7П

УДК 519.95

О.В. Порхун
Київський національний університет імені Тараса Шевченка, Україна
Україна, м. Київ, 01033, вул. Володимирська, 64

Застосування підходів мультикласифікації для встановлення діагнозу дерматологічних захворювань

O.V. Porkhun
Taras Shevchenko National University of Kyiv, Ukraine
Ukraine, 01033 *, c. Kyiv, Volodymyrska, 64*

Application of Multiclass Approaches for Determining Dermatology Diseases

Е.В. Порхун
Киевский национальный университет имени Тараса Шевченка, Украина
Украина, г. Киев, 01033, ул. Владимирская, 64

Применение подходов мультиклассификации для определения диагноза дерматологических заболеваний

У даній статті розглядаються існуючі підходи для вирішення задачі мультикласифікації з використанням моделі нейронної мережі «багатошаровий персептрон» та описано застосування розробленої системи мультикласифікації з реалізацією описаних підходів для вирішення задачі визначення діагнозу захворювання пацієнтів в галузі дерматології.
Ключові слова: мультикласифікація, розподілений вихідний код, матриця кодових слів, багатошаровий персептрон.

In the article the existent approaches for solving Multiclass Learning Problems with usage multilayer perceptron are considered and application of the developed multi-classification system implementing the described approaches for determining the dermatology diseases of patients is described.
Key words: multi-classification, distributed output code, matrix of codewords, multilayer perceptron.

В статье рассматриваются существующие подходы для решения задачи мультиклассификации с применением модели нейронной сети «многослойный персептрон», описано применение разработанной системы мультиклассификации с реализацией описанных подходов для решения задачи определения диагноза заболеваний пациентов в области дерматологии.
Ключевые слова: мультиклассификация, распределенный выходной код, матрица кодовых слов, многослойный персептрон.

Стрімке зростання обсягів інформації та швидкий розвиток інформаційних технологій зумовлюють надзвичайну актуальність питань розробки методів, алгоритмів та систем автоматичної класифікації. Галузі застосування методів класифікації найрізноманітніші, про що свідчить наявність великого числа баз даних машинного навчання, що містять зібрані дані з різних предметних областей – від розпізнавання рукописних символів до медичної діагностики. Зокрема, UCI Machine Learning Repository містить 239 наборів даних, зібраних для вирішення задач класифікації та розпізнавання образів [1].

Існує широкий спектр методів, які добре зарекомендували себе при вирішенні задач класифікації: дерева рішень, нейронні мережі, машини опорних векторів, Adaboost та ін. Дані методи достатньо успішно вирішують задачі бінарної класифікації, але далеко не всі можуть бути легко переналаштовані на випадок вирішення задачі мультикласифікації, тобто коли число класів $n > 2$.

Теоретичні підходи машинного навчання сфокусовані у більшій мірі на навчанні бінарних класифікаторів, у той час як питання розробки методів мультикласифікації не отримують достатнього розвитку та освітлення у наукових працях. У даній статті розглядаються існуючі підходи для вирішення задачі мультикласифікації з використанням моделі нейронної мережі багатошаровий персептрон та описано застосування розробленої системи мультикласифікації з реалізацією описаних підходів для вирішення задачі встановлення діагнозу дерматологічних захворювань у пацієнтів.

Підходи до вирішення задачі мультикласифікації

Задача мультикласифікації полягає у знаходженні невідомої функції $f(x)$, областю значень якої є дискретна множина, що містить k значень (класів) та $k>2$. Дана функція $f(x)$ визначається у процесі навчання на основі навчальної вибірки прикладів виду (x_i,d_i), $i=\overline{1,n}$, де $d_i = f(x_i)$ – відоме бажане значення для прикладу x_i.

Задачу мультикласифікації зводять до вирішення підзадач бінарної класифікації та результатом мультикласифікації є поєднання отриманих розв'язків. Стандартний підхід полягає в тому, щоб навчити k бінарних класифікаторів, бінарних функцій $f_1, f_2, \ldots, f_k$ по одній для кожного класу. Тобто, кожна бінарна функція f_i навчається для розпізнавання об'єктів лише одного класу, для об'єктів інших класів відповідь класифікатора f_i дорівнює 0. Після навчання кожна функція f_i оцінює належність до класу нового об'єкта x. Об'єкт x належить до того класу j, який відповідає найбільшому значенню активації класифікатора f_j. Даний підхід називається *one-per-class.* Необхідно відзначити, що у даному підході число класифікаторів та число класів співпадають.

