Список літератури

1. Нікітченко М. С. Математична логіка та теорія алгоритмів / М. С. Нікітченко, С. С. Шкільняк. – К. : ВПЦ Київський університет, 2008. – 528 с.
2. Шкільняк С. С. Секвенційні числення композиційно-номінативних логік квазіарних предикатів / С. С. Шкільняк // Пробл. програмування. – 2012. – № 2–3. – С. 33–43.
3. Шкільняк С. С. Секвенційні числення логік часткових та неоднозначних квазіарних предикатів / С. С. Шкільняк // Вісник Київського ун-ту. Серія: фіз.-мат. науки. – 2012. – Вип. 4. – С. 231–236.
4. Шкільняк С. С. Секвенційні числення реномінативних логік квазіарних предикатів / С. С. Шкільняк // Наукові записки НаУКМА. – 2012. – Т. 138. Комп'ютерні науки. – С. 23–29.
5. Шкільняк С. С. Відношення логічного наслідку в композиційно-номінативних логіках / С. С. Шкільняк // Пробл. програмування. – 2010. – № 1. – С. 15–38.
6. Нікітченко М. С. Логіки квазіарних предикатів кванторно-еквaційного рівня / М. С. Нікітченко, С. С. Шкільняк // Пробл. програмування. – 2012. – № 4. – С. 19–34.

S. Shkilniak

SEQUENT CALCULI FOR PURE FIRST-ORDER LOGICS OF QUASI-ARY PREDICATES

We construct special sequent calculi for pure first-order composition-nominative logics of partial single-valued, total multiple-valued and partial multiple-valued quasi-ary predicates with using of special variable definedness predicates. Such calculi are proposed for logics of quantifier level and for logics of quantifier-equational level. The soundness and completeness theorems for the introduced calculi are proved.

Keywords: logic, predicate, logical consequence. sequent calculi.

Матеріал надійшов 20.11.2012

УДК 519.95

Порхун О. В.

ВСТАНОВЛЕННЯ ДІАГНОЗУ ДЕРМАТОЛОГІЧНИХ ЗАХВОРЮВАНЬ ІЗ ЗАСТОСУВАННЯМ МОДЕЛЕЙ РОЗПОДІЛЕНОГО ВИХІДНОГО КОДУ ТА ПЕРСЕПТРОНУ

У статті розглянуто моделі розподіленого вихідного коду для вирішення задачі мультикласифікації з використанням нейронної мережі багатошаровий персептрон та описано застосування розробленої системи мультикласифікації з реалізацією описаних моделей для вирішення задачі визначення діагнозу захворювання пацієнтів в області дерматології.

Ключові слова: мультикласифікація, розподілений вихідний код, вичерпний код, матриця кодових слів, багатошаровий персептрон.

Вирішення проблем класифікації у різних предметних областях вимагає досліджень з розробки нових ефективних методів, алгоритмів та систем. Для багатьох практичних задач число класів об'єктів, які класифікуються, більше, ніж 2. Зокрема, у задачах медичної діагностики числом класів є кількість можливих діагнозів в області захворювання, при розпізнаванні рукописних символів кількість класів задається відповідно до розмірності алфавіту, що використовується, у задачах класифікації текстів класи задаються відповідно до обраного профілю класифікації, наприклад за тематикою, стилем написання (авторством) тощо.

Існуючі методи, які добре зарекомендували себе при вирішенні задач бінарної класифікації, далеко не всі можуть бути переналаштовані на випадок мультикласифікації. Нейронні мережі типу персептрон вирішують задачі мультикласифікації за допомогою використання вихідного шару нейронів, кількість яких задає число класів. Отже, у статті ми розглядаємо застосування різних моделей розподіленого вихідного коду *(distributed output code)* та багатошарового персептрону для вирішення задачі встановлення діагнозу дерматологічних захворювань.

