LNCS 765, Springer-Verlang, 1994, pp. 293 – 304. ***6.*** *B. Preneel, Cryptographic Hash Function, Kluwer Academic Publishing, 1995(to appear).*

УДК 681.3.06

ВИКОРИСТАННЯ НЕЙРОННОЇ МЕРЕЖІ КОХОНЕНА ДЛЯ РОЗПІЗНАВАННЯ СПАМУ

Ігор Терейковський
Державний університет інформаційно-комунікаційних технологій

Аннотация: ***Досліджена можливість використання сучасних модифікацій нейронної мережі Кохонена в системах захисту електронної пошти від спаму. Проведено адаптацію архітектури нейронної мережі типу пружинна карта для кластеризації електронних листів. Обґрунтовані принципи визначення та попередньої обробки вхідних параметрів.***
Summary: ***The opportunity of use of modern modifications of the neuron network Kohonena in systems of protection of electronic mail from spama is investigated. The adaptation of architecture of neuron network of a type a spring card for clusterisation of the electronic letters is carried out. The principles of definition and initial processing of entrance parameters are developed.***
Ключевые слова: ***Електрона пошта, спам, нейронні мережі, карта Кохонена, пружинна карта.***

Електронна пошта є одним із найбільш поширених та важливих сервісів як локальних комп'ютерних мереж, так і глобальної мережі Інтернет. Її роль пояснюється тим, що вона використовується не тільки для доставки приватних повідомлень, але й як важливий компонент системи електронного документообігу практично всіх підприємств та організацій. З цієї причини надійність та безпечність функціонування електронної пошти важлива як для приватних, так і для корпоративних користувачів. При цьому задача захисту електронної пошти від спаму на сьогодні є далекою від вирішення. Відзначимо, що в загальному випадку під терміном спам звичайно розуміють масово розповсюджувані листи, зміст яких носить рекламний або шахрайський характер. Як правило такі листи анонімні, а більшість користувачів не давали своєї згоди на отримання цієї розсилки. По даним [1, 2] спам складає близько 70 – 80 % від всього обсягу електронних листів в російськомовній зоні Інтервент. Основною причиною існування спаму є впевненість багатьох комерційних установ, що це є найбільш ефективним видом реклами товарів та послуг. Крім того, все частіше в спамі зустрічаються різноманітні обманні пропозиції. Наприклад, це може бути об'ява про зміну доменної адреси електронного магазину, в якому користувач може розрахуватись за допомогою безготівкового платежу по мережі Інтернет. Зрозуміло, що адреса є фіктивною, а користувач переказавши гроші товар не отримує. Такий вид спаму отримав назву “фішінг”.

Процес розповсюдження спаму є досить прибутковим бізнесом та поставлений на професійну основу. Розроблені та широкодоступними є спеціальні програмні засоби, що дозволяють оминаючи антиспамовий захист провести розсилку декількох мільйонів спам-листів протягом 2 – 3 годин. Наприклад, на Веб-сайтах *inattack.ru та www.izone.ru* розміщені, хоча і під іншою назвою та в дещо спрощеному варіанті, інсталяційні пакети прикладних програм для розсилки спаму. Тому масовими розсилками електронних листів все частіше починають займатись не тільки професійні спамери, але й звичайні користувачі Інтернет.

Якщо не враховувати психологічний аспект проблеми, то для кінцевого користувача основним негативним наслідком отримання спаму є збільшення обсягу нецільових листів, перегляд та класифікація яких потребує значного часу. Результати [1, 2] вказують, що не зважаючи на відповідний захист співробітники великих комерційних організацій, які відповідають за електронну пошту витрачають на обробку спаму до 30% свого робочого часу. Водночас збільшується Інтернет трафік та дисковий простір на поштових серверах та робочих станціях. Це вказує на актуальність проведення досліджень в напрямку застосування нових підходів для боротьби проти спаму.

І Недоліки загальнопоширених методів розпізнавання спаму

Сучасні методи боротьби зі спамом можливо розділити на законодавчі, організаційні та програмно-технічні. Розглянемо останні. Вони функціонують за принципом: розпізнання спаму – блокування (знищення) спаму. Відзначимо, що блокування розпізнаного спаму не викликає труднощів. При цьому задача розпізнавання спаму потребує доопрацювання [1 – 3], хоча це і суперечить рекламним заявам провідних виробників антиспамових засобів. Так в [1] зазначено, що термін блокування нового виду спаму

провідними поставниками поштових послуг в мережі Інтернет становить близько 20-30 хвилин. Проте при загальновідомій швидкості розсилки 2000000 – 3000000 листів за годину багато користувачів все ж отримають спам. Крім того, засоби захисту поставників поштових послуг орієнтовані в основному на розпізнавання масових розсилок, а практичний досвід свідчить, що обсяг сучасних розсилок спаму може знаходитись в межах 100 – 200 адресатів. Це може бути запрошення на тренінг чи семінар або реклама товару в межах окремого міста. Однак, велика кількість організацій, що розсилає такі запрошення та рекламні об'яви призводить до значного загального обсягу спаму.

