ESTIMATING PARAMETERS AND THE MIXTURE COMPONENT NUMBER OF A GMM IN THE PRESENCE OF UNOSERVED DATA

ƯỚC LƯỢNG THAM SỐ VÀ SỐ THÀNH PHẦN CỦA MÔ HÌNH GMM TRONG TRƯỜNG HỢP MỘT SỐ MẪU DỮ LIỆU KHÔNG QUAN SÁT ĐƯỢC

Vu Trung Kien[1,*], Tran Quang Viet[1]

ABSTRACT

The Gaussian Mixture Model (GMM) is one of powerful approaches to model data that is heterogeneous and stems from multiple populations. However, in some certain situations, a part of dataset is unobservable owing to censoring problem. This problem refers to the fact that the value of a measurement or observation is only partially known. For example, the sensors on smart phones are not able to measure WiFi Received Signal Strength Indication (RSSI) values below a fixed threshold (-100dBm with typical smart phones). In that cases, RSSI values which are less than or equal to -100dBm will return the same value as -100dBm. In this paper, a novel method is proposed in order to estimate the number of components of the GMM and its parameters with the existence of censored data by applying the Expectation Maximization algorithm (EM) and the Sum of Weighted Real elements in Logarithm of Characteristic Functions (SWRLCF). The experimental results using artificial data show that this proposal outperform the current approaches when collected data was suffered from censoring.

Keywords: *GMM, EM, SWRLCF, Censored Data.*

TÓM TẮT

Mô hình hỗn hợp Gauss là một công cụ được sử dụng một cách hiệu quả để mô tả phân bố của các tập dữ liệu không đồng nhất và thu thập từ nhiều đối tượng/điều kiện khác nhau. Tuy nhiên trong một số tình huống thực tế, một phần của tập dữ liệu có thể không quan sát được do bị “cắt”. Ví dụ, cảm biến trên các điện thoại thông minh không thể đo được chỉ số cường độ của tín hiệu phát ra từ một trạm thu/phát WiFi nếu chúng nhỏ hơn ngưỡng thu, ví dụ -100dBm. Khi đó tất cả các phép đo có giá trị nhỏ hơn hoặc bằng -100dBm sẽ được trả về với cùng một giá trị là -100dBm. Bài báo này đề xuất các thuật toán ước lượng các tham số của hình hỗn hợp Gauss và số thành phần Gauss dựa trên thuật toán cực đại hóa kỳ vọng và tổng phần thực của hàm đặc trưng. Các kết quả thực nghiệm với tập dữ liệu mô phỏng chứng minh hiệu quả của các thuật toán được đề xuất so với các công trình đã được công bố khi một phần của tập dữ liệu bị “cắt”.

Từ khóa: *Mô hình hỗn hợp Gauss (GMM), thuật toán cực đại hóa kỳ vọng (EM), tổng phần thực của hàm đặc trưng (SWRLCF), một số mẫu dữ liệu bị “cắt” và không quan sát được.*

1. INTRODUCTION

The GMM has been widely applied in the fields of signal processing. It is a model to represent normal distributed subsets within an overall dataset. The GMM does not require to know which sub-populations the data point belongs to, allowing the model to automatically find out which sub-populations. Since the demographic division is not known, this constitutes a form of unsupervised learning. For example, two Gaussian distributions with different means and variances are used to model two data sets which are RSSI values collected from two WiFi Access Points (AP) [1]. If we don't care which AP the data was gathered from, the distribution of all RSSIs must be the sum of two Gaussian components with different mixing weights (Figure 1). The model making this assumption is the GMM. In general, a GMM may have two or more than two components. The estimations of the individual normal distribution components’ parameters and the number of mixtured components are canonical problems in modeling data with GMMs.

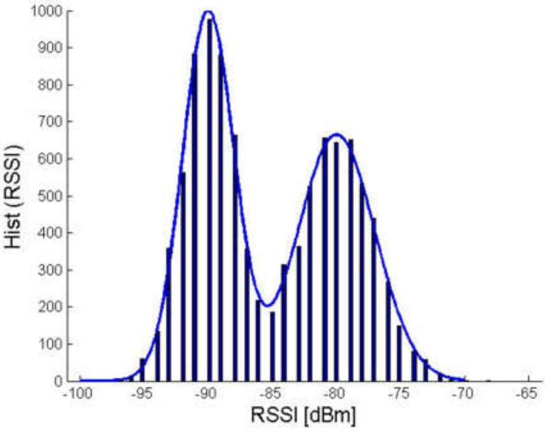

Figure 1. Complete data

In [2, 3], authors used GMMs to model WiFi RSSI data and applied the EM algorithm [4, 5] to estimate parameters but the censoring problem was not considered. Censoring (or clipping) means that the sensors on received devices are not able to measure RSSI values below a certain