Моделі розподіленого вихідного коду

Згідно з методом розподіленого вихідного коду, описаного в [1, с. 264], кожний клас задається бінарним рядком довжини n, «кодовим словом». Біти кодового слова відповідають окремим бінарним класифікаторам, які навчаються. Кожний бінарний класифікатор f_i може бути навчений розпізнаванню об'єктів більш ніж одного класу. В процесі навчання для прикладу класу i бажані відповіді даних n бінарних класифікаторів визначаються кодовим словом для класу i. Таким чином, для кожного класу формується кодове слово і, в результаті, будується матриця, рядки якої – це кодові слова, а стовпчики – бінарні класифікатори.

Після навчання новий об'єкт x класифікується оцінюванням кожного з n бінарних класифікаторів для отримання n-бітового кодового слова. Отримане кодове слово порівнюється з кожним із k кодових слів та об'єкт x належить класу, чиє кодове слово є найближчим згідно з обраною метрикою.

Моделі розподіленого вихідного коду будуються відповідно до різних представлень матриці кодових слів. Зокрема, в [2, с. 1007–1008] описано використання двох представлень матриці кодових слів: $M \in \{-1,1\}^{N_c \times n}$ і $M \in \{-1,0,1\}^{N_c \times n}$, де M – матриця кодових слів, N_c – кількість класів, n – кількість бінарних класифікаторів, тобто довжина кодового слова. Перше представлення називається *«один проти всіх» ("one-against-all")* і полягає у тому, що кожний бінарний класифікатор f_i навчається для розпізнавання об'єктів лише одного класу, для об'єктів інших класів бажана відповідь класифікатора f_i дорівнює –1. У другому представленні матриці M, що називається *«всі пари» ("All-Pairs")*, значення *0* означає, що об'єкти класу, якому воно відповідає, ігноруються класифікатором при навчанні, тобто навчальна вибірка формується лише з об'єктів класів, яким відповідають значення *1* і *–1*.

Для обох представлень матриці M при тестуванні нового об'єкту отримане для нього кодове слово порівнюється з усіма кодовими словами (рядками) матриці M та визначається номер «найближчого» до нього кодового слова.

Визначення мінімальної відстані між отриманим кодовим словом для об'єкту, що класифікується, та одним з кодових слів матриці при мультикласифікації розглядається як *процес декодування*. В [2, с. 1007] пропонується використання відстані Хемінга для реалізації процесу декодування. Зокрема, мінімальна відстань між отриманим кодовим словом $f(x) = (f_1(x), f_2(x), ..., f_n(x))$ і кодовими словами матриці M визначається за формулою:

$$y = \arg\min_r (d_H(M(r,.), f(x))), \quad (1)$$

$$d_H(M(r,.), f(x)) = \sum_{s=1}^{n} \frac{1 - sign(M(r,s) f_s(x))}{2}, \quad (2)$$

де $M(r,.)$ – позначає кодове слово r матриці M; s – біт кодового слова; y – клас, до якого належить об'єкт x.

Від вигляду матриці кодових слів залежить скільки помилок здатен виправити даний розподілений вихідний код у процесі декодування [1, с. 266]. Мірою якості розподіленого вихідного коду є мінімальна відстань Хемінга між кожною парою кодових слів. Якщо мінімальна відстань Хемінга дорівнює d, відповідний код може виправити до $\left\lfloor \frac{d-1}{2} \right\rfloor$ помилкових біт при декодуванні. Як описано в [1, с. 269], для забезпечення здатності виправлення помилок розподілений вихідний код повинен задовільняти властивостям:

1) роздільність рядків матриці – кожне кодове слово повинно мати достатню Хемінгову відстань до кожного з інших кодових слів;

2) роздільність стовпчиків – кожний бінарний класифікатор f_i повинен бути некорельований з класифікаторами f_j, навченими для інших бітових позицій, $j \neq i$. Це досягається за рахунок забезпечення великих значень Хемінгової відстані між стовпчиком i та іншими стовпчиками та Хемінгової відстані між стовпчиком i та доповненням до кожного з інших стовпчиків.