Розглянемо недоліки загальнопоширених методів розпізнавання спаму. Метод чорного, білого та сірого списків. Метод базується на аналізі зворотної IP-адреси листа. Всі листи, відправлені з IP-адрес, занесених в чорний список, знищуються ще на поштовому сервері. Адреса вноситься в чорний список на основі висновку користувача, що, лист є спамом. Від адресатів із білого списку отримання поштових повідомлень дозволено. В випадку, коли IP-адреса листа відсутня як в чорному, так і в білому списках, відправнику автоматично надсилається запит на авторизацію, а IP-адреса заноситься в тимчасовий сірий список. Вважається, спамер не буде надсилати підтвердження про відправку свого листа. Тому, якщо протягом визначеного терміну підтвердження про відправку листа від невідомого відправника не надходить, то його адреса заноситься в чорний список, а повідомлення знищується. Основний недолік даного методу полягає в тому, що IP-адреса не обов'язково вказує на джерело спаму. Наприклад, спам може прийти з динамічної IP-адреси, або розсилка здійснена без відома власника IP-адреси. Використання сірого списку доцільне тільки при невеликому обсязі листування з обмеженим колом осіб. В протилежному випадку ведення сірого списку потребує великих затрат на періодичну переконфігурацію. Крім того, сучасні засоби розповсюдження спаму дозволяють генерувати підтвердження відправки спам-листа.

Листи класифікуються як спам, якщо обсяг відправки електронної пошти з однієї адреси (з однієї підмережі) за короткий термін часу перевищує певну граничну величину, наприклад, 100000 листів за годину. Недоліками методу є необхідність контролю за практично всім простором поштових відправлень Інтернет, що потребує значних затрат. Крім того метод неефективний при невеликих спам-розсилках.

Метод розпізнання спаму по ключовим словам (словосполученням), які визначаються користувачем у вигляді набору правил. Даний метод не знайшов широкого розповсюдження через складнощі при формуванні вказаних правил.

Метод байєсовської фільтрації. Кожному слову або тегу HTML, що зустрічається в електронній переписці, присвоюється два значення: ймовірність його присутності в спамі (z) та ймовірність його присутності в звичайних листах ($1-z$). Величину z називають спамоцінкою слова. Для кожного нового листа за допомогою формули Байєса розраховується загальна спам-оцінка листа (Z):

$$Z = \frac{\prod_{i=1}^{N} z_i}{\prod_{i=1}^{N} z_i + \prod_{i=1}^{N} (1 - z_i)}, \qquad (1)$$

де N – кількість слів в листі.

Відзначимо, що в деяких антиспамових засобах застосовуються більш складні вирази розрахунку спам-оцінки. При цьому ймовірності z визначаються за допомогою спеціальних словників, або/та в процесі статистичного аналізу листів конкретного користувача. Якщо величина Z менша від деякого граничного значення (Δ), то лист класифікується як спам:

$$Z < \Delta, \qquad (2)$$

Як основний недолік байєсовської фільтрації є недостатня адекватність виразу (2) процесу розпізнавання спаму. Одним із наслідків цього є висока ймовірність пропуску спаму, якщо в листі мало слів з високою спам-оцінкою. Вказана обставина використовується спамерами для обходу та компрометації захисту. Так для обходу захисту рекламні листи модифікуються за рахунок використання слів синонімів та словосполучень ідентичних за змістом, але різних за набором слів. Скомпрометувати захист може безглуздий лист, що складається з набору нейтральних слів. Таким чином, жоден із існуючих методів розпізнавання не дозволяє адекватно реагувати на сучасні методи формування та розповсюдження спам-листів. В той же час навіть некваліфікований користувач легко проведе розпізнавання на основі співставлення своїх інтересів зі змістом листа. Тому в [3] пропонується методика автоматичного розпізнавання спаму за допомогою ймовірнісних та семантичних нейронних мереж (НМ), функціонування яких багато в чому повторює інтелектуальну діяльність людини. Ще одним типом НМ, застосування якого може підвищити ефективність розпізнавання спаму, є мережа (карта) Кохонена. На відміну від ймовірнісних та семантичних цей тип НМ пристосований не тільки для автоматичної класифікації, але й

для представлення значного обсягу образів у вигляді, зручному для їх класифікації користувачем. Потенційно це дозволяє використати як управляючий елемент людину-користувача, що може значно підвищити ефективність розпізнавання спаму.