[1] Faculty of Electronic Engineering, Hanoi University of Industry

[*] Email: kien.vu@haui.edu.vn

Received: 25/01/2021
Revised: 23/6/2021
Accepted: 15/11/2021

limitation, for example -100dBm (Figure 2). It occurs owing to the limited sensitivity of WiFi sensors on portable devices [6]. In [6, 7], an upgraded version of EM algorithm was proposed to deal with the censoring problem. The results showed that parameters were estimated more accurately when data suffered from censoring. However, data set was model by single Gaussian distributions but not GMMs.

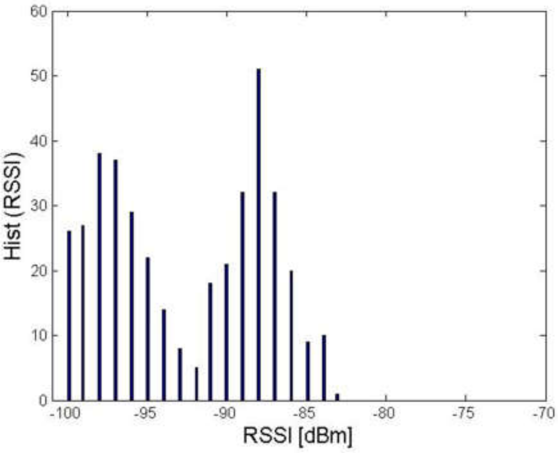

Figure 2. Censored data

The EM algorithm is one of popular methods to estimate parameters of GMMs but it has a drawback is that the mixture component number of GMM is not known. Therefore, it is required to develop a feasible method to estimate the mixture component number of GMM instead of assigning it to a fixed number [8]. The AIC [2] and BIC [9] were used to determine the best mixture component number for the GMM. These methods can reduce computational costs and produce relatively accuracy. A model selection method, basing on the SWRLCF, was proposed in [10]. This proposal is feasible with applications that have large amounts of data. Three approaches mentioned above can estimate the mixture component number of GMMs effectively when data are complete but the censoring problem was not noticed.

Taking the all problems mentioned above in to consideration, the target of this research is to estimate parameters and Gaussian component number of a collected data set following a mixture distribution and suffering from censoring. In the following, the algorithms for parameter estimation and the model selection are developed (section 2). The effectiveness of proposed methods is evaluated in section 3. The conclusion of this paper is mentioned in section 4.

2. METHODS

2.1. Parameter estimation using the EM algorithm

Definitions:

$\mathbf{y} = [y_1, y_2, ..., y_N]$ is the set of collected data (complete, non-censored data), $y_n \in \mathbb{R}$ are independent identically distributed random variables (n = 1 ÷ N), N is the total number of samples in dataset; c is the censored threshold;

$\mathbf{x} = [x_1, x_2, ..., x_N]$ is the set of censored data, $x_n = \begin{cases} y_n \text{ if } y_n > c \\ c \text{ if } y_n \le c \end{cases}$;

$\mathbf{\Theta} = [w_1, ..., w_J; \mu_1, ..., \mu_J; \sigma_1, ..., \sigma_J]$ is the set of parameters of a GMM; J is the Gaussian component number; $\theta_j = [\mu_j; \sigma_j]$ is the parameter of j^th Gaussian component; w_j are positive mixing weights which sum up to one (j = 1 ÷ J); p(...) is the probability density function.