Зокрема, для числа класів $3 \le k \le 7$ в [1, с. 270] пропонується метод вичерпних кодів (*Exhaustive codes*). Згідно з цим методом рядками матриці кодових слів є кодові слова довжини $2^{k-1}-1$. Перший рядок матриці заповнюється одиницями, далі i-й рядок матриці заповнюється 2^{k-i} нулями та 2^{k-i} одиницями, що чергуються, починаючи з нулів. Приклад матриці з вичерпним кодом для 4 класів наведено у таблиці 1.

Таблиця 1. **Вичерпний код для 4-х класів**

Клас	Кодові слова						
	f_1	f_2	f_3	f_4	f_5	f_6	f_7
C_1	1	1	1	1	1	1	1
C_2	0	0	0	0	1	1	1
C_3	0	0	1	1	0	0	1
C_4	0	1	0	1	0	1	0

Побудова матриці кодових слів, яка б задовільняла вищенаведеним умовам 1–2, для великого числа класів є відкритою дослідницькою проблемою. Зокрема, для випадку $8 \le k \le 11$ в [1, с. 270–272] пропонується метод вибору стовпчиків з вичерпних кодів (*Column Selection from Exhaustive Codes*) і для $k > 11$ – метод рандомізованого пошуку екстремуму (*Randomized Hill Climbing*) та метод BCH-кодів.

У даній роботі реалізовано 3 моделі розподіленого вихідного коду з представленнями матриці кодових слів *"one-against-all"*, *"All-Pairs"* та вичерпний код для вирішення задачі визначення діагнозу ериматозно-плоскоклітинних захворювань. Для останньої моделі в матриці кодових слів замість нулів використовуються –1, оскільки згідно з моделлю вичерпних кодів зразки всіх класів повинні брати участь у навчанні кожного з бінарних класифікаторів; значення 0 відповідно до моделі *"All-Pairs"* вказує, що зразок даного класу не бере участі у навчанні.

Набір даних

Як базу даних для навчання та тестування розробленої системи мультикласифікації, було використано набір даних, зібраних в області дерматології, Dermatology Data Set з UCI Machine Learning Repository. Дана вибірка містить 34 атрибути [3].

Визначення діагнозу ериматозно-плоскоклітинних захворювань є важливою проблемою в області дерматології. Існуючі діагнози мають такі клінічні ознаки, як еритема та лущення з дуже незначними відмінностями, тому відрізнити випадки захворювань є надзвичайно складною проблемою. Захворюваннями у такій групі є псоріаз, себорейний дерматит, червоний плоский лишай, рожевий лишай, хронічний дерматит та висівкоподібний лишай (lichen pilaris). Для постановки діагнозу, звичайно, необхідна біопсія, однак такі захворювання мають багато спільних гістопатологічних особливостей. Іншою перешкодою у постановці діагнозу є те, що хвороба може проявити ознаки одного захворювання на початковій стадії та мати інші характеристики на наступних етапах.

У вибірку даних Dermatology Data Set включено виміри 12 ознак, згрупованих як клінічні атрибути, а також виміри 22 гістопаталогічних ознак, отриманих шляхом аналізу зразків шкіри під мікроскопом. Усі атрибути з вибірки Dermatology Data Set описано в [3].

Мітками класів (діагнозів) захворювань є значення від 0 до 5: 0 – псоріаз; 1 – себорейний дерматит; 2 – червоний плоский лишай; 3 – рожевий лишай; 4 – хронічний дерматит; 5 – висівкоподібний лишай (lichen pilaris).

Система мультикласифікації

Розроблена система мультикласифікації реалізує вищеописані моделі *"one-against-all"*, *"All-Pairs"* і вичерпний код для вирішення задачі встановлення діагнозу дерматологічних захворювань; для декодування використовується відстань Хемінга [1]. Також у даній системі окремо реалізовано вирішення задачі з використанням багатошарового персептрону, де кількість класів задається числом нейронів у вихідному шарі.

При реалізації вичерпного коду для цієї задачі мінімальна відстань Хемінга між кодовими словами матриці дорівнює 8, отже, даний код дає можливість виправлення до трьох помилкових біт.

Як класифікатори у реалізаціях моделей *"one-against-all"*, *"All-Pairs"* та вичерпний код, використовується багатошаровий персептрон з єдиним нейроном у вихідному шарі.

