DOI: 10.18372/2310-5461.42.13758

УДК 621.396:51-74 (045)

О. Г. Голубничий**,** канд. техн. наук, доц.
Національний авіаційний університет
orcid.org/0000-0001-5101-3862
a.holubnychyi@nau.edu.ua

АНАЛІЗ ОСОБЛИВОСТЕЙ РЕАЛІЗАЦІЇ ЕМ-АЛГОРИТМУ ПРИ КЛАСТЕРИЗАЦІЇ СИСТЕМ СИГНАЛЬНИХ КОНСТРУКЦІЙ

Вступ

Одними з класів задач обробки сигналів та даних є задачі кластерного аналізу [1] та структурного прогнозування [2].

Задачі кластеризації полягають у алгоритмічній класифікації моделей, первинними даними для яких є результати спостережень, елементи даних та системи ознак, на групи (кластери).

Цей клас задач, як правило, відносять до складних комбінаторних задач, які характеризуються наявністю припущень та контекстністю розв'язуваної задачі, а також до задач статистичного розпізнавання образів, задач машинного навчання [1].

Задачі структурного прогнозування полягають у виявленні структурованості даних, вивченні функціональних залежностей між складовими елементами оброблюваних даних.

Цей клас задач відносять до задач машинного навчання [2].

Задачі кластерного аналізу та структурного прогнозування доповнюють одна одну при розв'язанні ряду практичних задач у галузі обробки сигналів та даних, пов'язаних з виявленням структур у досліджуваних системах даних та сигнально-кодових конструкцій, взаємозв'язків між ними, оцінюванні їх параметрів.

Одним з відомих та розповсюджуваних методів оцінювання параметрів моделей та кластеризації, який використовується при кластерному аналізі та структурному прогнозуванні, є expectation-maximization (EM) алгоритм [3–5].

Цей алгоритм та його модифікації використовуються для оцінювання параметрів компонент сумішей у вигляді гаусівських змішаних моделей, якими представляються досліджувані дані. ЕМ-алгоритм також дозволяє здійснювати кластеризацію об'єктів, якими є елементи досліджуваних даних, на основі отримуваних апостеріорних оцінок імовірностей приналежності об'єктів кластерам (компонентам гаусівської змішаної моделі) [5–7].

Постановка проблеми

До особливостей реалізації ЕМ-алгоритму при розв'язанні різних задач оцінювання параметрів, кластерного аналізу та структурного прогнозування, які суттєво можуть вплинути як на результат розв'язання задачі, так і на обчислювальну та алгоритмічну складність процесу розв'язання задачі, можна віднести такі:

1) чутливість ЕМ-алгоритму до початкового наближення оцінюваних параметрів моделі, що може проявлятися у розв'язанні задачі у локальних екстремумах логарифмічної функції правдоподібності (гіпотези щодо параметрів моделі), максимізація якої є метою ітераційної процедури ЕМ-алгоритму;

2) апріорна невизначеність кількості компонент суміші, наприклад гаусівської змішаної моделі, параметри якої оцінюються при реалізації ЕМ-алгоритму, якщо у задачі кластеризації, яка розв'язується з його використанням, кількість кластерів є апріорі невідомою;

3) необхідність введення та використання додаткових критеріїв додавання або видалення компонент (кластерів) у модифікаціях ЕМ-алгоритму з додаванням або видаленням компонент, від яких залежать ймовірності виникнення помилок 1-го та 2-го роду при кластеризації;

4) необхідність для багатьох задач враховувати контекст та фізичний зміст розв'язуваної задачі при інтерпретації отриманих розв'язків;

5) можливе виникнення при реалізації алгоритму невизначеностей (або сингулярностей), які підлягають розкриттю та контекстному поясненню відповідно до змісту розв'язуваної задачі з метою корекції та продовження реалізації ітераційної процедури ЕМ-алгоритму.

Стаття присвячена аналізу особливостей реалізації ЕМ-алгоритму, які можуть виникати при кластеризації систем сигнальних конструкцій та є пов'язаними з утворенням невизначеностей (на практиці ідентифікується програмно діленням на "0") у логарифмічній функції правдоподібності через виокремлення ЕМ-алгоритмом порожніх кластерів, що унеможливлює алгоритмічне продовження реалізації подальших ітерацій ЕМ-алгоритму до його збіжності.

Аналіз останніх досліджень і публікацій

До останніх досліджень та публікацій, в яких для розв'язання теоретичних та прикладних задач використовується ЕМ-алгоритм та його модифікації, можна віднести дослідження у галузі розробки методів адаптивного прогнозування надійності технічних систем [8], метод багатосенсорного відстеження космічних об'єктів, у якому ЕМ-алгоритм використовується для апроксимації статистичних характеристик у результатах спостережень [9], дослідження у галузі систем масового обслуговування, які пов'язані з розробкою методів оцінювання параметрів у системах типу $M_t/M/1/K$ з моделлю неоднорідного пуассонівського потоку, у яких ЕМ-алгоритм використовується для отримання оцінок за методом максимальної правдоподібності [10] тощо. До типових ознак цих та інших підходів та методів, у яких використовується ЕМ-алгоритм, можна віднести статистичний характер розв'язуваної задачі та отримання оцінок параметрів використовуваних для їх розв'язання статистичних моделей на основі принципу максимальної правдоподібності.