II Загальні принципи функціонування НМ типу карта Кохонена

Архітектура та функціонування мережі Кохонена базується на моделі Ліппмана-Хемінга, яка дозволяє класифікувати піддослідний образ ξ як один із бібліотечних образів (класів) [4]. В моделі Ліппмана-Хемінга кожен з образів однозначно описується відповідним K-вимірним вектором з бінарними компонентами, а критерієм класифікації є відстань Хеммінга від ξ до бібліотечних образів. Класифікація полягає в пошуку найближчого до ξ бібліотечного образу. Відстань між векторами розраховується відповідно до правила Хеммінга як кількість неоднакових компонент образів. На рис. 1 показана структура НМ Ліппмана-Хемінга для розпізнавання образу, що характеризується двома параметрами, як одного із трьох бібліотечних класів.

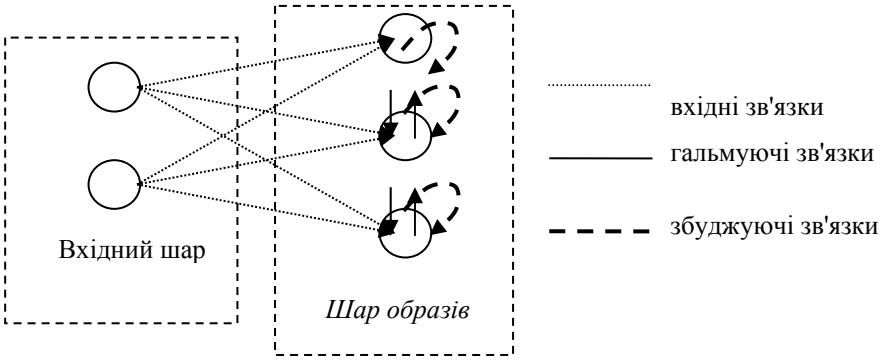


Рисунок 1 – Приклад структури НМ Ліппмана-Хеммінга

Структурно модель складається із вхідного шару нейронів та шару образів. В задачу вхідних нейронів входить лише розподіл вхідної інформації між нейронами шару образів. Кількість вхідних нейронів дорівнює кількості компонент в образі, тобто кількості параметрів, що характеризують образ. Кожен вхідний нейрон пов'язаний з кожним нейроном шару образів, кількість яких дорівнює кількості кластерів. Нейронам шару образів відповідає функція активації виду:

$$F(NET) = \begin{cases} NET, \exists\, NET > 0 \\ 0, \exists\, NET \le 0 \end{cases}, \tag{3}$$

де NET – вхід нейрону, що в загальному випадку розраховується за формулою:

$$NET = \sum_{i=1}^{I} x_i w_i , \tag{4}$$

де I – кількість вхідних зв'язків нейрону, x_i – величина i-го зв'язку, w_i – вага i-го зв'язку.

Процес навчання НМ Ліппмана-Хемінга полягає в тому, що вагові коефіцієнти нейронів шару образів встановлюються рівними нормованим компонентам бібліотечних образів:

$$w_n^m = \frac{x_n^m}{\sum_{i=1}^{K} x_i^m}, \tag{5}$$

де w_n^m – ваговий коефіцієнт n-го входу для m-го нейрону шару образів, x_n^m – n-та компонента для m-го нейрону (образу), K – кількість компонент в образі.

Компоненти невідомого образу також нормуються:

$$s_n = \frac{\xi_n}{\sum_{i=1}^{K} \xi_n}, \tag{6}$$

де s_n – нормована *n*-та компонента вхідного вектора ξ.

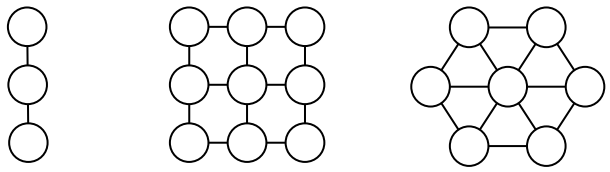
Подача невідомого образу призводить до початкових рівнів активації нейронів шару образів:

$$y^m(0) = f\left(0{,}5\sum_{n=1}^{M}\left(s_n w_n^m\right)\right), \tag{7}$$

де $y^m(0)$ – початковий рівень активації *m*-го нейрону шару образів, *M* – кількість бібліотечних образів, *f* – функція активації нейрону шару образів виду (3).

Після цього відбувається ітераційний процес вибору нейрону, який є найближчим до невідомого образу. Вибір реалізується за рахунок того, що гальмуючими зв'язками кожен з нейронів отримує негативне збудження від всіх інших нейронів. Величина негативного збудження від будь-якого нейрону пропорційна величині активації цього нейрону. Водночас кожен з нейронів отримує позитивне збудження від самого себе. Після деякої кількості ітерацій зостається єдиний активний нейрон-переможець, що вказує на клас, до якого належить піддослідний образ. Такий механізм вибору нейрону отримав назву "переможець забирає все"[4].

