УДК 004.93

Субботін С. О.

АВТОМАТИЧНЕ ВИДІЛЕННЯ ВИБІРОК ДЛЯ ПОБУДОВИ НЕЙРОМОДЕЛЕЙ

Запропоновано метод виділення вибірок, який для вихідної вибірки визначає індивідуальну значущість екземплярів, після чого послідовно нарощує підвибірку, відбираючи до неї найбільш індивідуально високоінформативні екземпляри в кожному класі і виключаючи екземпляри, які є надлишковими або такими, що погіршують класифікацію. Це дає змогу автоматизувати аналіз вибірки, скоротити розмірність навчальних даних, а також скоротити час і забезпечити прийнятну точність навчання нейромоделей. Проведено експерименти з дослідження запропонованого методу, результати яких дозволяють рекомендувати його для використання на практиці в задачах діагностування та розпізнавання образів.

Ключові слова: вибірка, відбір екземплярів, редукція даних, нейронна мережа, скорочення розмірності даних.

Вступ

Одним з найбільш популярних і поширених на практиці засобів побудови діагностичних і розпізнавальних моделей є штучні нейронні і нейро-нечіткі мережі [4].

Для побудови моделей на основі нейронних мереж необхідно мати навчальну вибірку даних. Формування навчальної вибірки мінімального обсягу є дуже важливим завданням, оскільки використання наявних на практиці великих вибірок даних для навчання нейромереж призводить до істотного збільшення часу побудови і вимог до ресурсів пам'яті, а також може призводити до отримання надлишкових моделей.

Відомі методи формування вибірок [3; 5–11] характеризуються низькою швидкістю роботи, а також невизначеністю критеріїв оцінки якості одержуваної підвибірки.

Метою цієї роботи було підвищення швидкості процесу формування та якості виділюваних навчальних вибірок для побудови нейромоделей.

Постановка завдання

Нехай задана вихідна вибірка у вигляді набору прецедентів $<x, y>$, де $x = \{x^s\}$, x^s – s-й екземпляр вибірки, $x = \{x_j\}$, $x^s = \{x^s_j\}$, $x_j = \{x^s_j\}$, x^s_j – значення j-ї діагностичної ознаки x_j, що характеризує екземпляр x^s, $y = \{y^s\}$, y^s – значення вихідної ознаки, зіставлене екземпляру x^s, $s = 1, 2, ..., S$, $j = 1, 2, ..., N$, S – кількість екземплярів у вихідній вибірці, N – число діагностичних ознак, що характеризують вибірку.

Для заданої вибірки прецедентів $<x, y>$ задача синтезу нейромоделі може бути подана як задача знаходження $<F(), w>$: $y^* = F(w, x^s)$, $f(F(), w, <x, y>) \rightarrow$ opt, де $F()$ – структура нейромоделі, яка на практиці зазвичай задається користувачем; w – набір керованих параметрів, що налаштовується на основі навчальної вибірки; $f()$ – користувальницький критерій, що характеризує якість аргументу щодо розв'язуваної задачі; opt – оптимальне (бажане чи прийнятне) значення функціонала $f()$ для розв'язуваної задачі.

Своєю чергою, задача формування підвибірки заданої вибірки $<x, y>$ полягає у знаходженні такого набору $X' = <x', y'>$: $x' \subset \{x^s\}$, $y' = \{y^s | x^s \in \}$, $S' < S$, $N' = N$, при якому $f(<x', y'>, <x, y>) \rightarrow$ opt, де N' – кількість ознак у підвибірці; S' – кількість екземплярів у підвибірці.

Аналіз літератури

Методи формування вибірок для побудови моделей прийняття рішень за прецедентами в [7; 8] поділяють на методи вибору прототипу і методи побудови прототипу, а також комбіновані методи. Тут під прототипом мається на увазі виділювана підвибірка відносно початкової вибірки.