Постановка завдання

Метою статті є аналіз особливостей реалізації ЕМ-алгоритму, які виникають через виокремлення ЕМ-алгоритмом порожніх кластерів при розв'язанні задачі кластеризації систем сигнальних конструкцій та пов'язані з утворенням невизначеностей (математичних сингулярностей) у логарифмічній функції правдоподібності, максимізація якої проводиться під час ітераційної процедури ЕМ-алгоритму.

Виклад основного матеріалу дослідження

Пояснимо проблемну ситуацію, яка зазначена у постановці завдання, на прикладі.

Нехай у системі обробки сигналів спостерігаються гармонічні сигнали $x_i(t)$, $i=\overline{1,10}$, які відрізняються один від одного фазами та можуть бути представлені моделлю (1).

$$x(t)=\sin\left(\frac{2\pi}{T}t+\Delta\varphi\right), \qquad (1)$$

де T – період коливань;

$$\Delta\varphi_1=-\frac{11}{180}\pi\ \left(-11^0\right),\ i=1;$$

$$\Delta\varphi_2=-\frac{3}{180}\pi\ \left(-3^0\right),\ i=2;$$

$$\Delta\varphi_3=+\frac{1}{180}\pi\ \left(+1^0\right),\ i=3;$$

$$\Delta\varphi_4=+\frac{6}{180}\pi\ \left(+6^0\right),\ i=4;$$

$$\Delta\varphi_5=+\frac{13}{180}\pi\ \left(+13^0\right),\ i=5;$$

$$\Delta\varphi_6=+\frac{42}{180}\pi=-\frac{33}{180}\pi+\frac{75}{180}\pi=$$
$$=3\Delta\varphi_1+\frac{75}{180}\pi\ \left(+42^0=-33^0+75^0\right),\ i=6;$$

$$\Delta\varphi_7=+\frac{66}{180}\pi=-\frac{9}{180}\pi+\frac{75}{180}\pi=$$
$$=3\Delta\varphi_2+\frac{75}{180}\pi\ \left(+66^0=-9^0+75^0\right),\ i=7;$$

$$\Delta\varphi_8=+\frac{78}{180}\pi=+\frac{3}{180}\pi+\frac{75}{180}\pi=$$
$$=3\Delta\varphi_3+\frac{75}{180}\pi\ \left(+78^0=+3^0+75^0\right),\ i=8;$$

$$\Delta\varphi_9=+\frac{93}{180}\pi=+\frac{18}{180}\pi+\frac{75}{180}\pi=$$
$$=3\Delta\varphi_4+\frac{75}{180}\pi\ \left(+93^0=+18^0+75^0\right),\ i=9;$$

$$\Delta\varphi_{10}=+\frac{114}{180}\pi=+\frac{39}{180}\pi+\frac{75}{180}\pi=$$
$$=3\Delta\varphi_5+\frac{75}{180}\pi\ \left(+114^0=+39^0+75^0\right),\ i=10.$$

Параметри сигналів у прикладі було обрано таким чином, щоб $x_i(t)$, $i=\overline{1,10}$, могли утворювати кластери $\{x_1(t),...,x_5(t)\}$ та $\{x_6(t),...,x_{10}(t)\}$ за ознакою фази, що також проілюстровано на рис. 1.

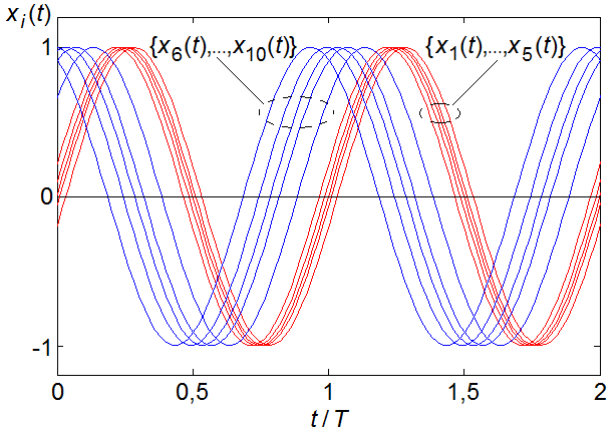

Рис. 1. Сигнали з прикладу, який розглядається

Під час аналізу сигналів при їх обробці в умовах апріорної невизначеності щодо структури та параметрів $x_i(t)$, $i=\overline{1,10}$, може бути поставлена задача кластеризації сигналів за такою ознакою кластеризації як взаємна кореляція між сигналами на спостережуваному інтервалі часу. Тоді вхідними даними для кластерного аналізу будуть значення нормованої взаємної кореляційної функції $r_{i,j}$, $i=\overline{1,10}$, $j=\overline{1,10}$, яка оцінюється, наприклад, на спостережуваному інтервалі часу $t\in(0,2T)$ з припущенням, що математичне сподівання процесів $M\{x_i(t)\}=0$:

$$r_{i,j}=\frac{1}{2\sigma\{x_i(t)\}\sigma\{x_j(t)\}T}\int_0^{2T}x_i(t)x_j(t)dt,$$

де $\sigma\{x_i(t)\}$, $\sigma\{x_j(t)\}$ — середньоквадратичні відхилення процесів $x_i(t)$ та $x_j(t)$ відповідно.