The likelihood of **y** is

$$\ell(\mathbf{y};\mathbf{\Theta}) = \prod_{n=1}^{N}\sum_{j=1}^{J} w_j\, p(y_n;\theta_j). \tag{1}$$

Let $\mathbf{A} = \begin{bmatrix} A_{11} & \cdots & A_{1J} \\ \vdots & \ddots & \vdots \\ A_{N1} & \cdots & A_{NJ} \end{bmatrix}$ is a set of latent variables, $A_{nj} = 1$ if y_n belongs to the j^th Gaussian component; otherwise, $A_{nj} = 0$. The equation (1) becomes:

$$\ell(\mathbf{y};\mathbf{\Theta},\mathbf{A}) = \prod_{n=1}^{N}\prod_{j=1}^{J}\left[w_j\, p(y_n;\theta_j)\right]^{A_{nj}}. \tag{2}$$

The log-likelihood is as follows:

$$\ln\left[\ell(\mathbf{y};\mathbf{\Theta},\mathbf{A})\right] = \sum_{n=1}^{N}\sum_{j=1}^{J} A_{nj}\left\{\ln(w_j) + \ln\left[p(y_n;\theta_j)\right]\right\}. \tag{3}$$

E-step: Calculate the conditional expectation of $\ln\left[\ell(\mathbf{y};\mathbf{\Theta},\mathbf{A})\right]$ given by **x** and parameters at k^th iteration $(\mathbf{\Theta}^{(k)})$:

$$\begin{aligned} &E\left\{\ln\left[\ell(\mathbf{y};\mathbf{\Theta},\mathbf{A})\right] \middle| \mathbf{x};\mathbf{\Theta}^{(k)}\right\} \\ &= \sum_{n=1}^{N}\sum_{j=1}^{J}\int_{-\infty}^{+\infty} A_{nj} \ln\left[w_j\, p(y_n;\theta_j)\right] p\left(y_n, A_{nj} \mid x_n;\Theta_j^{(k)}\right) dy_n \triangleq F\left(\mathbf{\Theta};\mathbf{\Theta}^{(k)}\right) \end{aligned} \tag{4}$$

≜For the case $A_{nj} = 0$, $F\left(\mathbf{\Theta};\mathbf{\Theta}^{(k)}\right) = 0$; when $A_{nj} = 1$, the equation (4) becomes:

$$\begin{aligned} F\left(\mathbf{\Theta};\mathbf{\Theta}^{(k)}\right) &= \sum_{n=1}^{N}\sum_{j=1}^{J}\int_{-\infty}^{+\infty} \ln\left[w_j\, p(y_n;\theta_j)\right] p\left(y_n, A_{nj} = 1 \mid x_n;\Theta_j^{(k)}\right) dy_n \\ &= \sum_{n=1}^{N}\sum_{j=1}^{J} (1 - z_n)\, \Upsilon\left(x_n;\Theta_j^{(k)}\right)\left[\ln(w_j) + \ln\left(\mathcal{N}(x_n;\theta_j)\right)\right] \\ &+ \sum_{n=1}^{N}\sum_{j=1}^{J} z_n \beta\left(\Theta_j^{(k)}\right)\left\{\ln(w_j) + \int_{-\infty}^{c} \ln\left[\mathcal{N}(y_n;\theta_j)\right] \frac{\mathcal{N}\left(y_n;\theta_j^{(k)}\right)}{I_0\left(\theta_j^{(k)}\right)} dy_n\right\}. \end{aligned} \tag{5}$$

In the equation(5), z_n(n = 1 ÷ N) are binary variables that indicate observable samples (z_n = 0) and unobservable samples (z_n = 1); $\mathcal{N}\left(...;\theta_j^{(k)}\right)$ is the Gaussian distribution parameterized by $\theta_j^{(k)}$; Functions $\Upsilon\left(x_n;\Theta_j^{(k)}\right)$, $\beta\left(\Theta_j^{(k)}\right)$ and $I_0\left(\theta_j^{(k)}\right)$ are as follows:

$$\Upsilon\left(x_n;\Theta_j^{(k)}\right)=\frac{w_j^{(k)}\mathcal{N}\left(x_n;\theta_j^{(k)}\right)}{\sum_{j=1}^{J}w_j^{(k)}\mathcal{N}\left(x_n;\theta_j^{(k)}\right)}; \tag{6}$$

$$\beta\left(\Theta_j^{(k)}\right)=\frac{w_j^{(k)}\,\mathrm{I}_0\left(\theta_j^{(k)}\right)}{\sum_{j=1}^{J}w_j^{(k)}\,\mathrm{I}_0\left(\theta_j^{(k)}\right)}; \tag{7}$$

$$\mathrm{I}_0\left(\theta_j^{(k)}\right)=\frac{1}{2}\mathrm{erfc}\left(-\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right). \tag{8}$$

M-step:

