MIXTURE MODEL CLUSTERING OF BINNED UNCERTAIN DATA

Hani HAMDAN

SUPELEC
Dept. of Signal Processing and Electronic Systems
3 rue Joliot-Curie, 91190 Gif-sur-Yvette, France
e-mail: Hani.Hamdan@supelec.fr

Abstract: This paper addresses the problem of taking into account data imprecision in the mixture model clustering of binned data. Binning (or grouping) data is common in data analysis and machine learning. Recently, we developed an original method which fitted the binning data procedure to imprecise data. The idea was to model imprecise data by multivariate uncertainty zones and to assign each uncertainty zone to several bins with proportions proportional to its overlapping volumes with the bins. The experimental results of this method when it was associated with the *binned-EM* algorithm (mixture approach) were encouraging. However, the *binned-EM* algorithm has the disadvantage of being sometimes computationally expensive. To overcome this problem, we propose in this paper to apply our binning data procedure with the classification approach based on *bin-EM-CEM* algorithm which is much faster than the *binned-EM* algorithm. The paper concludes with a brief description of a flaw diagnosis application using acoustic emission. The experimental results compare our binning data procedure with the classical one (when applied to imprecise data) in the classification approach framework, and with the *int-EM-CEM* algorithm, in the context of binned bivariate measurements of acoustic emission event localization.

Keywords: Imprecise data, uncertain data, binned data, binned uncertain data, mixture model, *bin-EM-CEM* algorithm, *int-EM-CEM* algorithm, clustering, semi fuzzy clustering.

1. INTRODUCTION

In this paper, we address the problem of taking into account data imprecision in the mixture model clustering of binned data. Binning data is common in data analysis and machine learning. Binned data are data collected or transformed into frequencies located in disjoined areas of $\mathbb{R}^p$ called bins. In other words, binned data correspond exactly to a multidimensional histogram and data in this form are also called grouped data. Such data occur systematically in a variety of application when a measurement instrument has finite resolution but it may also occur intentionally when real-valued variables are quantized to simplify data collection or to reduce data because of their large size.

Data in the form of histogram also play an important role in a variety of pattern recognition and machine learning problems. For example, in computer vision, color histograms are used for object recognition (Swain and Ballard, 1991). In image retrieval, many studies rely heavily on the use of color and feature histograms (see Flickner et al., 1995; Maybury, 1997, for instance). A number of techniques in approximate querying of databases and in data mining of massive data sets also use histogram representations (see Poosala, 1997; Matias et al., 1998; Lee et al., 1999, for instance).

More particularly, within the framework of a flaw diagnosis problem by real time acoustic emission control, we had to classify, under real time constraints, a set of points located in the plane (see Figure 1). This plane represents a tank (pressure equipment) and each point of the plane represents the localization of an acoustic emission event. The clustering *CEM* algorithm (Celeux and Govaert, 1992) applied using a diagonal Gaussian mixture model (Celeux and Govaert, 1995), provides a satisfactory solution if the positions of the acoustic emission events are quite precise but cannot react in real time, when the size of data becomes very large (*e.g.* more than 10000 points). As data sets become larger, data processing becomes increasingly complex and as a result, data analysis requires more computation time. To take into account these real time constraints, we propose to reduce and group data (see Figure 2) before their treatment. However, sometimes, the acoustic emission events can not be quite localized and their positions are then imprecise. Although the use of binned data in the form of histogram constitutes a natural way of taking into account the localization imprecision, Hamdan and Govaert (2004b) proposed an original method which fitted the binning data procedure to imprecise data. The idea was to model imprecise data by uncertainty zones, *i.e.*, to define uncertainty zones around the imprecise points provided by the acquisition system (the dimensions of each uncertainty zone are chosen according to the importance of the localization uncertainty which is also provided by the acquisition system), and then to assign each uncertainty zone to several bins with proportions proportional to its overlapping volumes (or surfaces in $\mathbb{R}^2$) with the bins. The preliminary results