Як і мережа Ліппмана-Хеммінга НМ (карта) Кохонена складається із двох шарів нейронів, вхідного і вихідного (топографічного). Кількість нейронів вхідного шару дорівнює кількості компонент вхідних образів. Кожен вхідний нейрон пов'язаний з кожним топографічним нейроном, який відповідає певному класу образів. Принциповою відмінністю від моделі Ліппмана-Хеммінга, де кожен клас характеризується заданими зовні параметрами одного бібліотечного образу, в карті Кохонена кожному класу, як правило, відповідають декілька бібліотечних образів. Розділ на класи здійснюється в процесі навчання та полягає в розрахунку вагових коефіцієнтів зв'язків топографічних нейронів. Кількість класів є параметром зовнішнім відносно мережі і визначається точністю, з якою необхідно виконати класифікацію набору бібліотечних образів. Навчання карти Кохонена відбувається методом "без вчителя", за допомогою механізму "переможець забирає все". Для підкреслення того, що межі класів визначаються самою мережею, замість терміну класифікація використовують термін кластеризація. На противагу моделі Ліппмана-Хеммінга, де положення нейрона-переможця в шарі образів не мало нічого спільного з координатами його вагових коефіцієнтів у вхідному просторі, в карті Кохонена близьким нейронам топографічного шару відповідають близькі вхідні образи. Для виявлення кореляції між топографічними нейронами карта Кохонена отримала деякі особливості як в структурі, так і в алгоритмі навчання. Особливості структури полягають в тому, що сітка зв'язків між нейронами побудована за певними правилами. На рис. 2 показано лінійну (а), квадратну (б) та гексагональну (в) сітки зв'язків.



а) **б)** **в)**

Рисунок 2 – Загальнопоширені сітки зв'язків між нейронами в топографічному шарі карти Кохонена

Навчання карти Кохонена відбувається методом послідовних наближень. Починається навчання з призначення матриці вагових коефіцієнтів випадкових значень. Після цього на вхід мережі послідовно подаються вектори, що відповідають навчальним образам. Для кожного з векторів розраховується відстань від нього до кластерного елементу:

$$d_m^j = \sqrt{\sum_{i=1}^{N}\left(w_i^m(t) - x_i\right)^2}\,,\ w_i^m \in W^m\,, \tag{8}$$

де d_m^j – відстань від *j*-го вхідного вектору до *m*-го нейрону, *t* – номер епохи навчання, *N* – кількість компонент вхідного вектору, $w_i^m(t)$ – вага зв'язку між *i*-м входом та *m*-м нейроном на *t*-ій епосі навчання, W^m – матриця вагових коефіцієнтів *m*-го нейрону, x_i – *i*-а компонента вхідного вектора.

Знаходиться нейрон, для якого ця відстань мінімальна, тобто нейрон-переможець. Після цього змінюються вагові коефіцієнти нейрону-переможця та сусідніх з ним нейронів:

$$w_i^k(t+1) = w_i^k(t) + \eta(t) \times (x_i - w_i^k(t)), \tag{9}$$

$$\begin{cases} w_i^n(t+1) = w_i^n(t) + \eta(t,n) \times (x_i - w_i^n(t)) \\ n \in \{k - r(t), \ldots k + r(t)\} \wedge n \in \{m\} \end{cases}, \tag{10}$$

де k – номер нейрону-переможця, η – коефіцієнт (норма) швидкості навчання, n – номер сусіднього нейрону, r – радіус навчання, $\{m\}$ – множина нейронів в топографічному шарі.

Відзначимо, що сусідніми вважаються нейрони, які мають між собою зв'язки, що входять в коло з центром, що відповідає нейрону-переможцю та заданим радіусом навчання. Наприклад при радіусі 1 для лінійної структури зв'язків будуть змінюватись вагові коефіцієнти нейрону-переможця, та двох найближчих до нього нейронів. Для квадратної структури крім нейрону-переможця змінюються вагові коефіцієнти чотирьох найближчих нейронів, а для гексагональної структури – шести найближчих нейронів. З кожною епохою швидкість і радіус навчання зменшуються, що призводить до точної настройки топографічної карти. Часто навчання спеціально розділяють на дві фази. Перша коротка фаза характеризується великою швидкістю та радіусом навчання. Друга фаза довга, характеризується малою швидкістю навчання та близьким до нуля радіусом. Рекомендована тривалість першої фази навчання становить біля 1000 епох, а тривалість другої фази 10000-100000 епох [4]. Загальна кількість навчальних епох має бути як мінімум в 10 разів більша, ніж кількість навчальних прикладів. В ідеальному варіанті зупинка навчання відбувається тоді, коли навчальні вектори не змінюють своїх кластерів при переході від однієї епохи до іншої. Інколи, в складних випадках, навчання закінчують після певної кількості епох. В результаті положення кожного центру кластеру встановлюється в деякій позиції, яка найкраще відповідає тим навчальним прикладам, для яких нейрон є переможцем. При цьому мережа організуються таким чином, що нейрони, які відповідають образам, розміщеним близько в просторі входів, будуть розміщені близько один від другого і на топографічній карті. Якщо в навчальному наборі даних були характерні точки з призначеними їм маркерами, тоді відповідні вузли карти також можуть бути маркірованими. Крім того, мережа вузлів може бути різнокольоровою, відповідно деякої ознаки, що дозволяє будувати трьохвимірні карти. Популярним способом показу на картах самих даних є застосування діаграм Хінтона, які передбачають зображення на кожному вузлі карти квадрату, розмір якого пропорційний кількості навчальних образів, найближчих до цього вузла. Після закінчення навчання на вхід мережі можна подавати нові образи для розпізнавання. При цьому можливо застосовувати так званий поріг доступу, який дорівнює максимальному рівню активації нейрона-переможця. Якщо рівень активації нейрона-переможця для класифікуємого образу нижчий ніж вказаний поріг, то класифікація проведена успішно. В випадку рівня активації вищого за поріг доступу, вважається, що мережа не прийняла рішення про класифікацію. Це може відбутись тоді, коли образ, що класифікується, значно відрізняється від навчальної вибірки.

