УДК 004.04 DOI: 10.31388/2220-8674-2018-2-42

МЕТОД ОБРОБКИ ДАНИХ В УМОВАХ СКЛАДНОГО ШУМОВОГО ОТОЧЕННЯ

Вовк С. М., к.ф.-м.н.
Дніпровський національний університет ім. Олеся Гончара,
e-mail: vovk_s_m@ukr.net
Гнатушенко В. В., д.т.н.
Дніпровський національний університет ім. Олеся Гончара,
e-mail: vvgnat@ukr.net

Анотація – запропоновано метод обробки даних в умовах складного шумового оточення, який ґрунтується на методі узагальненої максимальної правдоподібності та «супермножині» вартісних функцій. Ідея методу полягає в налаштуванні процесу обробки даних на поточне шумове оточення шляхом встановлення значень трьох вільних параметрів, пов'язаних з масштабом, важкістю «хвостів» й формою закону розподілу шуму, а також з фактом наявності аномальних значень. У загальному випадку запропонований метод вимагає рішення задачі оптимізації з неунімодальною цільовою функцією та пошуку відповідного локального мінімуму. Представлені результати порівняння ефективності запропонованого методу з методами арифметичного усереднення, медіанної, міріадної та меридіанної фільтрацій.

Ключові слова **- обробка даних, шумове оточення, робастні методи.**

Постановка проблеми. Актуальність обробки даних в умовах складного шумового оточення, обумовленого «забрудненнями» невідомої статистичної природи і, зокрема, аномальними значеннями, визначається різноманіттям моделей даних, шумів та аномалій [1]. Традиційний підхід до обробки даних в зазначених умовах заснований на методі М-оцінювання [2]. В його рамках звичайно виникає проблема вибору вартісної функції, яка має забезпечувати найкраще рішення для поточного шумового оточення й заданої моделі даних. На практиці рішення цієї проблеми ускладнюється через необхідність здійснювати такий вибір серед великої кількості відомих вартісних функцій, багато з яких є дуже схожими одна на одну (наприклад, такими є вартісні функції Тьюкі [2], Хампеля [3], Ендрюса [4] і Мєшалкіна [5]). Для усунення цього недоліку пропонується використати «супермножину» вартісних функцій [6], яка надає можливість генерувати різні вартісні функції з широким діапазоном їх властивостей шляхом встановлення відповідних значень її трьох вільних параметрів. В цьому випадку використання методу

узагальненої максимальної правдоподібності [2] та «супермножини» вартісних функцій з відповідним механізмом їх перетворень [7] призводить до побудови методу, який дозволяє оптимально налаштовувати обробку даних як на шум апріорно відомої статистичної природи з законами узагальнених розподілів Гаусса або Коші, так і на шум з «забрудненнями» невідомої статистичної природи. В останньому випадку налаштування здійснюється методом навчання.

Аналіз попередніх досліджень. Класичні методи обробки даних побудовані на використанні повного статистичного опису даних. Ці методи ґрунтуються на критерії максимуму правдоподібності, в якому функцією правдоподібності є функція спільної щільності імовірності випадкових значень [8]. Відомими результатами цього підходу є методи найменших квадратів та найменших модулів, які отримуються за умови, що випадкові значення є незалежними та ідентично розподіленими за законами Гаусса та Лапласа, відповідно [9]. Останніми досягненнями цього класичного підходу є методи міріадної [10] та меридіанної [8] фільтрації, які отримуються за умови, що випадкові значення є незалежними та ідентично розподіленими за законами Коші та «меридіанним» (або «гостровершинним») законом розподілу, відповідно.