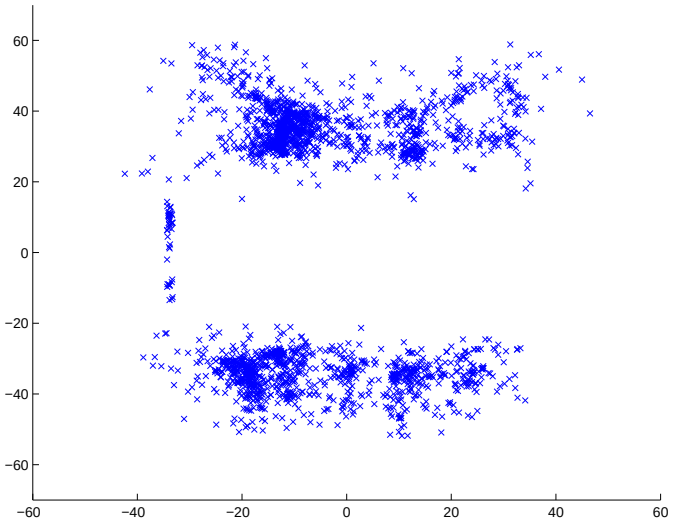

Fig. 1. Real data resulting from acoustic emission on the unfolded surface of a cylindrical pressure equipment.

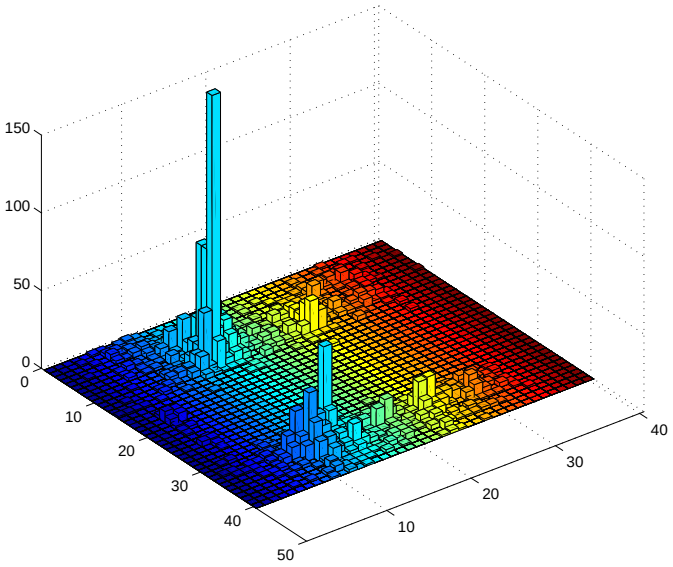

Fig. 2. Binned data corresponding to acoustic emission data of Figure 1.

of this method when it was associated with the *binned-EM* algorithm (Cadez et al., 2002) (mixture approach) were encouraging. However, the *binned-EM* algorithm has the disadvantage of being computationally expensive. In such cases and in this paper, we propose to apply our binning data procedure with the classification approach based on *bin-EM-CEM* algorithm (Samé, 2004) which is much faster.

To illustrate our concept of uncertainty zone data (or merely uncertain data), Figure 3 displays the uncertainty zones corresponding to acoustic emission data of Figure 1.

This paper is structured as follows. Section 2 presents the mixture model in two different contexts: that of binned data and that of uncertain data. Section 3 explains the classification approach in binned data mixture model clustering. In this section, the *bin-EM-CEM* algorithm (Samé, 2004; Samé et al., 2006) is described and summarized. In Section 4, we present the principle of fitting the binning

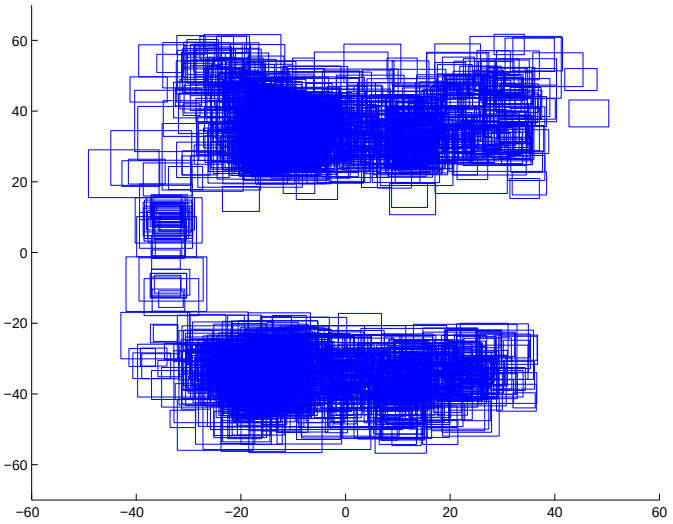

Fig. 3. Uncertain data corresponding to acoustic emission data of Figure 1.