На сьогодні існує досить багато різних модифікацій карти Кохонена. Серед них однією з найбільш досконалих є пружинна карта (ПК) [5]. Характерною рисою ПК є аналогія топографічного шару з пружною пластиною, що при лінійних та кутових деформаціях намагається зберегти свою початкову форму. Завдяки цьому вузли НМ з однієї сторони будуть притягуватись до точок даних, а з іншої намагатимуться мінімізувати своє розтягнення та прийняти максимально гладку форму (стати більш регулярними). Для реалізації такої аналогії в ПК використовується критерій оптимізації структури НМ, який представляє собою суму середнього квадрату відстані до вузла мережі та коефіцієнтів пружності з відповідними ваговими коефіцієнтами мережі.

Розглянемо алгоритм побудови прямокутної пружної сітки, наведений в [5]. Нехай p кількість вузлів сітки по горизонталі, q кількість вузлів по вертикалі. Пронумеруємо вузли мережі за допомогою індексів – $y_{i,j}$, i=1, **…,** p, j=1, **...,** q. Розділимо всю множину вхідних даних X на $p \times q$ підмножин (таксонів) $K_{i,j}$ (i=1,**…,** p, j=1, **…,** q), в межах кожної із яких точки знаходяться ближче до вузла мережі $y_{i,j}$, ніж до будь-якого іншого вузла:

$$K_{i,j} = \{x \in X, \ \forall \left| y^{i,j} - x \right|^2 \rightarrow \min_{i,j}\}. \tag{11}$$

При формуванні (11) як міри близькості вузлів мережі до даних використано величину середнього квадрату відстані від точки до найближчого вузла сітки. Кожен вузол (крім граничних) має чотирьох сусідів, з кожним із яких він з'єднаний відповідним зв'язком – ребром мережі. Чим більша середня довжина вузла, тим сильніше мережа розтягнута, що зумовлює необхідність мінімізації цієї величини. Таким чином, в мінімізуємий функціонал повинні ввійти різниці між положеннями сусідніх вузлів. Ступінь згину можливо визначити за допомогою точної оцінки величини другої похідної. В результаті критерій оптимізації можливо записати таким чином:

$$D = \frac{D_1}{N} + \lambda \times \frac{D_2}{p \times q} + \mu \times \frac{D_3}{p \times q} \to \min, \tag{12}$$

де N – кількість точок навчальної вибірки, X, λ та μ – коефіцієнти пружності, що відповідають за лінійну та кутову деформацію мережі, D_1 – міра близькості розміщення вузлів мережі до даних, D_2 – міра розтягнутості мережі, D_3 – міра кривизни мережі.

Складові D_1, D_2, D_3 розраховуються так:

$$D_1 = \sum_{i,j} \sum_{X_k \in K_{i,j}} \left\| X_k - y^{i,j} \right\|^2, \tag{13}$$

$$D_2 = \sum_{i=1}^{p} \sum_{j=1}^{q-1} \left\| y^{i,j} - y^{i,j+1} \right\|^2 + \sum_{i=1}^{p-1} \sum_{j=1}^{q} \left\| y^{i,j} - y^{i+1,j} \right\|^2, \tag{14}$$

$$D_3 = \sum_{i=1}^{p} \sum_{j=1}^{q-1} \left\| 2y^{i,j} - y^{i,j-1} - y^{i,i+1} \right\|^2 + \sum_{i=1}^{p-1} \sum_{j=1}^{q} \left\| 2y^{i,j} - y^{i-1,j} - y^{i+1,j} \right\|^2, \tag{15}$$

де $K_{i,j}$ – підмножина точок із множини X, для яких вузол мережі є найближчим (таксони).

Межі сумування в (14, 15) вибрані таким чином, щоб в функціоналах D_2 та D_3 ребро входило в суму тільки один раз. В [5] відзначено, що функціонал D є квадратичним щодо положення вузлів $y_{i,j}$, що дозволяє при заданому розділу множини точок на таксони провести його мінімізацію шляхом вирішення системи рівнянь розміром $pq \times pq$. Тому, для розрахунку мінімального значення функціоналу D необхідно:

- розмістити вузли мережі довільним чином;
- при заданих положеннях вузлів мережі провести поділ множини даних на таксони – $K_{i,j}$;
- при заданому поділі множин точок на таксони провести мінімізацію функціоналу D;
- етапи 1 та 2 слід повторювати доки функціонал D не перестане змінюватись в межах заданої точності.