Для прикладу сигнальних конструкцій, який розглядається з використанням моделі (1), значення $r_{i,j}$, що містять інформацію про кореляції між усіма парами різних сигнальних конструкцій, можуть бути представлені таким чином:

$$\begin{aligned}\tilde{\mathbf{r}}=(\tilde{r}_n)=(&r_{1,2};\ r_{1,3};\ r_{1,4};\ r_{1,5};\ r_{1,6};\ r_{1,7};\ r_{1,8};\ r_{1,9};\ r_{1,10};\\ &r_{2,3};\ r_{2,4};\ r_{2,5};\ r_{2,6};\ r_{2,7};\ r_{2,8};\ r_{2,9};\ r_{2,10};\ r_{3,4};\\ &r_{3,5};\ r_{3,6};\ r_{3,7};\ r_{3,8};\ r_{3,9};\ r_{3,10};\ r_{4,5};\ r_{4,6};\ r_{4,7};\\ &r_{4,8};\ r_{4,9};\ r_{4,10};\ r_{5,6};\ r_{5,7};\ r_{5,8};\ r_{5,9};\ r_{5,10};\ r_{6,7};\\ &r_{6,8};\ r_{6,9};\ r_{6,10};\ r_{7,8};\ r_{7,9};\ r_{7,10};\ r_{8,9};\ r_{8,10};\ r_{9,10}),\\ &n=\overline{1,N},\ N=45.\end{aligned}$$

При вказаних вище параметрах моделі (1) вектор-рядок $\tilde{\mathbf{r}}$ матиме такий вигляд:

$$\begin{aligned}\tilde{\mathbf{r}}=(&0{,}990;\ 0{,}978;\ 0{,}956;\ 0{,}914;\ 0{,}602;\\ &0{,}225;\ 0{,}017;\ -0{,}242;\ -0{,}574;\ 0{,}998;\\ &0{,}988;\ 0{,}961;\ 0{,}707;\ 0{,}358;\ 0{,}156;\\ &-0{,}105;\ -0{,}454;\ 0{,}996;\ 0{,}978;\ 0{,}755;\\ &0{,}423;\ 0{,}225;\ -0{,}035;\ -0{,}391;\ 0{,}993;\\ &0{,}809;\ 0{,}500;\ 0{,}309;\ 0{,}052;\ -0{,}309;\\ &0{,}875;\ 0{,}602;\ 0{,}423;\ 0{,}174;\ -0{,}191;\\ &0{,}914;\ 0{,}809;\ 0{,}629;\ 0{,}309;\ 0{,}978;\\ &0{,}891;\ 0{,}669;\ 0{,}966;\ 0{,}809;\ 0{,}934).\end{aligned}\qquad(2)$$

Використання для кластеризації ЕМ-алгоритму полягає у статистичному аналізі $\tilde{\mathbf{r}}$ як даних, які відповідають певній суміші розподілів. Однією з моделей таких сумішей є гаусівська змішана модель [5–7], яка складається з гаусівських компонент. Для випадку $\tilde{\mathbf{r}}$ ця модель формалізована виразом (3).

$$p(\tilde{r})=\sum_{k=1}^{K}\frac{\theta_k}{\sqrt{2\pi\sigma_k^2}}\exp\left[-\frac{(\tilde{r}-m_k)^2}{2\sigma_k^2}\right],\qquad(3)$$

де $p(\tilde{r})$ — щільність розподілу ймовірностей неперервної випадкової величини $\tilde{r}$; K — кількість компонент суміші; θ_k — ваговий коефіцієнт k-ї гаусовської компоненти у складі суміші $p(\tilde{r})$; $\sum_{k=1}^{K}\theta_k=1$; m_k — математичне сподівання k-ї гаусівської компоненти; σ_k — середньоквадратичне відхилення k-ї гаусівської компоненти; $k=\overline{1,K}$.

ЕМ-алгоритм дозволяє для спостережуваних даних $\tilde{\mathbf{r}}$ оцінити параметри гаусовської змішаної моделі $\boldsymbol{\omega}=\{\boldsymbol{\theta},\mathbf{m},\boldsymbol{\sigma}\}$, а також апостеріорні ймовірності $\gamma_{n,k}$ приналежності елемента $\tilde{r}_n$ k-й компоненті суміші, які використовуються для прийняття рішення про приналежність елемента тому чи іншому кластеру.