data procedure to imprecise data, and we define the binned uncertain data concept. Section 5 describes briefly an application of this approach to flaw diagnosis, on pressure equipments, using acoustic emission. The binned uncertain data approach, associated with the *bin-EM-CEM* algorithm, is then applied in the context of imprecise bivariate measurements of acoustic emission event localization. Comparisons to the classical binning data procedure (when we group the original 'raw' measurements), and to the *int-EM-CEM* algorithm (Hamdan and Govaert, 2004c,a), are also done.

2. THE MODEL

In this work, we suppose that $((\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n))$ is a sample resulting from the couple of random vectors $(\boldsymbol{X}, \boldsymbol{Z})$ associated to a distributions mixture defined on $\mathbb{R}^p$ by:

$$f(\boldsymbol{x}; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k) \qquad (1)$$

where $\boldsymbol{\Phi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and $\pi_1, \ldots, \pi_K$ are the proportions of the mixture and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ the parameters of each component density ; z_i $(1 \leqslant i \leqslant n)$ indicates the origin component of $\boldsymbol{x}_i$ $(1 \leqslant i \leqslant n)$ and we note $\mathbf{z} = (z_1, \ldots, z_n)$. We consider that every point $\boldsymbol{x}_i$ belongs to an uncertainty zone $\boldsymbol{\mathcal{R}}_i$ of $\mathbb{R}^p$ and we suppose that the only available knowledge of the sample $((\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n))$ is the set of uncertainty zones $\boldsymbol{\mathcal{R}}_i$ $(1 \leqslant i \leqslant n)$ as $\boldsymbol{x}_i \in \boldsymbol{\mathcal{R}}_i$ $(1 \leqslant i \leqslant n)$. We note $\mathbf{R} = (\boldsymbol{\mathcal{R}}_1, \ldots, \boldsymbol{\mathcal{R}}_n)$ the vector of these zones. On the other hand, we consider a partition $(\boldsymbol{\mathcal{H}}_1, \ldots, \boldsymbol{\mathcal{H}}_v)$ of the space $\mathbb{R}^p$ in v bins and *in the binned data framework*, we suppose that the only available knowledge of the sample $((\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n))$ is the set of the frequencies n_r of $\boldsymbol{x}_i$ belonging to $\boldsymbol{\mathcal{H}}_r$. We note $\mathbf{a} = (n_1, \ldots, n_v)$ the vector of these frequencies $(\sum_{r=1}^{v} n_r = n)$. Since there is no available information about the exact positions of the points $\boldsymbol{x}_i$ into the bins $\boldsymbol{\mathcal{H}}_1, \ldots, \boldsymbol{\mathcal{H}}_v$, we assume that in a same bin, observations have the same label, *i.e.*, all the observations of a given bin belong to the same cluster. Thus, the labels z_i $(1 \leqslant i \leqslant n)$ can be replaced by the labels z_r $(1 \leqslant r \leqslant v)$ of the bins if the bin membership of

each $\boldsymbol{x}_i$ $(1 \leqslant i \leqslant n)$ is known, and z_r $(1 \leqslant r \leqslant v)$ is coded in the form $z_r = (z_{r1}, \ldots, z_{rK})$ where z_{rk} is equal to 1 if $\boldsymbol{\mathcal{H}}_r$ is resulting from the component k and is equal to 0 elsewhere.