Для вирішення задачі визначення діагнозу ериматозно-плоскоклітинних захворювань систему мультикласифікації навчено та протестовано відповідно до кожної з чотирьох реалізацій.

Підготовка даних

Для прискорення процесу навчання методом зворотного розповсюдження помилки вхідні вектори та значення початкових згенерованих ваг та порогів мережі були пронормовані на діапазоні *[0,1]* і до отриманих значень застосовано додаткове нормування – зміщення середнього таким чином, щоб середнє значення дорівнювало нулю.

Вибірку даних поділено на навчальну та тестову у відношенні 80 : 20 %: навчальна вибірка складається з 284 прикладів, тестова – з 73 прикладів. Кількість прикладів у класах відповідних до діагнозів захворювань наведено у таблиці 2 для навчальної вибірки та у таблиці 3 для тестової вибірки.

Таблиця 2. **Розподіл зразків по класах у навчальній вибірці**

Діагноз (клас)	Кількість зразків
Псоріаз (0)	83
Себорейний дерматит (1)	49
Червоний плоский лишай (2)	56
Рожевий лишай (3)	39
Хронічний дерматит (4)	41
Висівкоподібний лишай (5)	16

Таблиця 3. **Розподіл зразків по класах у тестовій вибірці**

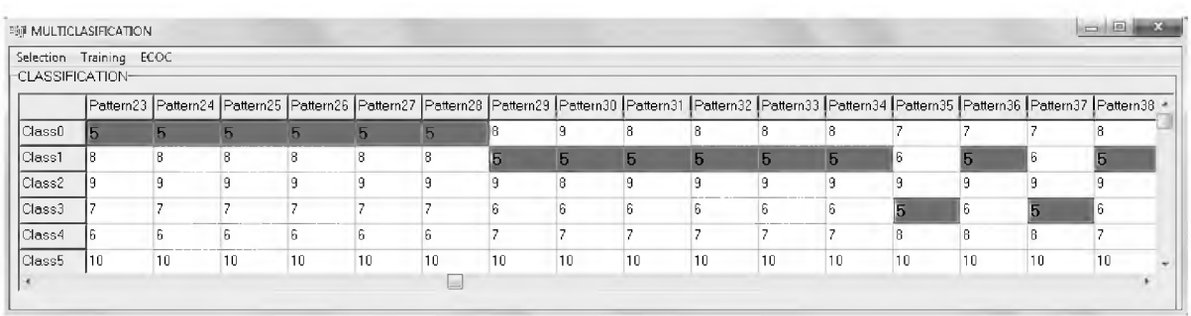
Діагноз (клас)	Кількість зразків
Псоріаз (0)	28
Себорейний дерматит (1)	10
Червоний плоский лишай (2)	15
Рожевий лишай (3)	9
Хронічний дерматит (4)	7
Висівкоподібний лишай (5)	4

Параметри системи мультикласифікації

Для навчання бінарних класифікаторів у реалізаціях розподіленого вихідного коду та багатошарового персептрону окремо застосовано метод зворотного розповсюдження помилки (Backpropagation learning algorithm) з фіксованим параметром швидкості навчання, що задається в опціях налаштування системи. При навчанні використовується послідовний режим, тобто корекція ваг мережі проводиться після надходження кожного прикладу. При цьому реалізовано можливість випадкової подачі прикладів у мережу, що дозволяє зробити пошук у просторі ваг стохастичним і, в свою чергу, зводить до мінімуму зупинку алгоритму у точці деякого локального мінімуму. Момент інерції при переобчисленні ваг не використовується. Критерієм зупинки навчання є досягнення заданої точності ε (тобто коли значення енергії середньоквадратичної помилки мережі не перевищує ε), або виконання заданої кількості епох навчання. Епохою вважається один повний цикл подачі набору прикладів. При завершенні навчання за кількістю епох та недосягненні заданої точності ε будуть збережені ваги мережі, що відповідають останньому локальному мінімуму помилки. При навчанні системи використано $\varepsilon = 10^{-5}$ та кількість епох, що дорівнює 4000.