Методи вибору прототипу [3; 5; 6; 10; 11] не модифікують, а тільки відбирають найбільш важливі екземпляри з вихідної вибірки. Ці методи залежно від стратегії формування рішень поділяють на методи з додаванням екземплярів [6; 3] (послідовно додають екземпляри з вихідної вибірки у формовану підвибірку), методи з видаленням екземплярів [3; 5; 6; 10] (послідовно видаляють екземпляри з вихідної вибірки, отримуючи

в підсумку підвибірку), методи фільтрації шуму [5; 10] (видаляють екземпляри, мітки класів яких не збігаються з мітками більшості сусідніх екземплярів), методи конденсації [3; 5; 6; 11] (додають екземпляри з вихідної вибірки у формовану підвибірку, якщо вони несуть нову інформацію, але не додають, якщо вони мають ті ж мітки класів, що й сусідні з ними екземпляри) і методи на основі стохастичного пошуку [11] (випадковим чином формують підвибірку з вихідної вибірки, можливо, перебираючи деяку множину варіантів і відбираючи найкращий з них). Загальними недоліками цих методів є висока ітеративність і тривалість пошуку, а також невизначеність вибору критеріїв якості одержуваної підвибірки.

Методи побудови прототипу на основі вихідної вибірки [9; 11] створюють штучні екземпляри, що дозволяють описувати вихідну вибірку. Спільними недоліками цих методів є висока ітеративність і значний час роботи, а також невизначеність у заданні параметрів методів.

Комбіновані методи [8] поєднують формування та відбір прототипів. Комбіновані методи мають недоліки як методів виділення, так і методів побудови прототипу.

Оскільки методи побудови прототипу і пов'язані з ними комбіновані методи є більш повільними, ніж методи вибору прототипу, то останні доцільно обрати як базис для вирішення завдання формування вибірок.

Для усунення недоліків, властивих цим методам, доцільно у формовану вибірку включати найбільш індивідуально інформативні екземпляри вихідної вибірки, а також зупиняти пошук при досягненні необхідної якості формованої вибірки. Це, відповідно, вимагає визначення показників, що дозволяють оцінювати індивідуальну інформативність екземплярів, а також критерію якості вибірки.

Оцінка індивідуальної значущості екземплярів вибірки

Заданий набір ознак $\{x_j\}$ утворює простір, у якому екземпляри вихідної вибірки являють собою точки. Розіб'ємо простір ознак $\{x_j\}$ на прямокутні області, обмеживши діапазон значень кожної ознаки x_j мінімальним і максимальним його значеннями. Тоді проекції розбиття на вісь ознаки дозволять виділити інтервали значень ознаки для кожного з прямокутних блоків. Інтервали можуть формуватися як проекції кластерів, або як регулярні ґрати, або на основі границь класів в одномірних проекціях вибірки на осі ознак [1].

Тоді кожний такий інтервал можна вважати термом і оцінити його значущість для прийняття рішень про віднесення екземпляра до кластера за допомогою ваги k-го терма j-ї ознаки для s-го екземпляра x^s щодо опису центра відповідного інтервалу за формулою: $w_{C\,jk}^{\,s} = \exp(-(0{,}5(r_{jk} - l_{jk}) - x^s_j)^2)$, а також ваги k-го терма j-ї ознаки для s-го екземпляра x^s щодо опису міжкластерних границь за формулою: $w_{B\,jk}^{\,s} = \exp(-(\min((r_{jk} - x^s_j), (x^s_j - l_{jk}))^2)$.

Відповідно, загальну значущість k-го терма j-ї ознаки для s-го екземпляра x^s щодо опису міжкластерних границь можна оцінити за допомогою ваги, що визначається за формулою: $w^s_{jk} = \max\{w_{C\,jk}^{\,s}, w_{B\,jk}^{\,s}\}$.

Визначивши для кожного s-го екземпляра значущості термів, можна також визначити ваги термів для всієї вибірки:

$$w_{jk} = \frac{S_{jk}}{SK_{jk}},$$

де K_{jk} – кількість класів, екземпляри яких потрапили у k-й інтервал значень j-ї ознаки; S_{jk} – кількість екземплярів, що потрапили у k-й терм j-ї ознаки.

Знаючи оцінки значущості термів, визначимо оцінки інформативності ознак за формулою:

$$w_j = \max_k\{w_{jk}\}.$$

На основі оцінок значущості термів і ознак можна визначити оцінки інформативності для кожного s-го екземпляра вибірки за формулою:

$$I(x^s) = \frac{\sum_{j=1}^{N}\left(w_j \sum_{k=1}^{k_j} w_{jk} w^s{}_{jk}\right)}{\sum_{j=1}^{N}\left(w_j \sum_{k=1}^{k_j} w_{jk} \max_p\{w^p{}_{jk}\}\right)}.$$

Чим більшим буде значення запропонованого показника, тим більш важливим буде екземпляр відносно вибірки, що його містить, і навпаки.