Альтернативний підхід, що описується в [1], називається *розподілений вихідний код (distributed output code)*. Згідно з даним підходом кожний клас задається бінарним рядком довжини n, «кодовим словом». Кожен біт кодового слова відповідає окремому бінарному класифікатору, який навчається. На відміну від підходу *one-per-class* 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.

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. Згідно з підходом *розподіленого вихідного коду* для кожного класу формується кодове слово та в результаті будується матриця, рядки якої – це кодові слова, а стовпчики – це бінарні класифікатори. Табл. 1 представляє приклад розподіленого вихідного коду довжини 6, що задає представлення для 5 класів.

Таблиця 1 – Розподілений вихідний код довжини 6

Клас	Кодові слова					
	f_1	f_2	f_3	f_4	f_5	f_6
0	0	0	1	1	0	0
1	1	0	0	0	1	1

					Продовження табл.1	
2	0	1	1	0	1	1
3	0	0	0	0	1	0
4	1	1	0	1	0	0

Побудова матриці кодових слів допускає й інші представлення. Зокрема, в [2] описано використання 2-х представлень матриці кодових слів: $M \in \{-1,1\}^{N_c \times n}$ та $M \in \{-1,0,1\}^{N_c \times n}$, 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)* та відповідає матриці в описаному вище підході *one-per-class*, де замість *0* використовується *-1*. У другому представленні матриці *M, «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* та *-1*. Зауважимо, що на відміну від представлення матриці $M \in \{-1,0,1\}^{N_c \times n}$ у підході *one-per-class* об'єкти класів, яким відповідає значення 0, приймають участь у навчанні кожного класифікатора. Для обох представлень матриці *M* при тестуванні нового об'єкта, отримане для нього кодове слово порівнюється з усіма кодовими словами (рядками) матриці *M,* та визначається номер «найближчого» до нього кодового слова, що відповідає ідеї розподіленого вихідного коду.

Визначення мінімальної відстані між отриманим кодовим словом об'єкта, що класифікується, та одним з кодових слів матриці розглядається як процес декодування у проблемі мультикласифікації. В [3] пропонується використання відстані Хемінга для реалізації процесу декодування. Зокрема, мінімальна відстань між отриманим кодовим словом $f(x)=(f_1(x), f_2(x), \ldots, f_n(x))$ та кодовими словами матриці M визначається за формулою:

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

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

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

Набір даних

В якості бази даних для навчання та тестування розробленої системи мультикласифікації був використаний набір даних, зібраних в області дерматології, Dermatology Data Set з UCI Machine Learning Repository. Дана вибірка містить 34 атрибути (http://archive.ics.uci.edu/ml/datasets/Dermatology).

Визначення діагнозу ериматозно-плоскоклітинних захворювань є реальною 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 даній групі є псоріаз, себорейний дерматит, червоний плоский лишай, рожевий лишай, хронічний дерматит та висівкоподібний лишай (lichen pilaris). Для постановки діагнозу як правило необхідна біопсія, однак дані захворювання мають достатньо багато спільних гістопаталогічних 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, які згруповані як клінічні атрибути, а також виміри 22 гістопаталогічних ознак, отримані шляхом аналізу зразків шкіри під мікроскопом. До групи клінічних атрибутів входить ознака родинної історії, яка приймає значення 1, якщо будь-яке із даних захворювань спостерігалося у родині, та 0 – у протилежному випадку. Також група клінічних атрибутів включає вік пацієнта. Всі інші ознаки мають область значень від 0 до 3, тут 0 означає, що ознака не присутня, 3 – вказує на найбільш можливе значення, 1 та 2 – вказують відносні проміжні значення. Всі атрибути з вибірки Dermatology Data Set наведено у табл. 2.

Таблиця 2 – Атрибути для визначення діагнозу ериматозно-плоскоклітинних захворювань

Клінічні атрибути (приймають значення 0,1,2,3)	
1: erythema 2: scaling 3: definite borders 4: itching 5: koebner phenomenon 6: polygonal papules	7: follicular papules 8: oral mucosal involvement 9: knee and elbow involvement 10: scalp involvement 11: family history (0 or 1) 34: Age (linear)
Гістопаталогічні атрибути (приймають значення 0,1,2,3)	
12: melanin incontinence 13: eosinophils in the infiltrate 14: PNL infiltrate 15: fibrosis of the papillary dermis 16: exocytosis 17: acanthosis 18: hyperkeratosis 19: parakeratosis 20: clubbing of the rete ridges 21: elongation of the rete ridges 22: thinning of the suprapapillary epidermis	23: spongiform pustule 24: munro microabcess 25: focal hypergranulosis 26:disappearance of the granular layer 27:vacuolisation and damage of basal layer 28: spongiosis 29: saw-tooth appearance of retes 30: follicular horn plug 31: perifollicular parakeratosis 32: inflammatory monoluclear inflitrate 33: band-like infiltrate