Робастні методи обробки даних побудовані на припущенні неповного статистичного опису даних. Вони призначені для рішення задач в умовах «забрудненого» шуму та ґрунтуються на застосуванні критерію максимуму узагальненої функції правдоподібності, причому ця функція може й не мати імовірнісного трактування [3]. З цієї причини відповідні методи називають методами побудови оцінок «типу максимальної правдоподібності», або скорочено – методами M-оцінювання [2]. Їх основою є вибір відповідної вартісної функції, яка має відповідати поточному шумовому оточенню та моделі даних. В умовах наявності аномальних значень вартісні функції мають відповідати умові B-робастності, коли супремум абсолютного значення функції впливу є обмеженим, та мати скінченне значення точки видалення (англ. rejection point), яка вказує на значення, починаючи з якого забруднення не буде впливати на отримуваний результат [3]. Основними вартісними функціями, які використовуються для цього, є вартісні функції Тьюкі [2], Хампеля [3], Ендрюса [4], Мєшалкіна [5], Демиденка [11] (ця функція також була запропонована в [12]), Гонсалеса-Арса [10] і Аясала-Барнера [8]. Кожна з цих функцій володіє зазначеними вище властивостями та може здійснювати обробку аномальних значень, причому дві останні з них налаштовані на шумове оточення імпульсного типу.

Об'єднання низки вартісних функцій в єдину «супермножину» [6] і побудова відповідного механізму їх перетворень [7] дозволяє одночасно усунути «надмірність» схожих вартісних функцій і забезпечити широкий діапазон можливих рішень, включаючи в якості їх граничних випадків моду, медіану, середнє арифметичне значення, «міріадне» значення і «меридіанне» значення. Можливість побудови «супермножини» визначається відомими формальними перетвореннями функцій (наприклад, перетворення функції Лоренца-Моффата в функцію Гаусса [13]), а також їх асимптотичною поведінкою. Корисність побудови такої супермножини обумовлена можливістю виконувати налаштування методу обробки на невідомий шум та/або перешкоду.

Формулювання цілей статті. Метою даної роботи є розробка методу обробки даних, який грунтується на методі узагальненої максимальної правдоподібності й «супермножині» вартісних функцій та який призначений для обробки даних в умовах складного шумового оточення.

Основна частина. Узвичаєним методом обробки даних, спотворених «брудним» шумом, є метод М-оцінювання [2]. Далі будемо використовувати модель даних у вигляді константи, що в статистичному сенсі призводить до необхідності рішення задачі оцінювання параметра зсуву θ елементів даних x_i; $i=1,...,N$. В рамках методу М-оцінювання ця задача має вигляд [2]:

$$\min_{\theta}[\sum_{i=1}^{N}\psi(x_i-\theta)], \qquad (1)$$

де $\psi(x)$ є заданою вартісною функцією [8], яку також називають функцією втрат або ваговою функцією [11]. Зауважимо, що позначення $\psi(x)$ взяте з [11] і відповідає позначенню $\rho(x)$ в [14]. З (1) видно, що вартісна функція формує цільову функцію задачі мінімізації.

Метод обробки даних, який пропонується, ґрунтується на виборі вартісної функції $\psi(x)$ з супермножини вартісних функцій, котра визначається формулою [6]:

$$\psi_S^{(\alpha,\beta,q)}(x)=k_S^{(\alpha,\beta,q)}[(1+|x/\alpha|^q)^{\beta/q}-1]; \quad -\infty<\beta\le 1, \qquad (2)$$

де $0<q<\infty$; $\beta<q$; $k_S^{(\alpha,\beta,q)}=1/[(1+|x_1/\alpha|^q)^{\beta/q}-1]$, $\psi_S^{(\alpha,\beta,q)}(x_1)=1$ і $x_1=1$. З (2) можна отримати, що супермножина включає велику кількість множин вартісних функцій, у складі яких є квадратична та

модульна вартісні функції, псевдо-Хьюберівська вартісна функція, вартісні функції Аясала-Барнера, Гонсалеса-Арса, Демиденка тощо [7]. Супермножина (2) може бути модифікована шляхом вирівнювання поведінки вартісних функцій в околиці нуля, що розширює діапазон вартісних функцій шляхом включення в нього узагальнених функцій Мєшалкіна [7]. Тоді виходячи з (1) та (2), для дискретного випадку пропонуємий метод полягає у рішенні задачі:

$$\min_{\theta}[\sum_{i=1}^{N}\psi_S^{(\alpha,\beta,q)}(x_i-\theta)]=\min_{\theta}\left\{k_S^{(\alpha,\beta,q)}\sum_{i=1}^{N}\left[\left(1+\left|\frac{x_i-\theta}{\alpha}\right|^q\right)^{\beta/q}-1\right]\right\}, \qquad (3)$$

де значення вільних параметрів α, β та q повинні бути налаштованими на поточне шумове оточення. У загальному вигляді обробка послідовності даних на основі (3) полягає у використанні вікна, яке ковзає послідовністю даних та виробляє вихідні значення.