Кількість операцій (ξ), необхідних для навчання карти Кохонена та ПК, можливо оцінити так:

$$\xi = k \times L_w \times P = k \times N_1 \times N_0 \times P, \tag{16}$$

де L_w – кількість синаптичних зв'язків N_1 – розмірність вхідного сигналу, N_0 – розмірність вихідного сигналу, P – кількість навчальних прикладів, k – константа.

Невелика кількість ітерацій дозволяє швидко перенавчати карту Кохонена та ПК, що зумовлює високу адаптивність цих НМ до зміни умов застосування.

III Застосування карти Кохонена та пружинної карти для класифікації електронних листів

Відповідно до [4, 5] алгоритм застосування карти Кохонена та ПК для вирішення конкретної задачі складається з таких етапів:

1. визначення номенклатури та допустимих величин вхідних параметрів;
2. підготовка навчальної вибірки;
3. нормування вхідних параметрів навчальної вибірки;
4. визначення кількості кластерів;
5. визначення виду сітки зв’язків між нейронами шару розпізнавання.
6. вибір параметрів навчання НМ;
7. навчання НМ;
8. візуалізація та верифікація результатів навчання;
9. якщо результати не задовільні необхідно провести навчання з новими параметрами НМ. Для цього повторити п. 4 – 8.

Таким чином, для розв'язання практичної задачі необхідно сформувати множину вхідних параметрів (п. 1 – 3), розробити архітектуру НМ (п. 4 – 6) та провести його навчання (п. 7 – 9).

Розглянемо приклад класифікації різнотематичних електронних листів. Як статистичний матеріал було використано 100 електронних листів за такими тематиками, як запрошення на різнопланові семінари, реклама побутових послуг та реклама промислових товарів. Листи були отримані автором протягом декількох тижнів 2006 року. Можна вважати, що на практиці листи однієї із вказаних тематик є цільовими, а інші листи – спам. Тому класифікація листів за тематиками в першому наближенні адекватна розпізнаванню спаму.

Основою формування множини вхідних параметрів НМ послужила методика [5, 6], яка передбачає:

1. формування з усіх піддослідних текстів словника інформативних слів; в словник не включаються малозначущі слова та слова-зв'язки;
2. приведення слів до деякої канонічної форми; причиною цього є те, що в українській та російській мовах більшість слів можуть бути представлені в декількох словоформах;
3. призначення кожному слову в канонічній формі порядкового номера;
4. розрахунок для кожного із текстів частоти зустрічі в ньому кожного з визначених в п. 2 слів:

$$\mu_i^j = \frac{n_i^j}{N_i}, \qquad (17)$$

де μ_i^j – частота зустрічі канонічної форми *j*-го слова в *i*-му тексті, n_i^j – кількість всіх словоформ *j*-го слова в *i*-му тексті, N_i – кількість всіх слів в і-му тексті.

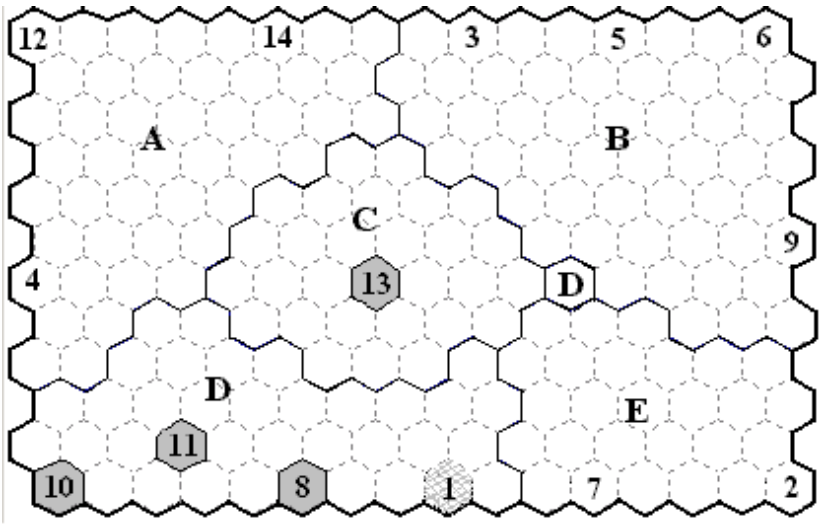
Таким чином, кількість слів словника дорівнює кількості вхідних параметрів НМ. Попередній аналіз статистичного матеріалу виявив, що кількість слів в канонічній формі в представлених електронних листах перевищує 1000. Можна зробити припущення, що збільшення тематик електронної кореспонденції призведе до пропорційного збільшення кількості слів і відповідно до (16) негативно вплине на швидкість навчання НМ. Ще однією передумовою зменшення кількості слів є обмеження доступних інструментальних засобів моделювання НМ. Так при моделюванні карти Кохонена використовувався пакет прикладних програм Deductor Studio 4.3, а при моделюванні ПК пакет – ViDaExpert, в яких кількість вхідних параметрів досить обмежена. Тому є сенс в попередній обробці електронних листів з метою зменшення кількості слів, що аналізуються звісно без втрати основного змісту тексту. Для цього можливо застосувати так зване реферування текстів [6]. Як інструментальний засіб реферування було використано пакет TextAnalyst 2.01, який дозволяє якісно реферувати тексти російською, українською та англійською мовами. Приклад реферату електронного листа з запрошенням на семінар з теми "Кодекс адміністративного судочинства України" показаний на рис. 3