Ітераційна процедура ЕМ-алгоритму [4, с.252] складається з кроку E (expectation), на якому оцінюються значення $\gamma_{n,k}$ при поточному наближенні параметрів моделі, та кроку M (maximization), на якому уточнюються параметри моделі, максимізуючи у поточній ітерації логарифмічну функцію правдоподібності (гіпотези щодо параметрів моделі), яка формалізована для гаусівської змішаної моделі таким чином:

$$L(\boldsymbol{\omega}\,|\,\tilde{\mathbf{r}})=\sum_{n=1}^{N}\ln\sum_{k=1}^{K}\frac{\theta_k}{\sqrt{2\pi\sigma_k^2}}\exp\left[-\frac{(\tilde{r}_n-m_k)^2}{2\sigma_k^2}\right].$$

Формалізуємо процедуру ЕМ-алгоритму на його s-й ітерації (s = 1, 2, 3,…) для задачі кластеризації, яка розглядається у статті:

1) Е-крок: при наближенні параметрів моделі

$$\begin{aligned}\boldsymbol{\omega}^{\langle s-1\rangle}&=\left\{\boldsymbol{\theta}^{\langle s-1\rangle},\mathbf{m}^{\langle s-1\rangle},\boldsymbol{\sigma}^{\langle s-1\rangle}\right\}=\\&=\left(\theta_1^{\langle s-1\rangle},...,\theta_K^{\langle s-1\rangle},m_1^{\langle s-1\rangle},...,m_K^{\langle s-1\rangle},\sigma_1^{\langle s-1\rangle},...,\sigma_K^{\langle s-1\rangle}\right)\end{aligned}$$

оцінюються значення

$$\gamma_{n,k}^{\langle s-1\rangle}=\frac{\dfrac{\theta_k^{\langle s-1\rangle}}{\sigma_k^{\langle s-1\rangle}\sqrt{2\pi}}\exp\left[-\dfrac{(\tilde{r}_n-m_k^{\langle s-1\rangle})^2}{2(\sigma_k^{\langle s-1\rangle})^2}\right]}{\sum_{l=1}^{K}\dfrac{\theta_l^{\langle s-1\rangle}}{\sigma_l^{\langle s-1\rangle}\sqrt{2\pi}}\exp\left[-\dfrac{(\tilde{r}_n-m_l^{\langle s-1\rangle})^2}{2(\sigma_l^{\langle s-1\rangle})^2}\right]},$$

а також оцінюється кількість $N_k^{\langle s-1\rangle}$ елементів $\tilde{\mathbf{r}}$, яка належить *k*-й компоненті суміші:

$$N_k^{\langle s-1\rangle} = \sum_{n=1}^{N} \gamma_{n,k}^{\langle s-1\rangle}. \qquad (4)$$

У ітерації $s = 1$ використовується початкове наближення параметрів моделі:

$$\boldsymbol{\omega}^{\langle 0\rangle} = \left\{\boldsymbol{\theta}^{\langle 0\rangle}, \mathbf{m}^{\langle 0\rangle}, \boldsymbol{\sigma}^{\langle 0\rangle}\right\} =$$
$$= \left(\theta_1^{\langle 0\rangle}, ..., \theta_K^{\langle 0\rangle}, m_1^{\langle 0\rangle}, ..., m_K^{\langle 0\rangle}, \sigma_1^{\langle 0\rangle}, ..., \sigma_K^{\langle 0\rangle}\right).$$

2) М-крок: визначаються уточнені оцінки параметрів моделі

$$\boldsymbol{\omega}^{\langle s\rangle} = \left\{\boldsymbol{\theta}^{\langle s\rangle}, \mathbf{m}^{\langle s\rangle}, \boldsymbol{\sigma}^{\langle s\rangle}\right\} =$$
$$= \left(\theta_1^{\langle s\rangle}, ..., \theta_K^{\langle s\rangle}, m_1^{\langle s\rangle}, ..., m_K^{\langle s\rangle}, \sigma_1^{\langle s\rangle}, ..., \sigma_K^{\langle s\rangle}\right),$$

які у *s*-й ітерації максимізують $L\left(\boldsymbol{\omega} \mid \tilde{\mathbf{r}}\right)$:

$$\theta_k^{\langle s\rangle} = \frac{N_k^{\langle s-1\rangle}}{N},$$

$$m_k^{\langle s\rangle} = \frac{1}{N_k^{\langle s-1\rangle}} \sum_{n=1}^{N} \gamma_{n,k}^{\langle s-1\rangle} \tilde{r}_n, \qquad (5)$$

$$\sigma_k^{\langle s\rangle} = \sqrt{\frac{1}{N_k^{\langle s-1\rangle}} \sum_{n=1}^{N} \gamma_{n,k}^{\langle s-1\rangle} \left(\tilde{r}_n - m_k^{\langle s\rangle}\right)^2}.$$

У випадку відсутності виникнення невизначеностей (математичних сингулярностей) у ітераційній процедурі алгоритму вона виконується до збіжності, практичним критерієм якої може бути умова $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \tilde{\mathbf{r}}\right) - L\left(\boldsymbol{\omega}^{\langle s-1\rangle} \mid \tilde{\mathbf{r}}\right) < \varepsilon$, де $\varepsilon > 0$ — число, при якому точність оцінювання параметрів моделі вважається достатньою в контексті збіжності алгоритму.