Calculating partial derivatives of $F\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)}\right)$ in terms of μ_j,σ_j,w_j then assigning zero we have re-estimated parameters at (k+1)th iteration:

$$\mu_j^{(k+1)}=\frac{\sum_{n=1}^{N}(1-z_n)\Upsilon\left(x_n;\Theta_j^{(k)}\right)x_n+\beta\left(\Theta_j^{(k)}\right)\frac{\mathrm{I}_1\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}\sum_{n=1}^{N}z_n}{\sum_{n=1}^{N}(1-z_n)\Upsilon\left(x_n;\Theta_j^{(k)}\right)+\beta\left(\Theta_j^{(k)}\right)\frac{\mathrm{I}_1\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}\sum_{n=1}^{N}z_n}; \tag{9}$$

$$\hat{\sigma}_j^{(k+1)}=\frac{\sum_{n=1}^{N}(1-z_n)\Upsilon\left(x_n;\Theta_j^{(k)}\right)\left(x_n-\mu_j^{(k)}\right)^2+\beta\left(\Theta_j^{(k)}\right)\left[\frac{\mathrm{I}_2\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}-\frac{2\mu_j^{(k)}\,\mathrm{I}_1\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}+\left(\mu_j^{(k)}\right)^2\right]\sum_{n=1}^{N}z_n}{\sum_{n=1}^{N}(1-z_n)\Upsilon\left(x_n;\Theta_j^{(k)}\right)+\beta\left(\Theta_j^{(k)}\right)\sum_{n=1}^{N}z_n}; \tag{10}$$

$$w_j^{(k+1)}=\frac{\sum_{n=1}^{N}(1-z_n)\Upsilon\left(x_n;\Theta_j^{(k)}\right)+\beta\left(\Theta_j^{(k)}\right)\sum_{n=1}^{N}z_n}{N}. \tag{11}$$

The notations $\mathrm{I}_1\left(\theta_j^{(k)}\right)$ and $\mathrm{I}_2\left(\theta_j^{(k)}\right)$ are given in the equations (12), (13):

$$\mathrm{I}_1\left(\theta_j^{(k)}\right)=\mu_j^{(k)}\mathrm{I}_0\left(\theta_j^{(k)}\right)-\frac{1}{\sqrt{2\pi}}\sigma_j^{(k)}\exp\left[-\left(\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)^2\right]; \tag{12}$$

$$\mathrm{I}_2\left(\theta_j^{(k)}\right)=\left[\left(\mu_j^{(k)}\right)^2+\left(\sigma_j^{(k)}\right)^2\right]\mathrm{I}_0\left(\theta_j^{(k)}\right)-\frac{1}{\sqrt{2\pi}}\sigma_j^{(k)}\left(c+\mu_j^{(k)}\right)\exp\left[-\left(\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)^2\right]. \tag{13}$$

2.2. Estimating the number of components of GMM

In this sub-section, the SWRLCF is proposed to estimate the Gaussian component number of GMM in the presence of censored data.

Let $\hat{w}_j$ and $\hat{\sigma}_j$ be the mixing weight and the standard deviation of jth Gaussian component of GMM (j = 1 ÷ J) obtained by applying the EM algorithm mentioned in previous sub-section. According to calculations in [10], the SWRLCF of a GMM with J Gaussian components is as follow:

$$\mathrm{SWRLCF}(J)=\sum_{j=1}^{J}\hat{w}_j\hat{\sigma}_j^2 \tag{14}$$

Figure 3 shows the proposed algorithm for estimating the Gaussian component number of GMM using the upgraded EM algorithm developed in sub-section 2.1 and the SWRLCF given in equation (14).

In the figure 3, a set of incomplete data (**x**) is inputted; ε is the convergence threshold of the EM algorithm; J_{max} is the number of Gaussian components for calculating SWRLCF(J); τ is the convergence threshold of the model selection algorithm. At Jth iteration, the algorithm outputs the estimated Gaussian component number $(\hat{J})$ and estimated parameters $(\hat{\boldsymbol{\Theta}}_j)$ using to model distribution of **x**.