Ідеологія Кодексу адміністративного судочинства.
На що забувають звертати увагу суди і учасники адміністративного процесу.
Завдання адміністративного судочинства.
Проблеми визначення юрисдикції адміністративних судів:
- справи адміністративної юрисдикції (поняття публічно-правового спору, суб'єкт владних повноважень);
- адміністративна юрисдикція і справи про адміністративні правопорушення.
Проблеми визначення підсудності справи (предметної, територіальної та інстанційної):
- порядок створення адміністративних судів;
- визначення підсудності після створення адміністративних судів.
Провадження в адміністративній справі першої інстанції, особливості.
Строки для звернення до адміністративного суду.

Рисунок 3 – Приклад реферату електронного листа

В результаті реферування загальна кількість слів була зменшена приблизно в 7 разів, при задовільній як для системи захисту якості змісту реферату. Однак реферування листа призвело і до негативного наслідку – втрати важливих атрибутів листа (заявленої тематики, зворотної адреси і т. і.). Аналіз кореспонденції показує, що вказані атрибути в основному представлені в перших 10 та останніх 30 словах листа. Тому при формуванні словника було використано, крім реферату, ще й вказані слова. Аналогічну обробку пройшли і інші 49 піддослідних листів. В результаті кількість слів в канонічній формі була зменшена до 108. Відзначимо, в багатьох випадках листи однієї тематики не значно відрізнялись між собою. Наприклад, було отримано 12 листів з запрошенням відвідати семінар з теми "Логістика". Різниця між листами полягала тільки в даті проведення семінару, а перелік інформативних слів залишився незмінним. Зрозуміло, що з точки зору розпізнавання спаму означені листи повинні відноситись до одного класу. Тому листи з однаковим набором інформативних слів були виділені в окремі групи. Темі «реклама побутових послуг» відповідає група листів №1, темі «запрошення на семінари» відповідають групи листів №2, 3, 4, 5, 6, 7, 9, 12, 13, 14, темі «реклама промислових товарів» – № 8, 10, 11. Частоти інформативних слів для кожної із груп листів були розраховані за допомогою (17) та використані як вхідні дані карти Кохонена та ПК.

При побудові карти Кохонена прийнято: форма сітки зв'язків – прямокутник (16×12), форма сітки зв'язків – гексагон, кількість кластерів – 5, кількість навчальних епох – 500, $\eta = 0.1$, $r = 6$ на початку навчання, $\eta = 0.005$, $r=1$ в кінці навчання. Розділена на кластери карта Кохонена представлена на рис. 4.

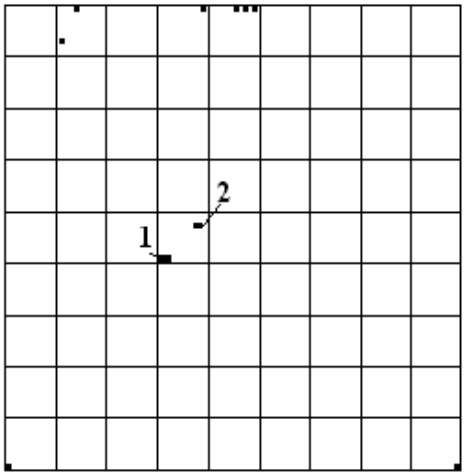


A, B, C, D, E – номери кластерів, 1, ... , 14 – номери груп листів з однаковим набором інформативних слів, ——— межі кластерів, - - - - - - межі комірок карти.

Рисунок 4 – Карта Кохонена

На рис. 4 номери 8, 10, 11, 13 відповідають групам листів з теми реклами промислових товарів, а номер 1 відповідає листам з теми реклам побутових послуг. Всі інші листи є запрошеннями на семінари. Можна зробити висновок, що карта Кохонена якісно розділила листи на дві основні теми – реклама (кластери C та D) та запрошення на конференції (кластери A, B, E). Проте не достовірно віднесла до одного кластеру листи з рекламою промислових товарів та листи з рекламою побутових послуг. Зазначимо, що якість відображення однотипних листів за допомогою класичної карти Кохонена дозволяє провести їх приблизну класифікацію самим користувачем.