Використовуватимемо у прикладі, який розглядається, такі параметри ($k = \overline{1, K}$):

1) $K = 11$;

2) $\theta_k^{\langle 0\rangle} = 1/K = 1/11$;

3) $m_k^{\langle 0\rangle} = -1 + (2k-1)/K = -1 + (2k-1)/11$;

4) $\sigma_k^{\langle 0\rangle} = 1/3K = 1/33$.

Вибір кількості компонент гаусівської змішаної моделі $K = 11$ було обрано таким чином, щоб вона мала достатньо велике значення у порівнянні з об'ємом вибірки ($N = 45$), що спричинить виявлення у ітераційній процедурі ЕМ-алгоритму принаймні одного порожнього кластеру та відповідної невизначеності (математичної сингулярності), яка робить неможливою подальшу алгоритмічну реалізацію ітераційної процедури ЕМ-алгоритму.

Вибір початкового наближення параметрів моделі $\boldsymbol{\omega}^{\langle 0\rangle}$ було обрано так, щоб вагові коефіцієнти $\theta_k^{\langle 0\rangle}$ були апріорі однаковими, у діапазоні можливих значень нормованої взаємної кореляційної функції розподіл математичних сподівань $m_k^{\langle 0\rangle}$ гаусовських компонент був рівномірним, а також використовувалося правило $\pm 3\sigma_k^{\langle 0\rangle}$ для їх середньоквадратичних відхилень.

Реалізація ЕМ-алгоритму для даних $\tilde{\mathbf{r}}$, які вказано у (2), а також при зазначених вище *K* та $\boldsymbol{\omega}^{\langle 0\rangle}$, дає такі результати процесу максимізації $L\left(\boldsymbol{\omega} \mid \tilde{\mathbf{r}}\right)$ у ітераціях: $L\left(\boldsymbol{\omega}^{\langle 0\rangle} \mid \tilde{\mathbf{r}}\right) = -68{,}9$, $L\left(\boldsymbol{\omega}^{\langle 1\rangle} \mid \tilde{\mathbf{r}}\right) = -8{,}64$, $L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \tilde{\mathbf{r}}\right) = 0/0$.

Виникнення невизначеності (математичної сингулярності) у $L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \tilde{\mathbf{r}}\right)$ пов'язане з тим, що утворюється порожній кластер $k = 1$, для якого виконується умова $\forall n \in \overline{1, N}$, $k = 1$, $\gamma_{n,k}^{\langle 1\rangle} \to 0$.

Проаналізуємо цю невизначеність для загального випадку утворення у *s*-й ітерації *k*-го порожнього кластеру, тобто $\lim\limits_{\substack{\gamma_{n,k}^{\langle s-1\rangle} \to 0, \\ n=\overline{1,N}}} L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \tilde{\mathbf{r}}\right)$.

Представляючи $\boldsymbol{\omega}^{\langle s\rangle}$ через $\boldsymbol{\gamma}^{\langle s-1\rangle}$ з використанням (4) та (5), отримаємо:

$$\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle} \to 0, \\ n=\overline{1,N}}} L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \tilde{\mathbf{r}}\right) = \lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle} \to 0, \\ n=\overline{1,N}}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \frac{\theta_k^{\langle s\rangle}}{\sqrt{2\pi\left(\sigma_k^{\langle s\rangle}\right)^2}} \exp\left[-\frac{\left(\tilde{r}_n - m_k\right)^2}{2\left(\sigma_k^{\langle s\rangle}\right)^2}\right] \equiv$$

$$\equiv \lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle} \to 0, \\ n=\overline{1,N}}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \frac{\left(\sum\limits_{p=1}^{N} \gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{N\sqrt{2\pi \sum\limits_{p=1}^{N} \gamma_{p,k}^{\langle s-1\rangle} \left(\tilde{r}_p - m_k^{\langle s\rangle}\right)^2}} \exp\left[-\frac{\left(\tilde{r}_n - m_k^{\langle s\rangle}\right)^2 \sum\limits_{p=1}^{N} \gamma_{p,k}^{\langle s-1\rangle}}{2\sum\limits_{p=1}^{N} \gamma_{p,k}^{\langle s-1\rangle} \left(\tilde{r}_p - m_k^{\langle s\rangle}\right)^2}\right].$$

Проаналізуємо $A_{n,k}=\lim\limits_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\dfrac{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{N\sqrt{2\pi\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}}\exp\left[-\dfrac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}}{2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}\right]$. Для зручності використаємо логарифмування та властивість границі неперервної функції, тобто знайдемо

$$A_{n,k}=\exp\left(\ln A_{n,k}\right)=\exp\left[\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\ln\left\{\frac{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{N\sqrt{2\pi\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}}\exp\left[-\frac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}}{2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}\right]\right\}\right]=$$

$$=\exp\left[\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\ln\frac{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{\frac{1}{2}}}-\ln N-\frac{1}{2}\ln\left(2\pi\right)-\frac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}}{2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}\right)\right].$$

