УДК 004.032.2

Студент Майданюк О.В., д.т.н., професор Дробишев Ю.П.

**Національний технічний університет України
«Київський політехнічний інститут»**

ОБРОБКА ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ З ПРОПУСКАМИ

Вступ

В експериментальних дослідженнях доволі часто дані мають вигляд таблиць, де рядки відповідають досліджуваним об’єктам, а стовпці – властивостям цих об’єктів (ТОВ - Таблиці «Об’єкт - Властивість»).

В умовах спостереження (вимірювань) з тих чи інших причин (наприклад, вихід з ладу вимірювального пристрою, зміна зовнішніх умов тощо) в ТОВ окремі дані відсутні («пропуски»). Обробка таких ТОВ традиційними математичними методами часто утруднена наявністю неповних рядків і стовпців. Наприклад, при кореляційному аналізі незрозуміло, що розуміти під коефіцієнтом кореляції між властивостями об’єктів, якщо стовпці не повні.

Видаляти неповні рядки (стовпці), як це було на перших етапах обробки таких таблиць з даними, неефективно, оскільки, якщо, наприклад, розмірність ТОВ 100×100 і в кожному рядку одне значення пропущено, то доведеться видалити всю таблицю, в якій 90% вірогідної інформації.

Виникають два шляхи вирішення цієї проблеми:

1) «Відновлення» пропусків (оцінка пропущених значень).
2) Використання методів обробки, нечутливих до пропусків.

Проблема обробки неповних даних вперше була сформульована у роботі Wilks S.S. у 1932 [1]. Далі для обробки даних з пропусками були розроблені різні підходи і методи: регресійний метод (Buck S.F. 1960 [2], Glasser 1964), метод головних компонент (Dear R.E. 1959, Gleason T.C. 1975), метод максимальної правдоподібності (Anderson T.W. 1957), метод пошагової регресії (Frane G.M. 1976), EM-алгоритм (Dempster A.P. and other 1977, Little R.J., Smith P.J. 1987). Короткий огляд зазначених методів дано в [3].

Загальний підхід до проблеми обробки неповних даних

Відновлені значення можуть бути отримані, певна річ, тільки на основі закономірностей, які існують для даної таблиці у вигляді зв'язків між її елементами. Це робить розв'язання задачі відновлення залежною від критерія фіксації закономірностей, що вносить певну частку суб'єктивності в відновлення пропущених даних.

Тому, мабуть доцільно, відновлювати пропуск тільки з єдиною метою: зробити можливим використання відомих ефективних математичних методів, непридатних для обробки неповних рядків та стовпців ТОВ.

Таким чином, закономірність, яка визначає відновлене значення, формулюється на основі лише вірогідних даних і вплив на неї «пропуску» має бути усунений. Тому, закономірність повинна формулюватися на основі найбільш загальних характеристик ТОВ, мало залежних від окремих елементів таблиці, і стійкою по відношенню до них.

Постановка задачі

Метою роботи є створення способу обробки великих масивів експериментальних даних. Елементи масивів представляють собою одновимірні функції визначені на різних інтервалах. Це не дозволяє використовувати традиційні математичні методи обробки.

Розробка алгоритму відновлення експериментальних даних

Дано масив G одновимірних функцій $\{f_i(t)\},\ \ i=1,2,...,N$

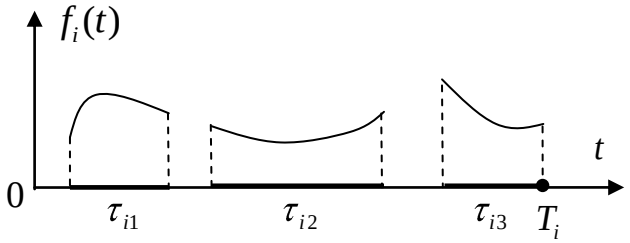

Рис.1. Графіки деяких одновимірних функцій $f_i(t)$

Функція $f_i(t)$ визначена лише на $\tau_i = \bigcup_{s=1}^{l} \tau_{is} < T_i,\ \ t \in [0, T_i]$.