3. THE CLASSIFICATION APPROACH IN BINNED DATA MIXTURE MODEL CLUSTERING

In this paragraph, we describe the classification approach (Samé, 2004) for simultaneous estimation of the mixture parameter $\boldsymbol{\Phi}$ and the partition $(z_1, \ldots, z_v)$ maximizing the log-likelihood criterion:

$$\begin{aligned} L(\boldsymbol{\Phi}; \mathbf{a}, \mathbf{z}) &= L(\boldsymbol{\Phi}; n_1, \ldots, n_v, z_1, \ldots, z_v) \\ &= \log P(\mathbf{a}, \mathbf{z}; \boldsymbol{\Phi}) \\ &= \sum_{r=1}^{v} \log \left(\pi_{z_r} \int_{\boldsymbol{\mathcal{H}}_r} f_{z_r}(\boldsymbol{x}; \boldsymbol{\theta}_{z_r}) d\boldsymbol{x} \right) \\ &\quad + \log \frac{n!}{\prod_{r=1}^{v} n_r!}. \end{aligned} \tag{2}$$

The maximization of $L(\boldsymbol{\Phi}; \mathbf{a}, \mathbf{z})$ with respect to $(\boldsymbol{\Phi}, \mathbf{z})$, knowing only the observed data $\mathbf{a} = (n_1, \ldots, n_v)$, can be performed on the basis of an initialization $\boldsymbol{\Phi}^{(0)}$ and then alternating, until convergence, the two following steps:

(1) computation of

$$\mathbf{z}^{(q)} = \underset{\mathbf{z}}{argmax}\ L(\boldsymbol{\Phi}^{(q)}; \mathbf{a}, \mathbf{z}).$$

(2) computation of

$$\boldsymbol{\Phi}^{(q+1)} = \underset{\boldsymbol{\Phi}}{argmax}\ L(\boldsymbol{\Phi}; \mathbf{a}, \mathbf{z}^{(q)}).$$

The development of these two steps leads to the *bin-EM-CEM* algorithm (Samé, 2004; Samé et al., 2006) which we summarized in Algorithm 1, in the case of diagonal Gaussian mixture model and while leaving from the initial parameter $\boldsymbol{\Phi}^{(0)}$, in three steps: E *(Expectation)*, C *(Classification)* and M *(Maximization)*. The maximization of $L(\boldsymbol{\Phi}; \mathbf{a}, \mathbf{z}^{(q)})$ with respect to $\boldsymbol{\Phi}$ in the M-step, is performed by an *internal binned-EM* algorithm (complete data are $((\boldsymbol{x}_1, z_1^{(q)}), \ldots, (\boldsymbol{x}_n, z_n^{(q)}))$) while leaving from the initialization $\boldsymbol{\Phi}^{(q)}$.

Algorithm 1 *bin-EM-CEM* in the case of diagonal Gaussian mixture model.