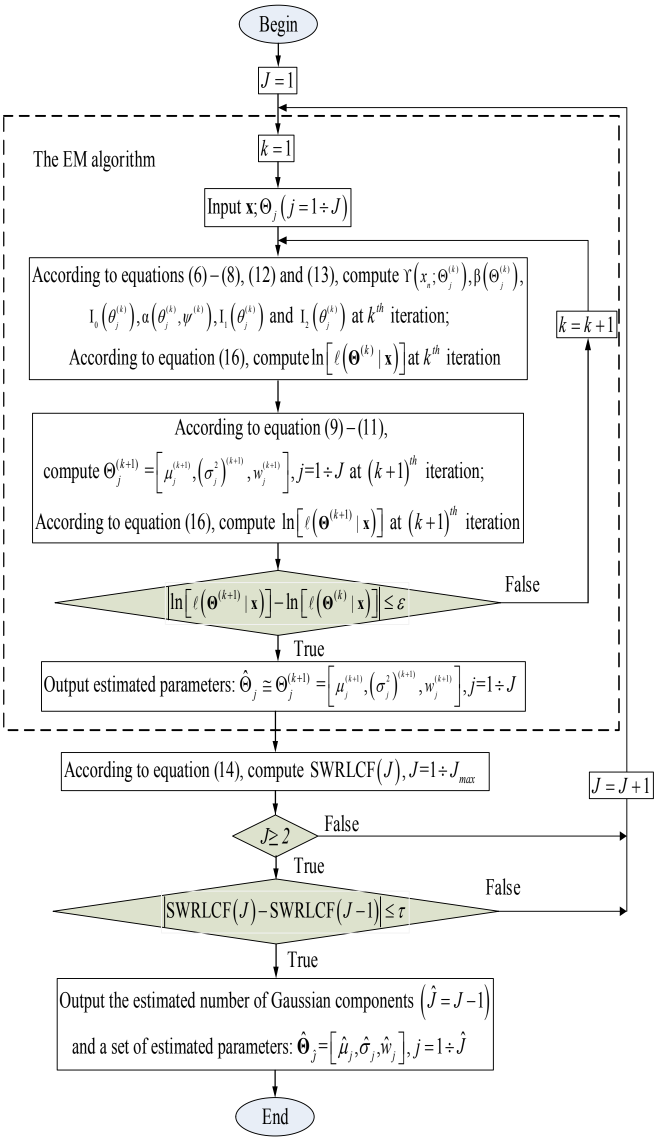
Begin
$J=1$
$k=1$
The EM algorithm
Input $\mathbf{x};\Theta_j\left(j=1\div J\right)$
According to equations (6) – (8), (12) and (13), compute $\Upsilon\left(x_n;\Theta_j^{(k)}\right),\beta\left(\Theta_j^{(k)}\right),$ $\mathrm{I}_0\left(\theta_j^{(k)}\right),\alpha\left(\theta_j^{(k)},\psi^{(k)}\right),\mathrm{I}_1\left(\theta_j^{(k)}\right)$ and $\mathrm{I}_2\left(\theta_j^{(k)}\right)$ at k^{th} iteration;
According to equation (16), compute $\ln\left[\ell\left(\boldsymbol{\Theta}^{(k)}\mid\mathbf{x}\right)\right]$ at k^{th} iteration
According to equation (9) – (11),
compute $\Theta_j^{(k+1)}=\left[\mu_j^{(k+1)},\left(\sigma_j^2\right)^{(k+1)},w_j^{(k+1)}\right], j=1\div J$ at $(k+1)^{th}$ iteration;
According to equation (16), compute $\ln\left[\ell\left(\boldsymbol{\Theta}^{(k+1)}\mid\mathbf{x}\right)\right]$ at $(k+1)^{th}$ iteration
$\left|\ln\left[\ell\left(\boldsymbol{\Theta}^{(k+1)}\mid\mathbf{x}\right)\right]-\ln\left[\ell\left(\boldsymbol{\Theta}^{(k)}\mid\mathbf{x}\right)\right]\right|\le\varepsilon$
False
$k=k+1$
True
Output estimated parameters: $\hat{\Theta}_j\cong\Theta_j^{(k+1)}=\left[\mu_j^{(k+1)},\left(\sigma_j^2\right)^{(k+1)},w_j^{(k+1)}\right], j=1\div J$
According to equation (14), compute $\mathrm{SWRLCF}(J), J=1\div J_{max}$
$J=J+1$
$J\geq 2$
False
True
$\left|\mathrm{SWRLCF}(J)-\mathrm{SWRLCF}(J-1)\right|\le\tau$
False
True
Output the estimated number of Gaussian components $\left(\hat{J}=J-1\right)$
and a set of estimated parameters: $\hat{\boldsymbol{\Theta}}_j=\left[\hat{\mu}_j,\hat{\sigma}_j,\hat{w}_j\right], j=1\div\hat{J}$
End

