання в усіх галузях людської життєдіяльності висувають все вищі вимоги до її надійності та експлуатаційної безпеки. Широке розповсюдження різноманітних програмно-апаратних комплексів та обчислювальних систем, що заміняють