$q \leftarrow 0$ (q indicates the current iteration)
Initialization of the proportions, centers and variances to an arbitrary value $\boldsymbol{\Phi}^{(0)}$
repeat
  {E-step: computation of probabilities
  $p_r^{(q)} = P(\boldsymbol{x} \in \boldsymbol{\mathcal{H}}_r | \boldsymbol{\Phi}^{(q)})$ and
  $p_{k/r}^{(q)} = P(z_k = 1 | \boldsymbol{x} \in \boldsymbol{\mathcal{H}}_r, \boldsymbol{\Phi}^{(q)})$}
  for $r = 1$ to v **do**
    $p_r^{(q)} \leftarrow \sum_{k=1}^{K} \pi_k^{(q)} \int_{\boldsymbol{\mathcal{H}}_r} f_k(\boldsymbol{x}; \boldsymbol{\theta}_k^{(q)}) d\boldsymbol{x}$
    for $k = 1$ to K **do**
      $p_{k/r}^{(q)} \leftarrow \frac{\pi_k^{(q)} \int_{\boldsymbol{\mathcal{H}}_r} f_k(\boldsymbol{x}; \boldsymbol{\theta}_k^{(q)}) d\boldsymbol{x}}{p_r^{(q)}}$
  {C-step: computation of the partition $\mathbf{z}^{(q)}$ by MAP (Maximum *A Posteriori*)}
  for $r = 1$ to v **do**
    $z_r^{(q)} \leftarrow \underset{k}{argmax}\ p_{k/r}^{(q)}$
  {M-step: computation of the parameter $\boldsymbol{\Phi}^{(q+1)}$}
  $\boldsymbol{\Phi}^{*} = \boldsymbol{\Phi}^{(q)}$ ($*$ indicates the current iteration of the *internal binned-EM* algorithm)
  repeat
    {Internal E-step: computation of probabilities
    $p_{r/k}^{*} = P(\boldsymbol{x} \in \boldsymbol{\mathcal{H}}_r | \boldsymbol{z} = k, \boldsymbol{\Phi}^{*})$}
    for $r = 1$ to v, $k = 1$ to K **do**
      $p_{r/k}^{*} \leftarrow \int_{\boldsymbol{\mathcal{H}}_r} f_k(\boldsymbol{x}; \boldsymbol{\theta}_k^{*}) d\boldsymbol{x}$
    {Internal M-step: computation of the parameter $\boldsymbol{\Phi}^{(q+1)}$}
    for $k = 1$ to K **do**
      $\pi_k^{**} \leftarrow \frac{\sum_{r=1}^{v} n_r {z_{rk}}^{(q)}}{n}$
      $\boldsymbol{\mu}_k^{**} \leftarrow \frac{\sum_{r=1}^{v} \frac{n_r {z_{rk}}^{(q)}}{p_{r/k}^{*}} \int_{\boldsymbol{\mathcal{H}}_r} \boldsymbol{x} f_k(\boldsymbol{x}; \boldsymbol{\theta}_k^{*}) d\boldsymbol{x}}{\sum_{r=1}^{v} n_r {z_{rk}}^{(q)}}$
      $\boldsymbol{\Sigma}_k^{**} \leftarrow \frac{\operatorname{diag}\left(\sum_{r=1}^{v} \frac{n_r {z_{rk}}^{(q)}}{p_{r/k}^{*}} \int_{\boldsymbol{\mathcal{H}}_r} (\boldsymbol{x} - \boldsymbol{\mu}_k^{**})(\boldsymbol{x} - \boldsymbol{\mu}_k^{**})^T f_k(\boldsymbol{x}; \boldsymbol{\theta}_k^{*}) d\boldsymbol{x} \right)}{\sum_{r=1}^{v} n_r {z_{rk}}^{(q)}}$
    $* \leftarrow **$
  until $\left| \frac{L(\boldsymbol{\Phi}^{**}; \mathbf{a}, \mathbf{z}^{(q)}) - L(\boldsymbol{\Phi}^{*}; \mathbf{a}, \mathbf{z}^{(q)})}{L(\boldsymbol{\Phi}^{*}; \mathbf{a}, \mathbf{z}^{(q)})} \right| < \varepsilon$ [or $* = *_{max}$]
  $\boldsymbol{\Phi}^{(q+1)} \leftarrow \boldsymbol{\Phi}^{**}$
  $q \leftarrow q + 1$
until unchanged partition ($\mathbf{z}^{(q)} = \mathbf{z}^{(q-1)}$) [or $q = q_{max}$]
$\hat{\mathbf{z}} \leftarrow \mathbf{z}^{(q)}$
$\hat{\boldsymbol{\Phi}} \leftarrow \boldsymbol{\Phi}^{(q+1)}$

4. BINNED UNCERTAIN DATA

Although the use of binned data in the form of histogram constitutes a natural way of taking into account the localization imprecision of data, we could improve the data discretization procedure by discretizing the uncertainty zones rather than imprecise points. This principle is illustrated in Figure 4 in the two-dimensional case. In this figure, we consider a rectangular surface, divided into 4 bins per dimension. Then, we consider a point $\boldsymbol{x}_i$ located in the bin $\boldsymbol{\mathcal{H}}_6$ and we suppose that due to some perturbation in measurement, the position of this point is measured as being in the bin $\boldsymbol{\mathcal{H}}_{11}$. By applying the traditional discretization procedure to the imprecise point $\tilde{\boldsymbol{x}}_i$, the frequency of the bin $\boldsymbol{\mathcal{H}}_{11}$ will be incremented by 1, while the frequencies of the other bins will not be incremented. This procedure leads in this case to an error, the true point being located in the bin $\boldsymbol{\mathcal{H}}_6$. In order to solve this problem, we propose a solution which discretizes, instead of $\tilde{\boldsymbol{x}}_i$, the uncertainty zone $\boldsymbol{\mathcal{R}}_i$ built around the imprecise point $\tilde{\boldsymbol{x}}_i$, and then to increment the frequencies of the bins $\boldsymbol{\mathcal{H}}_6$, $\boldsymbol{\mathcal{H}}_7$, $\boldsymbol{\mathcal{H}}_8$, $\boldsymbol{\mathcal{H}}_{10}$, $\boldsymbol{\mathcal{H}}_{11}$ and $\boldsymbol{\mathcal{H}}_{12}$, by proportions proportional to the surfaces of overlapping between the uncertainty zone $\boldsymbol{\mathcal{R}}_i$ and these bins. The advantage of this method compared to the traditional one, is the fact of incrementing the frequency of the bin $\boldsymbol{\mathcal{H}}_6$ containing the true point, by a value different from zero, contrary to the traditional method which, in this case, attributes a null frequency to this bin.