Figure 3. The proposed algorithm for estimating the Gaussian component number

3. RESULTS AND DISCUSSION

3.1. Parameter estimation

With the aim of evaluating the effectiveness of different EM algorithms proposed by authors and our proposal mentioned in sub-section 2.1, we generate 1000 samples of complete mixture data (**y**) basing on characteristics of gathered WiFi RSSI data [1, 11] with a set of true parameters given in Table 1. Observable (censored) data (**x**) was collected by applying function:

$$x_n = \begin{cases} y_n \text{ if } y_n > c \\ c \text{ if } y_n \le c \end{cases}. \qquad (15)$$

Table 1. Parameters used to generate artificial data

Parameter	w_1	σ_1	μ_1	w_2	σ_2	μ_2
Value	0.5	3	-80dBm	0.5	4	-90dBm

The EM algorithm convergence threshold was set to 10^{-6} ($\varepsilon = 10^{-6}$). Table 2 shows the mean and the standard deviation of Kullback Leibler (KL) divergence [12] between true parameters and estimated parameters proposed by authors. The EM-GMM is the EM algorithm for GMM introduced in [2, 3]. The EM-CD-G is the upgraded EM algorithm for estimating single Gaussian data suffering from censoring and dropping proposed in [6].

Table 2. Parameter estimation compared by mean and standard deviation of KL

c(dBm)	**-84**	**-87**	**-90**	**-93**	**-96**
Mean of KL					
EM-GMM	7.2847	5.6358	3.1491	0.0329	0.0018
EM-CD-G	0.0972	0.0886	0.0798	0.0679	0.0664
Proposed EM algorithm	0.0481	0.0126	0.0098	0.0034	0.0016
Standard deviation of KL					
EM-GMM	0.1984	0.1025	0.0351	0.0323	0.0151
EM-CD-G	0.1851	0.1325	0.1199	0.0175	0.0172
Proposed EM algorithm	0.0624	0.0451	0.0227	0.0176	0.0139

As can be seen in table 2, when the censored threshold (c) is -96dBM, the data are almost observable, the three methods produced the same results. However, once the censoring problem occurred, our method showed the best results among the considered algorithms. This can be clarified as follows: In the upgraded EM algorithm mentioned in sub-section 2.1, both observed data ($x_n = y_n$) and unobserved data (x_n = c) are contributed to the estimates.

3.2. Model selection

In this sub-section, the proposed method and other state-of-art approaches [2, 9, 10] are evaluated through different experiments on artificial data. The process of generating data is as follows:

- Generate complete data (**y**): 4 sets of mixture data with 1, 2, 3, and 4 Gaussian components, respectively (J = 1, 2, 3 and 4) were generated by using 4 set of parameters given in table 3. The number of samples in a data set is 250.

- Incomplete data (**x**) was gathered by using function in equation (15); the censored threshold (c) was changed to -90dBm, -92dBm and -94dBm (table 3).

The maximum Gaussian component number for calculating penalty functions and SWRLCF was set to 8 (J_{max} = 8). The convergence threshold of the model selection algorithm was set to 0.02 (τ = 0.02). After 1000 experiments, different levels between the true and estimated number of Gaussian components (J and $\hat{J}$) outputted by four approaches were recorded in Table 3.

Table 3. Model selection outputted by four algorithms.

Methods	Probability	Results		
		c =-94dBm	c =-92dBm	c =-90dBm
Using EM for GMM and AIC [2]	$J = \hat{J}$	0.28	0.01	0.01
	$\lvert J-\hat{J}\rvert = 1$	0.21	0.31	0.3
	$\lvert J-\hat{J}\rvert \ge 2$	0.51	0.68	0.69
Using EM for GMM and BIC [9]	$J = \hat{J}$	0.82	0.01	0.01
	$\lvert J-\hat{J}\rvert = 1$	0.15	0.39	0.38
	$\lvert J-\hat{J}\rvert \ge 2$	0.03	0.6	0.61
Using EM for GMM and SWRLCF [10]	$J = \hat{J}$	0.53	0.52	0.02
	$\lvert J-\hat{J}\rvert = 1$	0.27	0.39	0.75
	$\lvert J-\hat{J}\rvert \ge 2$	0.2	0.09	0.23
Proposed method	$J = \hat{J}$	0.85	0.81	0.76
	$\lvert J-\hat{J}\rvert = 1$	0.14	0.16	0.21
	$\lvert J-\hat{J}\rvert \ge 2$	0.01	0.03	0.03

Results in table 3 shows that our proposed approach introduced quite better results than other methods, particularly when almost data were censored (c = -92; -94dBm). This can be clarified as follows: Proposed method applied not only the upgraded EM algorithm but also extended SWRLCF, in which both observed data and unobserved data are contributed to the estimates (see in equations (9)-(11),(14)). On the other hand, in the penalty functions of AIC[2], BIC[9] and SWRLCF [10], there is almost no practical contribution of unobserved data while they really contributed to the SWRLCF in our approach, see in equation (10), (11), (14).