В табл.1 наведені результати рішення задачі оцінювання параметра зсуву оброблюваних даних запропонованим методом та іншими відомими методами для різних прикладів складного шумового оточення. Ці результати були отримані шляхом числового моделювання спотвореного фрагмента постійної функції, яка дорівнювала одиниці, де довжина фрагмента даних складала 101 дискретний відлік, а спотворення накладалися адитивно.

Табл.1 має наступну будову та позначення. Ліворуч у вертикальному стовпчику вказані імена методів обробки, які застосовувались. У верхньому рядку номерами позначені шість варіантів шумового оточення, які розглядалися. В комірках таблиці вказані значення середньоквадратичної помилки, яка обчислювалась за формулою: $\varepsilon=\sqrt{\frac{1}{J}\sum_{j=1}^{J}|\theta_j-\theta^*|^2}$, де $J=100$ є кількість випадкових реалізацій заданого оточення, $\theta^*=1$ та θ_j позначає j–ту отриману оцінку для θ^*. При цьому для етапів навчання та обробки після значення середньоквадратичної помилки вказані найкращі з отриманих значень вільних параметрів та використаний інтервал пошуку значення параметра зсуву θ.

Таблиця 1

Середньоквадратична помилка для 100 реалізацій

метод обробки	номер прикладу шумового оточення					
	1	2	3	4	5	6
mean	0,371	0,492	0,827	2,216	0,962	1,332
median	0,265	0,229	1,060	0,296	0,282	0,971
myriad training (β=0, q=2)	0,115 $\alpha=10^{-6}$ 0…2	0,081 $\alpha=10^{-3}$ 0…2	0,029 α=0,01 0…2	0,143 $\alpha=10^{-3}$ 0…2	0,085 $\alpha=10^{-4}$ 0…2	0,033 α=0,01 0…2
meridian training (β=0, q=1)	0,115 $\alpha=10^{-6}$ 0…2	0,081 $\alpha=10^{-4}$ 0…2	0,029 α=0,01 0…2	0,145 $\alpha=10^{-3}$ 0…2	0,085 $\alpha=10^{-4}$ 0…2	0,035 α=0,01 0…2
proposed method, training	0,083 α=0,1 β= -16 q=10 0…2	0,048 α=0,1 β= -16 q=10 0…2	0,016 α=0,1 β= -16 q=2 0…2	0,100 α=0,01 β= -16 q=1 0…2	0,054 α=0,01 β= -16 q=1,5 0…2	0,021 α=0,1 β= -4 q=2 0…2
myriad processing	0,108	0,074	0,032	0,154	0,093	0,032
meridian processing	0,109	0,075	0,031	0,154	0,093	0,035
proposed method, processing	0,085	0,062	0,016	0,109	0,072	0,023

Номери варіантів шумового оточення мають наступний сенс. Номер 1 позначає суму гауссівського шуму та завади у вигляді широкого гауссівського імпульсу одиничної амплітуди з рівномірно розподіленим в інтервалі від 30 до 60 дискретних відліків положенням його максимуму та півшириною, яка дорівнює протяжності 15 дискретних відліків. Номер 2 позначає суму гауссівського шуму та завади у вигляді послідовності вузьких додатних гауссівських імпульсів, в яких амплітуда, місцезнаходження та півширина розподілені за рівномірним законом в інтервалах [0, 2], [1, 101] та [0, 2], відповідно. Номер 3 позначає суму гауссівського шуму та завади у вигляді низки викидів з імовірністю їх появи p=0,56 та рівномірним законом розподілу амплітуд в інтервалі [2, 3]. Номери 4, 5 та 6

позначають суму шуму Коші з тими ж самими видами завад, які відповідають номерам 1, 2 та 3.