Метод відбору екземплярів для побудови нейромоделей

Запропонований показник оцінювання індивідуальної значущості екземплярів можна використовувати для автоматичного формування підвибірок із заданої вихідної вибірки на основі пропонованого нижче методу.

Етап ініціалізації: задати вихідну вибірку $<x, y>$ і занести її у множину нерозглянутих екземплярів $X = <x, y>$. Задати множину розглянутих екземплярів (сформовану підвибірку):

$X' = <x', y'> = \varnothing$. Задати прийнятне значення помилки ε.

Етап аналізу вибірки: сформувати розбиття простору ознак, наприклад, на основі методу [10], визначити індивідуальні оцінки інформативності екземплярів вихідної вибірки $\{I(x^s)\}$.

Етап додавання екземплярів. Із множини нерозглянутих екземплярів X відібрати по одному екземпляру кожного класу з найбільшою індивідуальною оцінкою значущості серед екземплярів свого класу, додати їх у X' і видалити з X:

$$X' = X' \cup \{< x^s, y^s > | < x^s, y^s > \in X, I(x^s) \geq I(x^p),$$
$$y^s = y^p, < x^p, y^p > \in X, s \neq p, s, p = 1,2,...,S\},$$
$$X = X \setminus \{< x^s, y^s > | < x^s, y^s > \in X, I(x^s) \geq I(x^p),$$
$$y^s = y^p, < x^p, y^p > \in X, s \neq p, s, p = 1,2,...,S\}.$$

Скорегувати відповідним чином S і S', де S' – обсяг вибірки X'.

Етап оцінювання якості підвибірки: оцінити якість поточної підвибірки X' щодо всієї вихідної вибірки $<x, y>$ за допомогою критерію помилки:

$$E = \sum_{s=1}^{S} \{1 \mid y^s \neq y'^z, z = \arg \min_{p=1,2,...,S'} \{R(x^s, x'^p)\}\},$$

$$R(x^s, x'^p) = R(x'^p, x^s) = \sum_{j=1}^{N} (x_j^s - x'^p_j)^2.$$

Етап перевірки закінчення пошуку. Якщо нове значення помилки E стало припустимим ($E \leq \varepsilon$) або відсутні нерозглянуті екземпляри ($X = \varnothing$), тоді закінчити пошук і повернути як результат сформовану вибірку X'. Якщо ж $E > \varepsilon$ і є нерозглянуті екземпляри ($X \neq \varnothing$), тоді якщо нове значення E стало більшим від попереднього значення, то перейти до етапу видалення екземплярів, у протилежному випадку – перейти до етапу додавання екземплярів.

Етап видалення екземплярів. Для кожного екземпляра сформованої підвибірки X' визначити його внесок у загальну помилку:

$$E(x'^p) = \sum_{s=1}^{S} \{1 \mid y^p \neq y'^s, R(x^s, x'^p) \leq R(x^s, x'^z), z = 1,2,...,S'\}.$$

У множині X' знайти екземпляр з найбільшим внеском у помилку і видалити його з X':

$$X' = X' \setminus \{< x'^s, y'^s > | < x'^s, y'^s > \in X', E(x'^s) \geq$$
$$\geq E(x'^p), < x'^p, y'^p > \in X', s \neq p, s, p = 1,2,...,S'\},$$
$$S' = S' - 1.$$

Оцінити помилку E для всієї вибірки щодо отриманої множини X'. Якщо вона зменшилася, то повторювати цей етап доти, доки в X' є хоча б один екземпляр з ненульовим індивідуальним внеском у помилку ($\exists x'^p : E(x'^p) > 0$), після чого перейти до етапу перевірки закінчення пошуку; у протилежному випадку – завершити пошук і повернути попередню підвибірку X'.

Запропонований метод прагне від самого початку розглянути найбільш перспективні екземпляри кожного класу для включення у формовану вибірку. У випадку, якщо одержувана підвибірка забезпечує прийнятний рівень помилки, метод завершує роботу. Якщо ж помилка не є прийнятною, але зменшується, то метод послідовно нарощує підвибірку, прагнучи забезпечити потрапляння до неї екземплярів-представників усіх класів порівно. Якщо ж помилка починає зростати, то метод виявляє екземпляри сформованої підвибірки, що погіршують класифікацію, і виключає їх із неї.