Для $\lim\limits_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\dfrac{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{\frac{1}{2}}}\right)$ використаємо правило Лопіталя, враховуючи також те, що для порожнього k-го кластеру $\widetilde{r}_n-m_k^{\langle s\rangle}\neq 0$:

$$\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\frac{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{\frac{1}{2}}}\right)=\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\frac{\dfrac{\partial}{\partial\gamma_{n,k}^{\langle s-1\rangle}}\left\{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}\right\}}{\dfrac{\partial}{\partial\gamma_{n,k}^{\langle s-1\rangle}}\left\{\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{\frac{1}{2}}\right\}}\right)=$$

$$=\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\frac{\frac{3}{2}\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{1}{2}}}{\frac{1}{2}\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{-\frac{1}{2}}}\right)=\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\frac{3\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{1}{2}}\left(\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{\frac{1}{2}}}{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2}\right)=0.$$

Для $\lim\limits_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\dfrac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}}{2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}$ також використаємо правило Лопіталя:

$$\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\frac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}}{2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}=\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\frac{\dfrac{\partial}{\partial\gamma_{n,k}^{\langle s-1\rangle}}\left\{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right\}}{\dfrac{\partial}{\partial\gamma_{n,k}^{\langle s-1\rangle}}\left\{2\sum\limits_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right\}}=\frac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2}{2\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2}=\frac{1}{2}.$$

З урахуванням цього

$$\lim_{\substack{\gamma_{n,k}^{\langle s-1\rangle}\to 0,\\ n=\overline{1,N}}}\left(\ln\frac{\left(\sum_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\right)^{\frac{3}{2}}}{\left(\sum_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2\right)^{\frac{1}{2}}}-\ln N-\frac{1}{2}\ln(2\pi)-\frac{\left(\widetilde{r}_n-m_k^{\langle s\rangle}\right)^2\sum_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}}{2\sum_{p=1}^{N}\gamma_{p,k}^{\langle s-1\rangle}\left(\widetilde{r}_p-m_k^{\langle s\rangle}\right)^2}\right)=-\infty,$$

$A_{n,k}=\exp\left(\ln A_{n,k}\right)=\exp(-\infty)=0$, $n=\overline{1,N}$, для окремо взятого k-го порожнього кластеру.

Таким чином, у випадку утворення в ітераційній процедурі ЕМ-алгоритму порожнього k-го кластеру та відповідних невизначеностей (математичних сингулярностей) для їх усунення необхідне виключення відповідної k-ї виродженої гаусівської компоненти на s-й ітерації ЕМ-алгоритму зі складу $L\left(\boldsymbol{\omega}^{\langle s\rangle}\,|\,\tilde{\mathbf{r}}\right)$, що не вплине на значення $L\left(\boldsymbol{\omega}^{\langle s\rangle}\,|\,\tilde{\mathbf{r}}\right)$ через те, що для окремо взятого порожнього k-го кластеру $A_{n,k}=0$, $n=\overline{1,N}$.

Зауважимо, що у цій статті не наводиться повне розв'язання задачі кластеризації для використаного прикладу аналізу сигнальних конструкцій, а він служить виключно для пояснення можливого випадку виникнення математичної сингулярності в ітераційній процедурі ЕМ-алгоритму через утворення порожніх кластерів з метою її аналізу. Також зазначимо, що в ітераційній процедурі ЕМ-алгоритму можуть виникати невизначеності (математичні сингулярності) інших типів, які не пов'язані з утворенням порожніх кластерів та аналіз яких потребує окремого розгляду.

Висновки

Однією з особливостей реалізації ЕМ-алгоритму при розв'язанні задач кластеризації систем сигнальних конструкцій є можливе виникнення стану математичної сингулярності у його ітераційній процедурі.

У статті показано приклад можливої задачі кластеризації сигнальних конструкцій, підхід до розв'язання якої полягає у аналізі взаємнокореляційних зв'язків між сигналами з використанням гаусівської змішаної моделі та оцінювання її параметрів і прихованих змінних (імовірностей приналежності елементів суміші певним компонентам цієї суміші, що є визначальним при прийнятті рішення про приналежність того чи іншого елемента певному кластеру) з використанням ЕМ-алгоритму.

Розглянута у статті задача характеризується математичною сингулярністю, яка пов'язана з утворенням порожніх кластерів. У результаті аналізу математичної сингулярності такого типу (невизначеність "0/0") показано, що можливе видалення утвореного порожнього кластеру таким чином, що структура та значення логарифмічної функції правдоподібності корегуються до таких, які були б у випадку апріорної відсутності зазначеного порожнього кластеру.

Проаналізована у статті особливість є більш властивою для модифікації ЕМ-алгоритму з видаленням компонент, а також у випадках відносно невеликої кількості елементів, які підлягають кластеризації, або відносно великої кількості кластерів, значення якої є структурним параметром ЕМ-алгоритму.