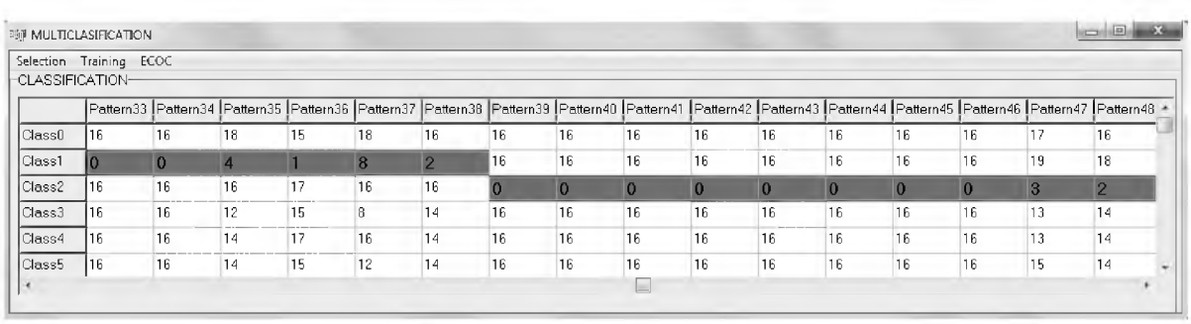
Структура нейронної мережі задається користувачем, який вибирає кількість шарів та кількість нейронів у шарі. Рекомендоване значення кількості нейронів у прихованих шарах дорівнює подвійному значенню розмірності вхідного вектора. У реалізації системи для визначення діагнозу ериматозно-плоскоклітинного захворювання сформовано структуру мережі *34–68–6* для персептрону з вхідним шаром з 34 нейронів (34 ознаки), прихованим шаром з 68 нейронів та вихідним шаром, нейрони якого відповідають 6 класам (діагнозам); для класифікаторів у розподіленому вихідному коді *"one-against-all"* та *"All-Pairs"* структура мережі має вигляд: *34–68–1*.

При навчанні мережі використовувалася логістична функція активації $f(x) = \frac{1}{1+e^{-ax}}$ з параметром $a=1$.



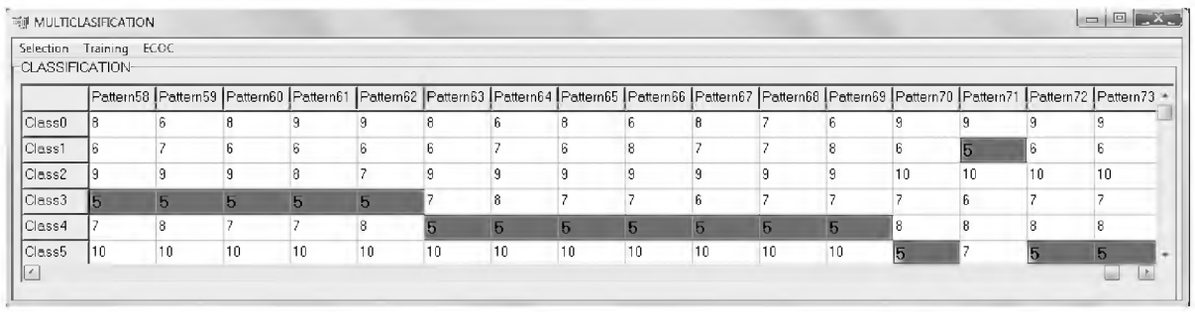
	Pattern23	Pattern24	Pattern25	Pattern26	Pattern27	Pattern28	Pattern29	Pattern30	Pattern31	Pattern32	Pattern33	Pattern34	Pattern35	Pattern36	Pattern37	Pattern38
Class0	5	5	5	5	5	5	8	9	8	8	8	8	7	7	7	8
Class1	8	8	8	8	8	8	5	5	5	5	5	5	6	5	6	5
Class2	9	9	9	9	9	9	9	8	9	9	9	9	9	9	9	9
Class3	7	7	7	7	7	7	6	6	6	6	6	6	5	6	5	6
Class4	6	6	6	6	6	6	7	7	7	7	7	7	8	8	8	7
Class5	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Рис. 1. Фрагмент результатів застосування моделі "All-Pairs" з відхиленням зразків класу 1 від правильного діагнозу