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 підходи *«one-against-all»* та *«All-Pairs»* з відповідними представленнями матриць $M \in \{-1,1\}^{N_c \times n}$ та $M \in \{-1,0,1\}^{N_c \times n}$ та для декодування використовує відстань Хемінга (1) та Евклідову відстань. Також у даній системі окремо реалізовано вирішення поставленої задачі з використанням багатошарового персептрона, де кількість класів задається числом нейронів у вихідному шарі.

Як класифікатор у реалізаціях підходів *«one-against-all»* та *«All-Pairs»* використовується багатошаровий персептрон з єдиним нейроном у вихідному шарі.

Для вирішення задачі визначення діагнозу ериматозно-плоскоклітинних захворювань система мультикласифікації була навчена та протестована відповідно до кожної з 3реалізацій.

Матриця кодових слів для представлення *«one-against-all»* має вигляд:

	f_1	f_2	f_3	f_4	f_5	f_6
C_0	1	-1	-1	-1	-1	-1
C_1	-1	1	-1	-1	-1	-1
C_2	-1	-1	1	-1	-1	-1
C_3	-1	-1	-1	1	-1	-1
C_4	-1	-1	-1	-1	1	-1
C_5	-1	-1	-1	-1	-1	1

Тобто кожен класифікатор f_i використовує всі приклади з навчальної вибірки, але розпізнає лише один клас прикладів. Тут число рядків відповідає числу класів – діагнозів, число стовпчиків – кількості бінарних класифікаторів (персептронів).

Матриця кодових слів для реалізації представлення *«All-Pairs»*має вигляд:

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
C_0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
C_1	-1	0	0	0	0	1	1	1	1	0	0	0	0	0	0
C_2	0	-1	0	0	0	-1	0	0	0	1	1	1	0	0	0
C_3	0	0	-1	0	0	0	-1	0	0	-1	0	0	1	1	0
C_4	0	0	0	-1	0	0	0	-1	0	0	-1	0	-1	0	1
C_5	0	0	0	0	-1	0	0	0	-1	0	0	-1	0	-1	-1

У цьому випадку кожен бінарний класифікатор навчається на прикладах двох класів, значення 0 позначає, що приклади даного класу ігноруються при навчанні.

Підготовка даних

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

Вибірка даних була поділена на навчальну та тестову у відношенні 80:20%: навчальна вибірка складається з 284 прикладів та тестова – з 73 прикладів. Кількість прикладів у класах, відповідних діагнозам захворювань, наведена у табл. 3 для навчальної вибірки, та у табл. 4 для тестової вибірки.

Таблиця 3 – Розподіл прикладів по класам 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 4 – Розподіл прикладів по класам 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 був застосований метод зворотнього 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 реалізована можливість випадкової подачі прикладів у мережу, що дозволяє зробити пошук у просторі ваг стохастичним і в свою чергу зводить до мінімуму зупинку алгоритма у точці деякого локального мінімума. 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 було 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 була сформована структура мережі *34 – 68 – 6* для персептрону з вхідним шаром з 34 нейронів (34 ознаки), з прихованим шаром з 68 нейронів та з 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 $a=1$.

Результати роботи системи мультикласифікації

У процесі тестування системи відповідно до трьох реалізованих моделей для тестової вибірки з 73 прикладів (випадків захворювання), 70 було прокласифіковано вірно, тобто точність встановлення діагнозу для тестової вибірки склала 95,8904%.

Після навчання багатошарового персептрону із структурою *34 – 68 – 6* загальна помилка класифікації становить 0,0256.

При реалізації підходу *«one-against-all»* значення загальної помилки для кожного бінарного класифікатора наведено у табл. 5.

Таблиця 5 – Результати роботи бінарних класифікаторів з використанням матриці кодових слів *«one-against-all»*

	f_1	f_2	f_3	f_4	f_5	f_6
Total error	9,9973E-5	0,03828	9,9573E-5	0,05278	9,9514E-5	9,9526E-5

У випадку реалізації підходу *«All-Pairs»* значення загальної помилки для кожного з 15 бінарних класифікаторів наведено у табл. 6.