Методи обробки мають такі умовні назви: mean – метод обчислення арифметичного середнього значення; median – метод обчислення медіанного значення; myriad training – метод обчислення міріадного значення на етапі навчання; meridian training – метод обчислення меридіанного значення на етапі навчання; proposed method, training – запропонований метод на етапі навчання; myriad processing – метод обчислення міріадного значення на основі результатів навчання; meridian processing – метод обчислення меридіанного значення на основі результатів навчання; proposed method, processing – запропонований метод на основі результатів навчання. Під час моделювання етап навчання (training) відрізнявся від етапу обробки (processing) різним початковим значенням датчика випадкових чисел перед формуванням того чи іншого шумового оточення та випадкової завади. Крім того, на етапі навчання виконувався пошук найкращих значень вільних параметрів, які потім використовувались на етапі обробки.

В табл.1 доцільно звернути увагу на те, що значення помилок, які отримані для методів myriad, meridian та proposed на етапі обробки, мають той самий порядок малості, що й значення помилок, які отримані для них на етапі навчання. Крім того, з табл.1 видно, що для перших двох комбінацій метод proposed навчається певній «інтервальній» фільтрації, де параметр згладжування α збігається з параметром масштабу шуму, а значення параметрів $\beta = -16$ та $q = 10$ відповідають вартісній функції, яка за формою є близькою до форми «прямокутної ями». Для третьої комбінації метод proposed навчається фільтрації Мєшалкіна, про що кажуть значення $\beta = -16$, $q = 2$ й $q = 2{,}5$. Нагадаємо, що ці перші три комбінації шуму та завади пов’язані з шумом Гаусса. Для наступних трьох комбінацій шуму та завади, які пов’язані з шумом Коші, налаштування вільних параметрів методу proposed також приводить до гарних результатів. Так, його використання дає приблизно в півтора рази кращий результат, ніж результат налаштування методів myriad та meridian, а також в три, в чотири та в сорок два рази кращий результат, ніж медіанна фільтрація, коли вона застосовується до комбінації шуму Коші з широким гауссівським імпульсом, з низкою вузьких гауссівських імпульсів та з викидами значень (номери 4, 5 та 6, відповідно). Найбільш значущими тут можна вважати результати обробки комбінацій шуму Гаусса та шуму Коші з викидами (номери 3 та 6, відповідно). Оскільки при цьому кількість викидів перевищувала половину кількості елементів даних (бо імовірність появи викидів дорівнювала 0,56), то метод

медіанної фільтрації в якості рішення кожен раз давав значення одного з цих викидів та, відповідно, давав велику помилку. З табл.1 також видно, що налаштування методу proposed для випадку комбінації шуму Коші з викидами (номер 6) привело не до методу міріадної фільтрації, а до методу фільтрації на основі узагальненої вартісної функції Демиденко [7] з параметрами $\alpha = 0{,}1$, $\beta = -4$ та $q = 2$. В цілому можна зазначити, що налаштування вільних параметрів методу proposed забезпечує найкращі результати.

Висновки. Запропонований метод може бути використаний для обробки даних в умовах складного шумового оточення шляхом налаштування його вільних параметрів. Якщо шумове оточення є простим та має відомі статистичні характеристики, то запропонований метод приводить до відповідних для них оптимальних оцінок. Якщо шумове оточення є складним й не має повного статистичного опису, то запропонований метод дозволяє отримувати оцінки, які є більш ефективними у порівнянні з оцінками методів арифметичного усереднення, медіанної, міріадної та меридіанної фільтрацій.

Література

1. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15-58.
2. Huber, P., & Ronchetti, E. M. (2009). *Robust statistics*. 2nd ed. Hoboken: Wiley. doi: 10.1002/9780470434697.
3. Робастность в статистике. Подход на основе функций влияния / Ф. Хампель. – М.: Мир, 1989. – 512 с.
4. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers,W. H., & Tukey, J.W. (1972). *Robust Estimates of Location*. Princeton Univ. Press, Princeton.
5. Meshalkin, L. D. (1973). Some mathematical methods for the study of non-communicable diseases. *Uses of epidemiology in planning of health services. Proc. of the Sixth International Scientific Meeting,* (Primosten, August 29 –September 3, 1971), 250–256.
6. Borulko, V.F., & Vovk, S. M. (2016). Minimum-duration filtering. *Радіоелектроніка, інформатика, управління*, *1*, 7-14. doi: 10.15588/1607-3274-2016-1-1.
7. Vovk, S. M. (2016). General approach to building the methods of filtering based on the minimum duration principle. *Radioelectronics and Communications Systems*, *59*(7), 281-292. doi: 10.3103/S0735272716070013.
8. Aysal, T. C., & Barner, K. E. (2007). Meridian filtering for robust signal processing. *IEEE Trans. on Signal Processing*, *55*(8), 3949–3962.

9. Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, *32*(7), 685–694. doi: 10.1080/02664760500079464.
10. Gonzalez, J. G., & Arce, G. R. (2001). Optimality of the myriad filter in practical impulsive-noise environments. *IEEE Trans. on Signal Processing, 49*(2), 438–441. doi: 10.1109/78.902126.
11. Демиденко Е. З. Оптимизация и регрессия / Е. З. Демиденко. – М.: Наука. – 1989. – 296 с.
12. Вовк С. М. Метод минимума длительности для восстановления финитных сигналов / С. М. Вовк, В. Ф. Борулько // Методы представления и обработки случайных сигналов и полей: тезисы докладов Всесоюзной научно-технической конференции (Туапсе, 10-12 октября 1989 г.). – Харьков, 1989. – С. 98.
13. Trujillo, I., Aguerri, J.A.L., Cepa, J., Gutierrez, C.M. (2001). The effects of seeing on Sersic profiles – II. The Moffat PSF. *Monthly Notices of the Royal Astronomical Society*, *328*(3), 977–985.
14. Huber, P. J. (1984). Finite sample breakdown of M- and P-estimators. *The Annals of Statistics*, *12*(1), 119-126.

МЕТОД ОБРАБОТКИ ДАННЫХ В УСЛОВИЯХ СЛОЖНОГО ШУМОВОГО ОКРУЖЕНИЯ

Вовк С. М., Гнатушенко В. В.

Аннотация

Предложен метод обработки данных в условиях сложного шумового окружения, который базируется на методе обобщенной максимальной правдоподобности и «супермножестве» оценочных функций. Идея метода состоит в настраивании процесса обработки данных на поточное шумовое окружение путем определения значений трех свободных параметров, связанных с масштабом, тяжестью «хвостов» и формой закона распределения шума, а, так же, с фактом наличия аномальных значений. В общем случае предложеный метод требует решения задачи оптимизаци с неунимодальной целевой функцией и поиску соответствующего локального минимума. Представлены результаты сравнения эффективности предложенного метода с методами арифметического усреднения, медианной, мириадной и меридианной фильтраций.

DATA PROCESSING METHOD FOR COMPLEX NOISE ENVIRONMENT

S. Vovk, V. Hnatushenko

Summary

The use of the robust method of M-estimation to data processing under conditions of complex noise environment is complicated by the problem of choosing a cost function that should provide the best solution. To eliminate this shortcoming, it is proposed to use a "superset" of cost functions, which enables the generation of cost functions in a wide range of their properties by setting the corresponding values of three free parameters.

Objective. The purpose of this work is to develop a method of data processing, which is based on the method of generalized maximum likelihood and "superset" of cost functions and which is intended for data processing in conditions of complex noise environment.

Method. The proposed method of data processing under conditions of complex noise environment is based on the method of generalized maximum likelihood and "superset" of cost functions. The idea of the method is to tune the data processing onto the current noise environment by setting the values of three free parameters, which are related to the scale, the heaviness of tails and the form of noise distribution law, as well as the fact of the presence of abnormal values. In the general case, the proposed method requires a solution of the optimization problem with a non-unimodal objective function and finding the appropriate local minimum.

Results. The simulation of the problem of estimating the location parameter of the processed data for various examples of complex noise environment confirmed the performance of the proposed method.

The proposed method can be used to data processing under conditions of complex noise environment by adjusting its free parameters. If the noise environment is simple and has known statistical characteristics, the proposed method leads to the optimal estimates for them. If the noise environment is complex and does not have a complete statistical description, the proposed method allows to obtain estimates that are more effective than those of arithmetic averaging, median, myriad and meridian filtering.

Keywords: data processing, noise environment, robust methods.