	Pattern33	Pattern34	Pattern35	Pattern36	Pattern37	Pattern38	Pattern39	Pattern40	Pattern41	Pattern42	Pattern43	Pattern44	Pattern45	Pattern46	Pattern47	Pattern48
Class0	16	16	18	15	18	16	16	16	16	16	16	16	16	16	17	16
Class1	0	0	4	1	8	2	16	16	16	16	16	16	16	16	19	18
Class2	16	16	16	17	16	16	0	0	0	0	0	0	0	0	3	2
Class3	16	16	12	15	8	14	16	16	16	16	16	16	16	16	13	14
Class4	16	16	14	17	16	14	16	16	16	16	16	16	16	16	13	14
Class5	16	16	14	15	12	14	16	16	16	16	16	16	16	16	15	14

Рис. 2. Фрагмент результатів застосування моделі "All-Pairs" з відхиленням зразка класу 5 від правильного діагнозу



	Pattern58	Pattern59	Pattern60	Pattern61	Pattern62	Pattern63	Pattern64	Pattern65	Pattern66	Pattern67	Pattern68	Pattern69	Pattern70	Pattern71	Pattern72	Pattern73
Class0	8	6	8	9	9	8	6	8	6	8	7	6	9	9	9	9
Class1	6	7	6	6	6	6	7	6	8	7	7	8	6	5	6	6
Class2	9	9	9	8	7	9	9	9	9	9	9	9	10	10	10	10
Class3	5	5	5	5	5	7	8	7	7	6	7	7	7	6	7	7
Class4	7	8	7	7	8	5	5	5	5	5	5	5	8	8	8	8
Class5	10	10	10	10	10	10	10	10	10	10	10	10	5	7	5	5

Рис. 3. Фрагмент результатів застосування моделі вичерпного коду. Зразок 37 має однакову Хемінгову відстань, що дорівнює 8, до класів 1 та 3

Результати роботи системи мультикласифікації

Результати застосування моделі персептрон із структурою *34–68–6* без використання розподіленого вихідного коду та моделей розподіленого вихідного коду *"one-against-all"* та *"All-Pairs"* з бінарними класифікаторами типу персептрон *34–68–1* виявилися однаковими і складають 95,9 % точності встановлення діагнозу (70 із 73 зразків тестової вибірки прокласифіковано правильно).

Після тестування персептрону *34–68–6* помилковий діагноз отримали 2 зразки класу 1 та один зразок класу 3. У випадку застосування моделі *"one-against-all"* помилковий діагноз встановлено для одного зразка класу 1 та двох зразків класу 3.

Фрагменти результатів класифікації системи з реалізацією моделі *"All-Pairs"* показано на рис. 1–2 і наведено всі зразки з відхиленням від правильного діагнозу. Зразки у тестовій вибірці впорядковано відповідно до їх класів з розподілом, наведеним вище у табл. 3.

Комірки таблиць, що відповідають класу об'єкта, виділено іншим кольором, значення комірок – найменші значення відстані Хемінга між кодовим словом кожного з об'єктів і рядком матриці кодових слів. Як бачимо на рис. 1, різниця у відстані Хемінга для зразків 35 і 37 до класів 1 і 3 становить лише 1 біт. Решту зразків, не зображених на рис. 1–2, прокласифіковано правильно.

Застосування моделі вичерпного коду на даній тестовій вибірці дозволило правильно прокласифікувати 72 зразки і для зразка 37 було отримано однакову мінімальну відстань Хемінга до класів 1 і 3, що свідчить про значну схожість значень виділених ознак даного зразка для обох класів (діагнозів). Отже, точність встановлення діагнозу з використанням моделі вичерпного коду становить близько 99,5 %. Фрагмент результатів роботи системи з реалізацією моделі вичерпного коду наведено на рис. 3.