У результаті виконання цього методу з вихідної вибірки $<x, y>$ буде сформована навчальна підвибірка X', що містить найбільш значущі екземпляри класів, що дозволяють забезпечити побудову найбільш точної моделі. Залишок екземплярів вихідної вибірки, що не увійшов у сформовану підвибірку, можна розглядати як тестову підвибірку.

Експерименти і результати

Для дослідження запропонованого методу було розроблено комп'ютерну програму, яка використовувалася при проведенні експериментів з розв'язання задач діагностування і розпізнавання образів на основі нейронних мереж. Характеристики вибірок даних наведено в табл. 1.

Таблиця 1. **Характеристики вихідних вибірок даних**

Задача	*N*	*S*	*K*
Діагностування лопаток газотурбінних авіадвигунів [10]	512	32	2
Розпізнавання автотранспортних засобів за характеристиками зображень [11]	26	800	4

Запропонований метод порівнювався з методом еволюційного пошуку, у якому якість рішень визначалася шляхом побудови нейромоделі з подальшою оцінкою її помилки, а процес пошуку продовжувався до одержання прийнятного рішення.

Далі для кожної сформованої вибірки будувалася модель на основі нейронної мережі прямого поширення, що навчалася за допомогою методу Левенберга – Марквардта [4]. Усі нейрони мережі використовували вагову (постсинаптичну)

функцію – зважена сума, і функцію активації – логістичний сигмоїд.

Після навчання кожна модель тестувалася на всій вихідній вибірці, за якою визначалася помилка E як кількість екземплярів відповідної вибірки, для яких розрахункове і фактичне значення цільової ознаки не збіглися. Результати проведених експериментів наведено в табл. 2.

Таблиця 2. **Результати експериментів**

Задача	Запропонований метод				Метод еволюційного пошуку			
	S'	S'/S	E	t, хв	S'	S'/S	E	t, хв
Діагностування лопаток газотурбінних авіадвигунів	6	0,19	0	13,1	8	0,25	0	32,4
Розпізнавання автотранспортних засобів за характеристиками зображень	119	0,15	1	17,9	143	0,18	7	44,1

Як видно з табл. 2, запропонований метод дає змогу одержувати модель із прийнятною точністю за менший час у порівнянні з методом еволюційного пошуку, що пояснюється, з одного боку, незалежністю запропонованого методу від моделі, а з другого боку, врахуванням у процесі пошуку інформації про індивідуальну значущість екземплярів. При цьому розмір вибірки, сформованої запропонованим методом, виявився меншим за рахунок виключення надлишкових прикладів.

Проведені експерименти підтвердили роботоздатність запропонованого методу і програмного забезпечення, що його реалізує, і дозволяють рекомендувати цей метод для використання на практиці при розв'язанні задач діагностування і розпізнавання образів.

Висновки

Вирішено актуальне завдання розроблення математичного забезпечення для автоматизації формування вибірок при побудові діагностичних і розпізнавальних моделей за прецедентами. Уперше запропоновано метод формування навчальних вибірок, який для вихідної вибірки визначає індивідуальну значущість екземплярів, після чого послідовно нарощує підвибірку, відбираючи до неї найбільш індивідуально високоінформативні екземпляри в кожному класі і виключаючи екземпляри, які є надлишковими або такими, що погіршують класифікацію. Це дає змогу автоматизувати аналіз вибірки, скоротити розмірність навчальних даних, а також скоротити час і забезпечити прийнятну точність навчання нейромоделей.

Практична цінність отриманих результатів полягає в тому, що розроблено програмне забезпечення, яке реалізує запропонований метод, а також проведено експерименти з його дослідження, результати яких дозволяють рекомендувати цей метод для використання на практиці в задачах діагностування та розпізнавання образів.