4. CONCLUSION

In this paper, the authors introduced the novel methods to take the censoring problem presented in collected data into consideration. Simulation results using generated data showed that our upgraded EM algorithm allows to estimate the parameters of GMMs more accurately than existing methods, especially when collected data were suffered from censoring problem. By applying our proposal, errors of estimated parameters have been reduced. Moreover, by utilizing the extended SWRLCF, both the number of components and parameters of GMMs are estimated accurately. This leads the fact that the performance in modelling the distribution of data set is better. In future works, the proposed EM and model selection algorithm will be applied to WiFi RSSI based indoor positioning systems so as to recduce the positioning errors and the calculating time.

REFERENCES

[1]. Luo Jiayou, Zhan Xingqun, 2014. *Characterization of Smart Phone Received Signal Strength Indication for WLAN Indoor Positioning Accuracy Improvement.* Journal of Networks, vol. 9(3), pp. 439-746. DOI: 10.4304/jnw.9.3.739-746

[2]. C. H. Tseng, J. Yen, 2016. *Enhanced Gaussian Mixture Model for Indoor Positioning Accuracy.* in International Computer Symposium (ICS), pp. 462-466.DOI: 10.1109/ICS.2016.0099

[3]. M. Alfakih, M. Keche, H. Benoudnine, 2015. *Gaussian mixture modeling for indoor positioning WIFI systems*. in 3rd International Conference on Control, Engineering & Information Technology (CEIT), pp. 1-5. DOI: 10.1109/CEIT.2015.7233072

[4]. A. Dempster, N. Laird, D. B. Rubin, 1977. *Maximum Likelihood From Incomplete Data Via The EM algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, pp. 1-38.

[5]. G. Lee, C. Scott, 2012. *EM algorithms for multivariate Gaussian mixture models with truncated and censored data.* Computational Statistics & Data Analysis, vol. 56, pp. 2816-2829, DOI: https://doi.org/10.1016/j.csda.2012.03.003

[6]. M. K. Hoang, R. Haeb-Umbach, 2013. *Parameter estimation and classification of censored Gaussian data with application to WiFi indoor positioning*. in IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3721-3725.DOI: 10.1109/ICASSP.2013.6638353

[7]. M. K. Hoang, J. Schmalenstroeer, R. Haeb-Umbach, 2015. *Aligning training models with smartphone properties in WiFi fingerprinting based indoor localization.* in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1981-1985.DOI: 10.1109/ICASSP.2015.7178317

[8]. G. Celeux, C. Biernacki, 1998. *Choosing Models in Model-Based Clustering and Discriminant Analysis.* Journal of Statistical Computation and Simulation, vol. 64, pp. 1-22, DOI: 10.1080/00949659908811966

[9]. T. Huang, H. Peng, K. Zhang, 2013. *Model Selection for Gaussian Mixture Models.* Statistica Sinica, vol. 27, pp. 147-169. DOI: 10.5705/ss.2014.105

[10]. C. Xie, J. Chang, Y. Liu, 2013. *Estimating the number of components in Gaussian mixture models adaptively*. Journal of Information & Computational Science, vol. 124, pp. 6216-6221. DOI: 10.1016/j.ijleo.2013.05.028

[11]. K. Kaemarungsi, 2006. *Distribution of WLAN received signal strength indication for indoor location determination.* in 2006 1st International Symposium on Wireless Pervasive Computing, pp. 6-16.DOI: 10.1109/ISWPC.2006.1613601

[12]. J. R. Hershey, P. A. Olsen, 2007. *Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models.* in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, pp. IV-317-IV-320.DOI: 10.1109/ICASSP.2007.366913

THÔNG TIN TÁC GIẢ

Vũ Trung Kiên, Trần Quang Việt

Khoa Điện tử, Trường Đại học Công nghiệp Hà Nội