При побудові ПК прийнято: форма сітки зв'язків – квадрат (9×9), форма сітки зв'язків – квадрат, кількість кластерів – 10. Навчання ПК було розділене на дві фази. Тривалість першої фази 500 навчальних епох, тривалість другої – 3000 навчальних епох. На початку навчання $\eta = 0.07$, $r = 10$, в кінці навчання $\eta = 0.003$, $r = 1$. Відображення багатовимірної площини навчальних даних на площину головних компонент параметрів навчальних даних показано на рис. 5. Кожній точці на рис. 5 відповідає своя група листів.



1 – листи з теми реклами промислових товарів, 2 – листи з теми реклами побутових послуг

Рисунок 5 – Відображення багатовимірної множини навчальних даних на площину

На рис. 5 видно, що групи листів з теми «реклама промислових товарів» та «реклама побутових послуг» відображаються досить компактно. Це вказує на можливість їх швидкої та якісної класифікації користувачем шляхом візуального аналізу ПК. Зазначимо, на рис. 5 межі кластерів не показані через те, що кластери ПК зайняли несуміжні комірки і їх відображення ускладнює процес візуального аналізу. Однак в автоматичному режимі ПК достовірно розділила листи на три відповідні групи (теми).

IV Висновки

Підвищити ефективність розпізнавання спаму можливо за рахунок використання в антиспамових засобах блоку автоматизованої класифікації листів за допомогою карт Кохонена та ПК.

Перспективним шляхом підвищення рівня захисту електронної пошти є адаптація існуючих методів реферування текстів до використання в системах розпізнавання спаму.

Література: ***1.*** *Цветков В. Я., Булгаков С. В. Спам и некоторые методы борьбы с ним. // http://vio.fio.ru.* ***2.*** *Спам 2004: аналитический отчет – http://www.ashmanov.com.* ***3.*** *Терейковский И. А. Применение семантического анализа содержимого электронных писем в системах распознавания спама / Защита информации – 2006.– № 4, с. 49-60.* ***4.*** *Ежов А. А., Шумский С. А. Нейрокомпьютинг и его применения в экономике и бизнесе / М.: МИФИ, 1998. – 224 с.* ***5.*** *Зиновьев А. Ю. Визуализация многомерных данных / М.: СК Пресс, 2005. - 180 с.* ***6.*** *Заболеева-Зотова А. В. Естественный язык в автоматизированных системах. Семантический анализ текстов: Монография / ВолгГТУ. – Волгоград 2002. – 228 с.*

УДК 004.056.55 (076.5)

РЕАЛІЗАЦІЯ ЗАХИСТУ ІНФОРМАЦІЇ В КОМ'ЮТЕРНИХ СИСТЕМАХ ТА МЕРЕЖАХ НА ОСНОВІ ОПЕРАЦІЙНОЇ СИСТЕМИ FREEBSD

Богдан Корнієнко, Леонід Щербак
Національний авіаційний університет

Анотація: **Методологія курсу "Захист інформації в комп`ютерних системах та мережах", що складається із лекцій та лабораторних робіт, має за мету дати студентам фахові знання з основ захисту інформаційної взаємодії у комп`ютерних мережах при їх підключенні до відкритих комунікацій. Лабораторні роботи реалізуються на базі FreeBSD – технологічно зрілої та досконалої операційної системи, що відображено у її стійкості, захищеності та підтримці галузевих стандартів.**

Summary: **The methodology of a rate « Protection of the information in computer systems and networks », that will consist of lectures and laboratory works, has for the purpose to give students a professional knowledge on bases of protection of information interaction in computer networks. Laboratory works are realized on base FreeBSD - technologically and perfect operational system that is displayed in her stability, security and support of branch standards.**

Ключові слова: **Захист інформації, комп`ютерні мережі, операційна система FreeBSD.**

Вступ

Підготовка у вищих навчальних закладах України фахівців за освітнім напрямом «Інформаційна безпека» та розробка методології ряду навчальних дисциплін цього напряму є актуальною задачею, що викликає активну дискусію серед професорсько-викладацького складу та фахівців із захисту інформації. Про це свідчить значна кількість публікацій [1 – 5]. Дана робота присвячена розробці методики викладання курсу "Захист інформації в комп`ютерних системах та мережах" на основі використання операційної системи FreeBSD.

Методологія курсу "Захист інформації в комп`ютерних системах та мережах", що складається з лекцій та лабораторних робіт, має за мету дати студентам фахові знання з основ захисту інформаційної взаємодії у комп`ютерних мережах при їх підключенні до відкритих комунікацій. Курс охоплює основні методи та засоби міжмережного екранування для захисту локальних мереж від несанкціонованого доступу, базові протоколи безпеки та засоби побудови захищених віртуальних мереж. Лабораторні роботи реалізуються на базі FreeBSD – технологічно зрілої та досконалої операційної системи, що відображено у її стійкості,