Таблиця 6 – Загальна помилка бінарних класифікаторів для представлення «All-Pairs»

Classificator	Total error	Classificator	Total error	Classificator	Total error
f_1	9,9962E-5	f_7	0,03359	f_{13}	9,9754E-5
f_2	9,9871E-5	f_8	9,9992E-5	f_{14}	9,9891E-5
f_3	9,9673E-5	f_9	0,04563	f_{15}	9,9944E-5
f_4	9,9973E-5	f_{10}	9,9985E-5		
f_5	9,9485E-5	f_{11}	9,9971E-5		
f_6	9,9981E-5	f_{12}	9,9591E-5		

Для представлень *«one-against-all»* та *«All-Pairs»* на етапі декодування помилка класифікації склала 3 зразки (для *«one-against-all»*: 1 зразок класу 1 та 2 зразки класу 3; для *«All-Pairs»*: 2 зразки класу 1 та 1 зразок класу 5).

При реалізації розподіленого вихідного коду для представлення *«All-Pairs»* в процесі навчання для 13 бінарних класифікаторів було досягнуто глобальний мінімум помилки за число епох в середньому < 1200; для моделі з представленням *«one-against-all»* глобальний мінімум помилки було досягнуто для 4 бінарних класифікаторів за число епох в середньому < 1400. Незважаючи на більшу кількість класифікаторів у представленні *«All-Pairs»,* система навчилася значно швидше, ніж у випадку застосування моделей представлення *«one-against-all»* та багатошарового персептрону 34 – 68 – 6. Це пов'язано з тим, що у даному представленні кожен класифікатор використовує об'єкти лише двох класів і лише для 2 бінарних класифікаторів умовою зупинки навчання було досягнення заданого числа епох. Таким чином, перевагою представлення *«All-Pairs»*для вирішення поставленої задачі є швидкість навчання системи, яка значно вища порівняно з представленням *«one-against-all»* та з моделлю персептрона *34 – 68 – 6.*

Процес навчання та результати роботи системи з реалізацією підходу *«All-Pairs»* показано на рис. 1-2.

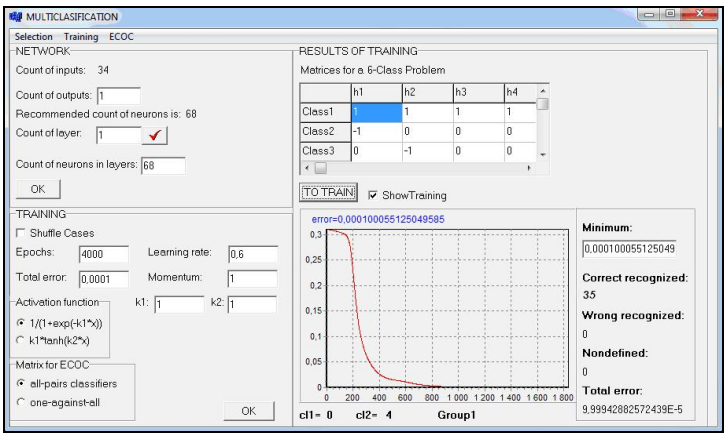
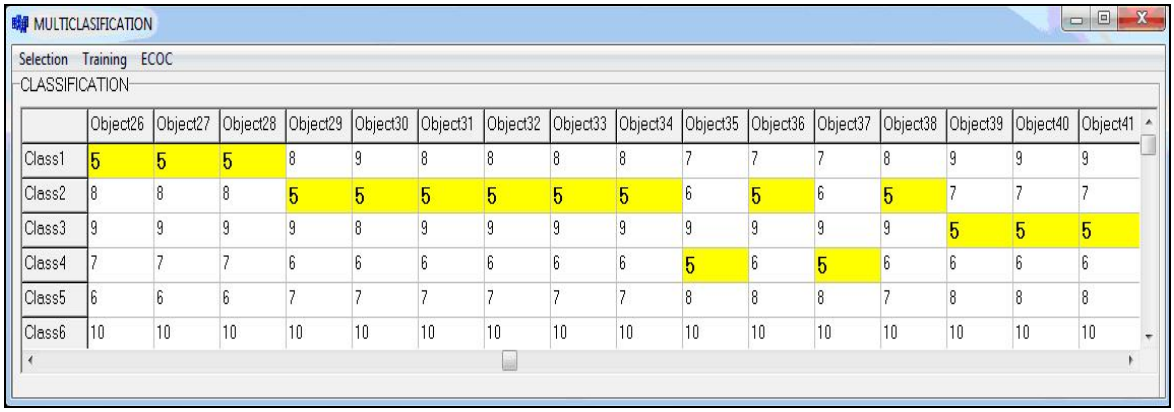


Рисунок 1 – Навчання системи на основі підходу «*All-Pairs*».

