

УДК 004.93

МЕТОДИ ВИЗНАЧЕННЯ ФРАКТАЛЬНОЇ РОЗМІРНОСТІ ВИБІРОК ДАНИХ І МОДЕЛЕЙ

С. О. Субботін, д. т. н., професор

Запорізький національний технічний університет

subbotin@zntu.edu.ua

Запропоновано показники, що характеризують якість підвибірок відносно вихідної вибірки на основі принципів фрактального аналізу. Розроблені показники досліджені при вирішенні завдань формування навчальних вибірок даних.

Subbotin S. A. Methods of data sample and model fractal dimension evaluation. The indicators allowing to characterize the quality of subsamples with respect to the original sample on the basis of the principles of fractal analysis is proposed. The developed indicators studied in the training data sample forming tasks.

Ключові слова: ФРАКТАЛЬНА РОЗМІРНІСТЬ, ВИБІРКА, МОДЕЛЬ.

Keywords: FRACTAL DIMENSION, SAMPLE, MODEL.

Нехай ми маємо вихідну вибірку $\langle x, y \rangle$ – набір S прецедентів про залежність $y(x)$, $x = \{x^s\}$, $y = \{y^s\}$, $s = 1, 2, \dots, S$, що характеризуються набором N вхідних ознак $\{x_j\}$, $j = 1, 2, \dots, N$, де j – номер ознаки, і вихідною ознакою y . Кожен s -й прецедент подамо як $\langle x^s, y^s \rangle$, $x^s = \{x_j^s\}$, де x_j^s – значення j -ї вхідної, а y^s – значення вихідної ознаки для s -го прецедента (екземпляра) вибірки, $y^s \in \{1, 2, \dots, K\}$, де K – кількість класів, $K > 1$. Тоді задача синтезу моделі залежності $y(x)$ буде полягати у визначенні такої структури і параметрів моделі net , при яких буде задовільнено критерій якості моделі $E(net, \langle x, y \rangle) \rightarrow \min$.

У випадку, коли вихідна вибірка має велику розмірність, перед побудовою нейромоделі необхідно вирішити задачу виділення навчальної вибірки меншого обсягу, що вимагає критерію, який характеризує якість формованої підвибірки.

Перспективним напрямком для визначення властивостей даних є фрактальна розмірність – коефіцієнт, що описує фрактальні або структурні множини на основі кількісної оцінки їхньої складності як коефіцієнт зміни в деталях зі зміною масштабу [1]. Відомі методи оцінки фрактальної розмірності даних [2] і моделей [3] засновані на різних принципах. Що не дозволяє інтегрувати їхні результати.

Метою цієї роботи було створення показників якості вибірок на основі принципів фрактального аналізу.

Для аналізу фрактальної розмірності вибірки пропонується використовувати такий метод.

Етап ініціалізації. Задати навчальну вибірку $\langle x, y \rangle$, пронормувати значення ознак, відобразивши їх на інтервал $[0, 1]$. Задати L – кількість інтервалів, на які будуть розбиватися діапазони значень ознак.

Етап кластеризації. Розбити діапазон значень кожної ознаки на L інтервалів довжиною $l = L^{-1}$. Сформувати кластери як прямокутні блоки на перетинанні інтервалів різних ознак.

Етап аналізу даних. Визначити кількість екземплярів, що потрапили до кожного прямокутного гіперблоку, утвореного інтервалами ознак. Визначити кількість гіперблоків зі стороною розміром l , що покривають вибірку в просторі N ознак $n(l)$.

Етап визначення фрактальної розмірності. Визначити при заданому l фрактальну розмірність вибірки: $D = n(l) / \log(l^{-1})$.

Єдиним керованим параметром методу є задана кількість інтервалів L , на які розбиваються діапазони значень ознак. Недоліками методу є невизначеність у виборі значення параметра L , а також відсутність зв'язку методу з якістю синтезованої моделі.

Для усунення відзначених недоліків пропонується використовувати метод визначення фрактальної розмірності вхідної множини для навчання розпізнавальної моделі.

Етап ініціалізації. Задати нормовану навчальну вибірку $\langle x, y \rangle$, метод синтезу моделі, критерій якості навчання моделі як функцію помилки E , а також максимально прийнятне значення помилки ε .

Етап формування й аналізу розбиття даних. Послідовно змінюючи значення $L=2, \dots, S$: визначити довжину інтервалу l ; квантувати ознаки вибірки, розбиваючи діапазони їхніх значень на L інтервалів; визначити $n(l)$; побудувати модель net за допомогою заданого методу; оцінити помилку E побудованої моделі net .

Етап визначення фрактальної розмірності. Для всіх l , для яких помилка моделі є прийнятною, визначити фрактальну розмірність даних щодо точності (помилки) синтезованої моделі: $D_{net} = \left\{ n(l) / \log(l^{-1}) \right\} E(net) \leq \varepsilon$.

Єдиним керованим параметром методу є граничне значення помилки моделі ε . Очевидно, що чим меншим буде задане ε , тим більш детальною повинна бути модель, тобто буде потрібно виділити більшу кількість кластерів, а, отже, і більшим буде число L . Перевагою методу є те, що він пов'язаний з показником якості синтезованої моделі, а також в автоматичному режимі встановлює оптимальне значення параметра L .

Розроблені показники реалізовані програмно і досліджені при рішенні задач формування навчальних вибірок даних з вихідних вибірок різної розмірності.

Проведені експерименти підтвердили працездатність запропонованого математичного забезпечення і дозволяють рекомендувати його для використання на практиці при рішенні задач побудови моделей по прецедентах.

Література

1. Dubuc B. Evaluating the fractal dimension of profiles / B. Dubuc, J. Quiniou, C. Roques-Carmes [et al.] // Physical Review A. – 1989. – Vol. 39, № 3. – С. 1500-1512.
2. Camastra F. Data Dimensionality Estimation Methods: A survey / F. Camastra // Pattern Recognition. – 2003. – Vol. 36, No. 12. – P. 2945-2954.
3. Crişan D. A. Fractal Dimension Spectrum as an Indicator for Training Neural Networks // D. A. Crişan, R. Dobrescu // University Politehnica of Bucharest Sci. Bull., Series C. – 2007. – Vol. 69, No. 1. – P. 23-32.