Залежно від задачі кластеризації, розглянута особливість ЕМ-алгоритму може бути використана при кластерному аналізі складних конструкцій сигналів та даних, наприклад мультиплікативно комплементарних бінарних сигнально-кодових конструкцій [11], структур їх кореляційних функцій [12], систем корельованих сигнальних конструкцій при неортогональному множинному доступі (NOMA) у системах мобільних телекомунікацій 5G [13] тощо.

ЛІТЕРАТУРА

1. **Jain A. K.,** Murty M. N., Flynn P. J. Data clustering: a review. *ACM Computing Surveys*. 1999. Vol. 31. No. 3. P. 264-323. DOI: 10.1145/331499.331504.

2. **Bakir G. H.** *et al*. Predicting Structured Data (Neural Information Processing series) 1st Edition. The MIT Press, 2007. 360 p.

3. **Dempster A. P.,** Laird N. M., Rubin D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*. 1977. Vol. 39. No. 1. P. 1-38.

4. **Gupta M. R.,** Chen Y. Theory and Use of the EM Algorithm. *Foundations and Trends® in Signal Processing*. 2011. Vol. 4. No. 3. P. 223-296. DOI: 10.1561/2000000034.

5. **Vlassis N.,** Likas A. A Greedy EM Algorithm for Gaussian Mixture Learning. *Neural Processing Letters*. 2002. Vol. 15. P. 77-87.

6. **Yu D.,** Deng L. Gaussian Mixture Models. *Automatic Speech Recognition. A Deep Learning Approach*. London: Springer-Verlag, 2015. P. 13-21. DOI: 10.1007/978-1-4471-5779-3_2

7. **Huang T.,** Peng H., Zhang K. Model Selection for Gaussian Mixture Models. *Statistica Sinica*. 2017. Vol. 27. P. 147-169.

8. **Wang Xi** *et al*. An Adaptive Prognostic Approach for Newly Developed System with Three-Source Variability. *IEEE Access*. 2019. Vol. 7. P. 53091-53102. DOI: 10.1109/ACCESS. 2019. 2911307

9. **Wei B.,** Nener B. D. Multi-Sensor Space Debris Tracking for Space Situational Awareness with Labeled Random Finite Sets. *IEEE Access*. 2019. Vol. 7. P. 36991-37003. DOI: 10.1109/ACCESS.2019. 2904545

10. **Li C.,** Okamura H., Dohi T. Parameter Estimation of M_t/M/1/K Queueing Systems with Utilization Data. *IEEE Access*. 2019. Vol. 7. P. 42664-42671. DOI: 10.1109/ACCESS.2019.2906796

11. **Голубничий А. Г.,** Конахович Г. Ф. Мультипликативно комплементарные бинарные сигнально-кодовые конструкции. *Известия высших учебных заведений. Радиоэлектроника*. 2018. Т. 61. № 10. С. 551-565. DOI: 10.20535/ S0021347018100011.

12. **Голубничий О. Г.** Синтез аналітичних форм опису автокореляційної функції узагальнених бінарних послідовностей Баркера типу 1 на основі її декомпозиції з використанням лінійних складових. *Наукоємні технології.* 2019. Т. 41. № 1. С. 10–15. DOI: 10.18372/2310-5461.41.13523.

13. **Голубничий О. Г.** Синтез систем корельованих сигналів з використанням доповненої процедури Грама-Шмідта. *Наукоємні технології.* 2018. Т. 40. № 4. С. 405-408. DOI: 10.18372/2310-5461.40.13265.

Голубничий О. Г.

АНАЛІЗ ОСОБЛИВОСТЕЙ РЕАЛІЗАЦІЇ ЕМ-АЛГОРИТМУ ПРИ КЛАСТЕРИЗАЦІЇ СИСТЕМ СИГНАЛЬНИХ КОНСТРУКЦІЙ

ЕМ-алгоритм (expectation-maximization algorithm) є відомим статистичним методом, який використовується в області обробки даних та сигналів для кластерного аналізу, оцінювання параметрів, а також в інших методах машинного навчання. ЕМ-алгоритм характеризується деякими специфічними особливостями, наприклад чутливістю до початкових параметрів. Метою статті є аналіз особливостей ЕМ-алгоритму, які виникають внаслідок виявлення порожніх кластерів при розв'язанні задачі кластеризації сигнальних конструкцій. Ці особливості проявляються в утворенні невизначеностей (математичних сингулярностей) у логарифмічній функції правдоподібності, максимізація якої виконується під час ітеративної процедури ЕМ-алгоритму. У статті наведено приклад можливої задачі кластеризації в області аналізу сигналів. У цьому прикладі використовується підхід, що базується на аналізі взаємних кореляцій між сигналами при представленні цих кореляцій гаусівською змішаною моделлю, а також на оцінюванні параметрів гаусівської змішаної моделі та прихованих змінних (ймовірностей приналежності елементів суміші до певних компонент цієї суміші, що є визначальним при прийнятті рішень про приналежність конкретного елемента до певного кластеру) за допомогою ЕМ-алгоритму. В результаті аналізу розглянутого типу математичної сингулярності показано, що можна видалити порожній кластер таким чином, що структура і значення логарифмічної функції правдоподібності коригуються до тих, які були б у випадку апріорної відсутності зазначеного порожнього кластера. Особливість, яка проаналізована у статті, більш характерна для модифікації ЕМ-алгоритму з видаленням компонент гаусівської змішаної моделі. Ця особливість може також виникати у випадках відносно невеликого числа елементів, які підлягають кластерному аналізу, або у випадку відносно великої кількості компонент (кластерів) гаусівської змішаної моделі, яка є структурним параметром ЕМ-алгоритму.