Розпізнавання зразків класу 0 та класу 4



	Object26	Object27	Object28	Object29	Object30	Object31	Object32	Object33	Object34	Object35	Object36	Object37	Object38	Object39	Object40	Object41
Class1	5	5	5	8	9	8	8	8	8	7	7	7	8	9	9	9
Class2	8	8	8	5	5	5	5	5	5	6	5	6	5	7	7	7
Class3	9	9	9	9	8	9	9	9	9	9	9	9	9	5	5	5
Class4	7	7	7	6	6	6	6	6	6	5	6	5	6	6	6	6
Class5	6	6	6	7	7	7	7	7	7	8	8	8	7	8	8	8
Class6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Рисунок 2 – Результат роботи системи на основі підходу *«All-Pairs»*

На рис. 2 комірки таблиці, що відповідають класу об’єкта, виділено жовтим кольором, значення комірок – найменші значення відстані Хемінга (1) між кодовими словом кожного з об’єктів та рядком матриці кодових слів.

Висновки

1. На основі існуючих підходів вирішення задачі мультикласифікації розроблена система мультикласифікації з реалізацією підходу розподіленого вихідного коду для двох представлень матриці кодових слів: $M \in \{-1,1\}^{N_c \times n}$ та $M \in \{-1,0,1\}^{N_c \times n}$ та із застосуванням багатошарового персептрону.

2. Розроблена система мультикласифікації була застосована для вирішення задачі встановлення діагнозу ериматозно-плоскоклітинних захворювань у пацієнтів та отримана точність діагнозу для тестової вибірки склала 95,8904%.

3. Результати застосування описаних моделей мультикласифікації збігаються у точності встановлення діагнозу, однак модель із реалізацією представлення *«All-Pairs»* має перевагу у швидкості навчання.

Література

1. UCI Machine Learning Repository [Електронний ресурс]. – Режим доступу: http://archive.ics.uci.edu/ml/datasets.html
2. Dietterich T.G. Solving Multiclass Learning Problems via Error-Correcting Output Codes / T.G. Dietterich, G. Bakiri // Artificial Intelligence Research. – 1995. – vol. 2. – P. 263-286.
3. Oriol Pujol. Discriminant ECOC: A Heuristic Method for Application Dependent Design of Error Correcting Output Codes / Oriol Pujol, Petia Radeva, Jordi Vitriá // IEEE Transaction on pattern analysis and machine intelligence. – 2006. – vol. 28, №. 6. – P. 1107-1012.
4. Хайкин С. Нейронные сети. Полный курс. Второе издание / Хайкин С. – М. : «Вильямс», 2008. – С. 115-126.

Literatura

1. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html
2. Dietterich T.G. and G. Bakiri. Solving Multiclass Learning Problems via Error-Correcting Output Codes. – Artificial Intelligence Research, vol. 2, 1995. – P. 263-286.

3. Oriol Pujol, Petia Radeva, and Jordi Vitriá. Discriminant ECOC: A Heuristic Method for Application Dependent Design of Error Correcting Output Codes. – IEEE Transaction on pattern analysis and machine intelligence, vol. 28, № 6, 2006. – P.1107-1012.
4. Simon Haykin. Neural Networks. A Comprehensive Foundation. Second Edition. – M. : «Viliams», 2008. – 1104 p.

RESUME

O.V. Porkhun

Application of Multiclass Approaches for Determining Dermatology Diseases

In given article the existent approaches for solving Multiclass Learning Problems with usage multilayer perceptron are considered. Based on these approaches the multi-classification system implementing the distributed output code for two representation of the codewords matrix: $M \in \{-1,1\}^{N_c \times n}$ and $M \in \{-1,0,1\}^{N_c \times n}$ using multilayer perceptron was developed.

This system has been applied to solve the problem of determining erythemato-squamous diseases of patients and the diseases accuracy obtained for the test data set was 95.8904%.

The diseases accuracy with the usage described multi-classification models coincide, but the model implementing the representation *"All-Pairs"* has the advantage of training rate.

Стаття надійшла до редакції 23.04.2013.