Роботу виконано в межах держбюджетної науково-дослідної теми Запорізького національного технічного університету «Інтелектуальні інформаційні технології автоматизації проектування, моделювання, управління та діагностування виробничих процесів і систем» (номер держ. реєстрації 0112U005350) за підтримки міжнародного проекту «Centers of Excellence for young RESearchers» Європейської Комісії (№ 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

Список літератури

1. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : монография / [С. А. Субботин, Ан. А. Олейник, Е. А. Гофман и др.] ; под ред. С. А. Субботина. – Харьков : Компания СМИТ, 2012. – 318 с.
2. Субботин С. А. Автоматическая система обнаружения и распознавания автотранспортных средств на изображении / С. А. Субботин, К. Ю. Бойченко // Программные продукты и системы. – 2010. – № 1. – С. 114–116.
3. Aha D. W. Instance-based learning algorithms / D. W. Aha, D. Kibler, M. K. Albert // Machine Learning. – 1991. – № 6. – P. 37–66.
4. Engelbrecht A. Computational intelligence: an introduction / A. Engelbrecht. – Sidney : John Wiley & Sons, 2007. – 597 p.
5. Gates G. The reduced nearest neighbor rule / G. Gates // IEEE Transactions on Information Theory. – 1972. – Vol. 18, № 3. – P. 431–433.
6. Hart P. E. The condensed nearest neighbor rule / P. E. Hart // IEEE Transactions on Information Theory. – 1968. – Vol. 14. – P. 515–516.
7. Jankowski N. Comparison of instance selection algorithms I. Algorithms survey / N. Jankowski, M. Grochowski // Artificial Intelligence and Soft Computing : 7th International Conference ICAISC-2004, Zakopane, 7–11 June, 2004 : proceedings. – Berlin : Springer, 2004. – P. 598–603. – (Lecture Notes in Computer Science, Vol. 3070).
8. Reinartz T. A unifying view on instance selection / T. Reinartz // Data Mining and Knowledge Discovery. – 2002. – № 6. – P. 191–210.
9. Schikuta E. Grid-clustering: a fast hierarchical clustering method for very large data sets / E. Schikuta // Pattern Recognition : 13th International Conference, Vienna, 25–29 August 1996 : proceedings. – Los Alamitos : IEEE, 1996. – Vol. 2. – P. 101–105.

10. Wilson D. L. Asymptotic properties of nearest neighbor rules using edited data / D. L. Wilson // IEEE Transactions on Systems, Man, Cybernetics. – 1972. – Vol. 2, № 3. – P. 408–421.

11. Wilson D. R. Reduction techniques for instancebased learning algorithms / D. R. Wilson, T. R. Martinez // Machine Learning. – 2000. – Vol. 38, № 3. – P. 257–286.

S. Subbotin

THE AUTOMATIC SAMPLE EXTRACTION FOR NEURAL NETWORK MODEL BUILDING

The sample selection method is proposed, which for the original sample determines the individual instance significance, followed by successive increase of a subsample, selecting the most highly informative individually instances in each class, and excluding instances that are redundant or worse a classification. This allows to automate the sample analysis, to reduce the training data dimensionality, to reduce the time and to provide the acceptable accuracy of neural network training. The experiments to investigate the proposed method are conducted. Their results allow recommend method to use in practice for the diagnosis and pattern recognition.

Keywords: sample, instance selection, data reduction, neural network, data dimensionality reduction.

Матеріал надійшов 01.09.2014

УДК 519.7

Галкін О. А.

НЕПАРАМЕТРИЧНІ ОЦІНКИ УСЕРЕДНЕНИХ ЯДЕРНИХ ВІДОБРАЖЕНЬ УМОВНИХ РОЗПОДІЛІВ ДЛЯ ЗАДАЧ РОЗПІЗНАВАННЯ ОБРАЗІВ

Статтю присвячено непараметричним оцінкам усереднених ядерних відображень умовних розподілів, що є неявними відображеннями розподілу в потенційно нескінченновимірний простір характеристик, а також комплексному ядерному підходу для розв'язання широкого класу задач розпізнавання образів. Ключова ідея полягає у відображенні умовного розподілу в нескінченновимірний простір характеристик з використанням функції ядра. Запропонований підхід може бути використаний для побудови більш простих та ефективних статистик для оцінки такого неперервного мультимодального розподілу, як функція глибини.

Ключові слова: ядерне відображення, функція ядра, оцінка щільності.

Вступ

Розглядаючи неперервні генеральні сукупності, припустимо, що задано випадкову величину Z, генеральну сукупність H, розподіл $D(Z)$, а також щільність $d(Z)$. Будемо вважати, що випадкова величина X також належить генеральній сукупності H, а також припустимо, що запропонований підхід має місце лише у випадку, коли Z та X належать різним генеральним сукупностям.