Ключові слова: кластерний аналіз; ЕМ-алгоритм; математична сингулярність; гаусівська змішана модель; машинне навчання; обробка даних.

Holubnychyi A. G.

ANALYSIS OF PECULIARITIES OF THE EM-ALGORITHM IN THE IMPLEMENTATION OF CLUSTERING OF SIGNAL CONSTRUCTIONS

The expectation-maximization (EM) algorithm is a well-known statistical method, which is used in the field of data and signal processing for cluster analysis, parameter estimation and other machine learning techniques. The EM-algorithm is characterized by some specific peculiarities, e.g., a sensitivity to initial parameters. The article aims to analyze peculiarities of the EM-algorithm, which arise due to the detection of empty clusters in solving the problem of clustering of signal constructions. These peculiarities boil down to the formation of uncertainties (mathematical singularities) in the log-likelihood function, the maximization of which is performed during the iterative procedure of the EM-algorithm. An example of a possible clustering problem in the field of signal analysis is shown. The example uses an approach based on an analysis of correlations between signals using the Gaussian mixture model for these correlations and an estimation of the Gaussian mixture model parameters and hidden variables (the probabilities for belonging of elements of the mixture to certain components of this mixture, which is decisive for a criteria whether a particular element belongs to a particular cluster) using the EM-algorithm. As a result of the analysis of considered type of mathematical singularity, it is shown that it is possible to remove an empty cluster in such a way that the structure and

values of the log-likelihood function are adjusted to those that would be in the case of the a priori absence of this specified empty cluster. The peculiarity, which is analyzed in the article, is more typical for the modification of the EM-algorithm with deletion the components of the Gaussian mixture model. This peculiarity may also arise in cases of relatively small number of elements that need to be analyzed through clustering or relatively large number of components of the Gaussian mixture model (clusters), which is the structural parameter of the EM-algorithm.

Keywords: cluster analysis; expectation-maximization algorithm; mathematical singularity; Gaussian mixture model; machine learning; data processing.

Голубничий А. Г.

АНАЛИЗ ОСОБЕННОСТЕЙ РЕАЛИЗАЦИИ ЕМ-АЛГОРИТМА ПРИ КЛАСТЕРИЗАЦИИ СИСТЕМ СИГНАЛЬНЫХ КОНСТРУКЦИЙ

ЕМ-алгоритм (expectation-maximization algorithm) является известным статистическим методом, который используется в области обработки данных и сигналов в целях кластерного анализа, оценки параметров, а также в других методах машинного обучения. ЕМ-алгоритм характеризуется некоторыми специфическими особенностями, например, чувствительностью к начальным параметрам. Целью статьи является анализ особенностей ЕМ-алгоритма, которые возникают в результате обнаружения пустых кластеров при решении задачи кластеризации сигнальных конструкций. Эти особенности проявляются в образовании неопределенностей (математических сингулярностей) в логарифмической функции правдоподобия, максимизация которой выполняется при итерационной процедуре ЕМ-алгоритма. В статье приведен пример возможной задачи кластеризации в области анализа сигналов. В этом примере используется подход, основанный на анализе взаимных корреляций между сигналами при представлении этих корреляций гауссовской смешанной моделью, а также на оценивании параметров гауссовской смешанной модели и скрытых переменных (вероятностей принадлежности элементов смеси определенным компонентам этой смеси, которые являются определяющими при принятии решений о принадлежности конкретного элемента к определенному кластеру) с помощью ЕМ-алгоритма. В результате анализа рассмотренного типа математической сингулярности показано, что является возможным удаление пустого кластера таким образом, что структура и значение логарифмической функции правдоподобия корректируются так, если априори отсутствовал бы указанный пустой кластер. Особенность, которая проанализирована в статье, более характерна для модификации ЕМ-алгоритма с удалением компонент гауссовской смешанной модели. Эта особенность может также возникать в случаях относительно небольшого числа элементов, подлежащих кластерному анализу, или в случае относительно большого количества компонент (кластеров) гауссовской смешанной модели, которое является структурным параметром ЕМ-алгоритма.

Ключевые слова: кластерный анализ; ЕМ-алгоритм; математическая сингулярность; гауссовская смешанная модель; машинное обучение; обработка данных.

Стаття надійшла до редакції 28.04.2019 р.
Прийнято до друку 31.05.2019 р.