Гавриш О. С., к.ф.-м.н., доцент
Гончаров А. В., к.т.н., доцент
Могілей С. О., phD
Баранов А. Д., студент
Балакін О. М., студент
Черкаський державний технологічний університет

МЕТОД ВИЗНАЧЕННЯ АПРОКСИМУЮЧОЇ МОДЕЛІ ЗАВАДИ З МНОЖИНИ БЛИЗЬКИХ ДО ГАУССІВСЬКИХ ВИПАДКОВИХ ВЕЛИЧИН

Вибір оптимальної моделі апроксимуючої завади може здійснюватися за різними критеріями. В роботі [1] запропоновано обирати апроксимуючу модель за величиною похибки апроксимації розподілу, яка може бути представлена як квадратичне відхилення центрованих нормованих щільностей ймовірностей. Щільність розподілу ймовірностей вихідної випадкової величини може бути задана або аналітично або отримана шляхом апроксимації гістограми, отриманої при обробці вхідних даних. Щільність апроксимуючого розподілу розкладається в ряд Еджворта для її подання послідовністю кумулянтів.

Альтернативним методом вибору моделі близької до гауссівської випадкової величини (БГВВ), що найкращим чином апроксимує реальну заваду, є аналіз чисельних характеристик гістограми розподілу.

Метод визначення моделі з множини БГВВ складається з кількох частин (рис.1):

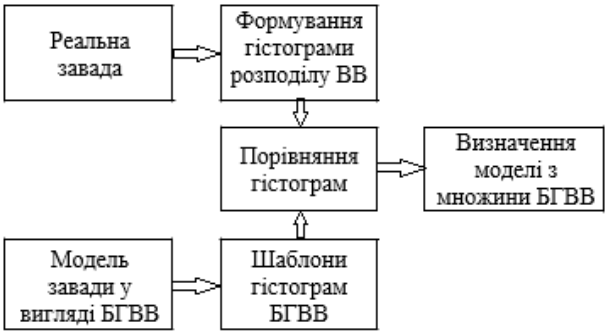

Рисунок 1 – Схема методу визначення моделі завади з множини БГВВ

1. модуль формування гістограми розподілу випадкової величини;
2. модуль, в якому містяться шаблони можливих апроксимуючих моделей, одна з яким буде найкращим чином апроксимувати розподіл реальної завади.
3. модуль порівняння гістограм, в якому розраховується мінімальна відстань між гістограмами;
4. вибір найбільш схожої моделі випадкової величини, з класу БГВВ, на основі даних, отриманих в блоці порівняння гістограм.

Формування гістограми розподілу випадкової величини здійснюється класичним способом і полягає у побудові функції, яка на кожному частинному інтервалі (x_{i-1}; x_i) набуває сталого значення щільності розподілу частот $n_i/(x_i-x_{i-1})$. Таким чином, сформована гістограма розподілу, який описується двовимірним вектором, що містить просторові координати стовпця і значенням ознаки стовпця.

Еталона гістограма формується з врахуванням конкретних значень кумулянтних коефіцієнтів до 6-го порядку. При цьому частина цих параметрів «зануляється» і користувач має кілька еталонних гістограм, одна з яким краще за інші буде описувати реальний розподіл випадкової величини. Таким чином, можна вважати, що в розпорядженні користувача є дві гістограми – тестована (по результатам реальної вибірки) і еталонна (отримана з врахуванням певного набору конкретних значень кумулянтних коефіцієнтів до 6-го порядку). Очевидно, що кількість еталонних гістограм залежить від кількості моделей БГВВ, що використовуються для апроксимації розподілу.

Постає питання, яким чином оцінити, подібність гістограм. Існує кілька підходів до вирішення цього завдання. Припустимо, що відомі обидві гістограми. Часто близькість еталонної і тестованої гістограм вимірюється за допомогою деякої тест статистики, яка забезпечує кількісний вираз «відстань» між гістограмами. Чим менша ця відстань, тим більше подібні гістограми. У літературі існує кілька визначень таких відстаней, наприклад, відстань по Колмогорову, відстань Кульбака-Лейблера, повна варіація функції, хі-квадрат відстань. Зазвичай, це тест статистики, розподіл яких можливо задати формулами чи побудувати методом Монте Карло. Інший шлях - це перетворення гістограм у функції щільності ймовірності та проведення порівняння вже щільностей. Цей підхід заснований на припущенні, що гістограми отримані при вимірі випадкових змінних, які забезпечують основу для оцінки емпіричного розподілу щільності ймовірності. Обчислення відстані між двома щільностями можна розглядати подібно до обчислення байесівської ймовірності. Наприклад, для відстані між двома статистичними сукупностями використовують відстань Бхаттачарія або Хеллінгера.

В даній роботі як дистанція між двома гістограмами використовується метрика earth mover's distance. Earth's mover distance – метрика, яка базується на мінімальній вартості переходу однієї гістограми в іншу. Якщо уявити два розподіли, один як масу землі, яку потрібно розподілити, іншій як набір отворів. Тоді EMD визначає

найменшу кількість роботи, необхідної для заповнення отворів землею. Обчислення EMD базується на вирішенні транспортної задачі лінійного програмування, для вирішення якою існують ефективні алгоритми.

Припустимо, що кілька постачальників кожен із заданою кількістю товарів повинні поставити його кільком споживачам, кожний з обмеженою пропускною здатністю. Для кожної пари постачальник-споживач задана вартість транспортування одиниці товару. Тоді транспортне завдання можна сформулювати як пошук найменш дорогого потоку товарів, що відповідають запитам споживачів.

Порівняння гістограм легко зводиться до транспортної задачі, якщо позначити одну гістограму як постачальника та іншу як споживачів. Цей процес можна розуміти, як пошук мінімальної роботи, необхідної для перетворення однієї гістограми на іншу.

Сформулюємо завдання лінійного програмування: нехай $P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$ – перша гістограма з m стовпців, p_i – стовпець, w_{p_i} – вага стовпця. $Q = \{(q_1, w_{q_1}), ..., (q_n, w_{q_n})\}$ – друга гістограма з n стовпців.

$D = [d_{ij}]$ – матриця, де d_{ij} вартість переходу одиничного значення зі стовпця p_i в стовпець q_j.

Тоді потрібно знайти потік $F = [f_{ij}]$, де f_{ij} - потік між p_i і q_j, що мінімізує загальну вартість $\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}$, що відповідає обмеженням

$$f_{ij} \geq 0, \quad 1 \leq i \leq m, \ 1 \leq j \leq n \tag{1}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{p_i} \quad 1 \leq i \leq m \tag{2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{q_j} \quad 1 \leq j \leq n \tag{3}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right). \tag{4}$$

Обмеження (1) дозволяє пересуватися постачальникам від P до Q і навпаки.

Вимога (2) обмежує сумарні поставки, які можуть бути послані в P їхньою вагою.

Обмеження (3) обмежує Q від кількості прийому постачання більшого ніж їхня вага.

Обмеження (4) визначає максимальну сумарну кількість можливих поставок – сумарний потік.

Вирішивши транспортну задачу, знайдемо оптимальний потік F, тоді EMD визначено як роботу, нормалізовану за сумарним потоком:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}. \tag{5}$$