Висновки

Розроблено систему мультикласифікації з реалізацією моделі персептрон із структурою *34–68–6* і моделей розподіленого вихідного коду: *"one-against-all"*, *"All-Pairs"* та *вичерпного коду* з використанням персептрону *34–68–1* в якості бінарних класифікаторів.

Розроблену систему мультикласифікації застосовано для вирішення задачі встановлення діагнозу еритозно-плоскоклітинних захворювань у пацієнтів.

Точність діагнозу для тестової вибірки із застосуванням моделей персептрону *34–68–6* і моделей *"one-against-all"* і *"All-Pairs"* склала 95,9 %. Точність встановлення діагнозу з використанням моделі вичерпного коду становить близько 99,5 %, що відповідає здатності даного коду виправляти до трьох помилкових біт.

Отримані результати свідчать про високу ефективність застосування вищеописаних методів і моделей мультикласифікації для вирішення поставленої задачі медичної діагностики за даною вибіркою атрибутів і про перспективність досліджень в області побудови ефективних моделей розподіленого вихідного коду та їх застосування на практиці у прикладних системах штучного інтелекту.

Список літератури

1. Dietterich T. G. Solving Multiclass Learning Problems via Error-Correcting Output Codes / T. G. Dietterich, G. Bakiri // Artificial Intelligence Research. – Vol. 2. – 1995. – P. 263–286.
2. Pujol O. Discriminant ECOC: A Heuristic Method for Application Dependent Design of Error Correcting Output Codes / O. Pujol, P. Radeva, J.Vitriá // IEEE Transaction on pattern analysis and machine intelligence. – Vol. 28. – № 6. – 2006. – P. 1107–1012.
3. UCI Machine Learning Repository. Dermatology Data Set [Електронний ресурс]. – Режим доступу: http://archive.ics.uci.edu/ml/datasets/Dermatology. – Назва з екрана.

O. Porkhun

DETERMINING THE DERMATOLOGY DISEASES USING THE DISTRIBUTED OUTPUT CODE MODELS AND PERCEPTRON

In the article the distributed output code models for solving multiclass learning problems with usage multilayer perceptron are considered and application of the developed multi-classification system implementing the described models for determining the dermatology diseases is described.

Keywords: multi-classification, distributed output code, exhaustive codes, matrix of codewords, multilayer perceptron.

Матеріал надійшов 10.06.2013

УДК 681.3

Марченко О. О.

ПОБУДОВА ЛЕКСИКО-СИНТАКСИЧНОЇ МОДЕЛІ ПРИРОДНОЇ МОВИ ІЗ ЗАСТОСУВАННЯМ СУЧАСНИХ МЕТОДІВ ОБРОБКИ ВЕЛИКИХ ТЕКСТОВИХ КОРПУСІВ

Статтю присвячено розробці алгоритму формування моделі лексико-синтаксичних структурних зв'язків природної мови на основі частотно-синтаксичного аналізу речень великого текстового корпусу. Для запису універсальних структур необмеженої складності та довжини використано модель керуючих просторів синтаксичних структур речень природної мови. Для ефективного та економного представлення даних розріджені масиви трансформовано за допомогою методів невід'ємної факторизації матриць та тензорів.

Ключові слова: обробка текстів природною мовою, керуючі простори синтаксичних структур, невід'ємна факторизація тензорів.

Вступ

Разом із значним зростанням обчислювальної потужності сучасних комп'ютерів і появою нових інтелектуальних алгоритмів обробки великих масивів інформації останнім часом розробка нових методів розв'язку багатьох задач штучного інтелекту вийшла на якісно новий рівень. Серед таких фундаментальних алгоритмів обробки великих масивів інформації виділяється універсальний і потужний підхід – невід'ємна тензорна факторизація.

Невід'ємна тензорна факторизація сьогодні широко затребувана в таких областях, як машинне навчання, обробка зображень, інформаційний пошук, обробка природної мови, та інших напрямках. Такий підхід є одним з найбільш перспективних для виявлення та аналізу взаємозв'язків і від-