Нехай розв'язується задача компактного зображення масиву G шляхом розкладення функцій $f_i(t)$ за деяким лінійно-незалежним (ортогональним) базисом, наприклад, базисом Карунена-Лоєва, для

одержання якого необхідно знати оцінку коваріаційної функції $K(t,u)$ масиву G [4].

$$K(t,u)=\frac{1}{n-1}\sum_{i=1}^{n}[f_i^{(0)}(t)\cdot f_i^{(0)}(u)], \qquad (1)$$

де $f_i^{(0)}(t)$ - центрована $f_i(t)$, тобто $f_i^{(0)}(t)=f_i(t)-\overline{f_i}$,

f_i - середнє значення $f_i(t)$ на τ_i.

Оскільки, $f_i(t)$ отримані на різноманітних інтервалах τ_i, знаходження $K(t,u)$ на багатьох частинах аргументів t неможливе. Для розв'язання цієї задачі пропонується доповнити "пропущені" на $[0,T]$ значення функцій $f_i(t)$, але таким чином, щоб це не вплинуло на кореляційну функцію, визначену тільки за вірогідними даними масиву G.

Введемо інтервал $[0,T]$, $T=\max_i T_i$. Виберемо деякий повний базис, ортонормований на $[0,T]$, тобто

$$\int_0^T \varphi_k(t)\cdot\varphi_j(t)dt=\delta_{kj}=\begin{cases}1, & k=j\\ 0, & k\neq j\end{cases}$$

Якщо $f_i(t)$ визначена на $[0,T]$, то з необхідною точністю її можна виразити відрізком узагальненого ряду Фурьє

$$f(t)\approx\sum_{k=1}^{m}a_k\cdot\varphi_k(t), \qquad (2)$$

де a_k - коефіцієнти Фурьє.

$$a_k=\int_0^T f(t)\cdot\varphi_k(t) \qquad (3)$$

Оскільки $f_i(t)$ визначена на $\tau_i<T$, то формула (3) незастосовна для визначення вектора коефіцієнтів $\overrightarrow{a^{(i)}}$. Проте, можна спробувати знайти $\overrightarrow{a^{(i)}}$ з умови, наслідком якої і є узагальнена формула Фурьє, а власне

$$\min_{\overrightarrow{a^{(i)}}}\int_{\tau_i}[(f_i(t)-\sum_{k=1}^{m}a_k^{(i)}\cdot\varphi_k(t)]^2dt \qquad (4)$$

Мінімізація функціоналу (4) приводить до матричного рівняння відносно вектора $\overrightarrow{a^{(i)}}$ для кожної функції $f_i(t)$, $i=1,2,...,n$

$$Q^{(i)}\cdot\overrightarrow{a^{(i)}}=\overrightarrow{b^{(i)}} \qquad (5)$$

Матриця $Q^{(i)}=\{q_{kj}^{(i)}\}$, $\overrightarrow{b^{(i)}}=\{b_j^{(i)}\}$.

Відповідно $q_{kj}^{(i)}=\int_{\tau_i}\varphi_k(t)\cdot\varphi_j(t)dt$, $b_j^{(i)}=\int_{\tau_i}f_i(t)\cdot\varphi_j(t)dt$.

Знайшовши вектор $\overrightarrow{a^{(i)}}$ з (5), будуємо аналітичне продовження $\tilde{f}(t),\ t \in [0,T]$ функції $f_i(t)$ за формулою $\tilde{f}(t) = \sum_{k=1}^{m} a_k^{(i)} \cdot \varphi_k(t),\ t \in [0,T]$.

Масив функцій $\{\tilde{f}(t)\}$ використовується для побудови коваріаційної функції $K(t,u)$ на квадраті $[T \times T]$ замість $\{f_i(t)\}$.

Висновки

Метод відновлення даних дозволяє застосовувати існуючі математичні методи, не орієнтовані на обробку таблиць з неповними даними, для обробки експериментальних даних з пропусками.

Перспективи подальших розвідок у цьому напрямку стосуються розробки методів обробки, нечутливих до пропусків.

Література

1. *Wilks S.S* "Moments and distributions of estimates of population from fragmentary samples" // Ann. Math. Statist. 1932. V. 3. P.163-195.
2. *S. F. Buck* "A method of Estimation of Missing Values in Multivariate Data suitable for use with an Electronic Computer", 1960.
3. *Н.Г. Загоруйко* "Прикладные методы анализа данных и знаний", Новосибирск, Издательство Института математики, 1999 г.
4. *Свешников А.А.* "Прикладные методы теории случайных функций", Издательство Наука ГРФМЛ, М. 1968.