The generalization of this principle to the multidimensional case is direct. In fact, we define *binned uncertain data* as the binned data obtained by discretizing the mul-

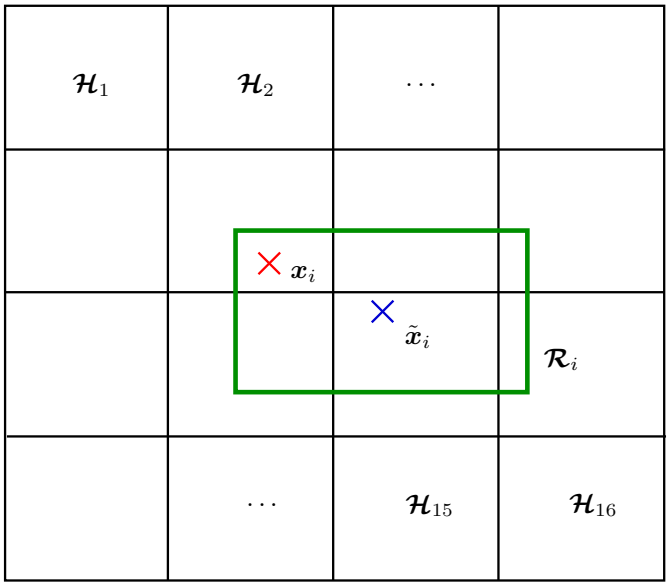

Fig. 4. Principle of discretization of uncertain data.

tivariate uncertainty zones $\mathcal{R}_1, \ldots, \mathcal{R}_n$ (cf. Section 2) according to the principle we defined above. The discretization we propose, is carried out by sharing each uncertainty zone $\mathcal{R}_i$ between several bins, *i.e.*, by incrementing the frequencies of the bins by proportions proportional to the volumes of intersections (overlappings) between the uncertainty zone $\mathcal{R}_i$ and the bins $\mathcal{H}_1, \ldots, \mathcal{H}_v$.

Figure 5 represents an example of simulated data set having a size of 50 rectangular uncertainty zones, obtained from a mixture of two Gaussian components of the space $\mathbb{R}^2$. To illustrate our new concept of binned uncertain data, figures 6 and 7 display the corresponding binned data obtained respectively by the classical discretization (discretization of imprecise points) and the new one (discretization of uncertainty zones) with a 20×20 grid (20 bins per dimension). Thus, we notice that the new data discretization procedure is very promising. Indeed, the new method tends to smooth the histogram of the observed data and the shape of the two Gaussian components of the mixture of figure 5, appears very clearly with this method (see figure 7).

Figure 8 shows the binned uncertain data corresponding to the uncertain data of Figure 3.

At the convergence of the *bin-EM-CEM* algorithm applied to binned uncertain data, the proportions of the uncertainty zone overlapping volumes (or surfaces in $\mathbb{R}^2$) with the bins $\mathcal{H}_1, \ldots, \mathcal{H}_v$, can be considered as membership degrees of $\mathcal{R}_i$ only to the corresponding bin clusters providing thus a semi fuzzy clustering. To obtain a partition of uncertain data, we need only to arrange each individual in the cluster maximizing its membership degree (MAP: Maximum *A Posteriori*).

5. AN APPLICATION TO ACOUSTIC EMISSION CONTROL

Our work was motivated by a non destructive control application for anticipating and detecting structures defects. During a pressurization control, the acoustic emission

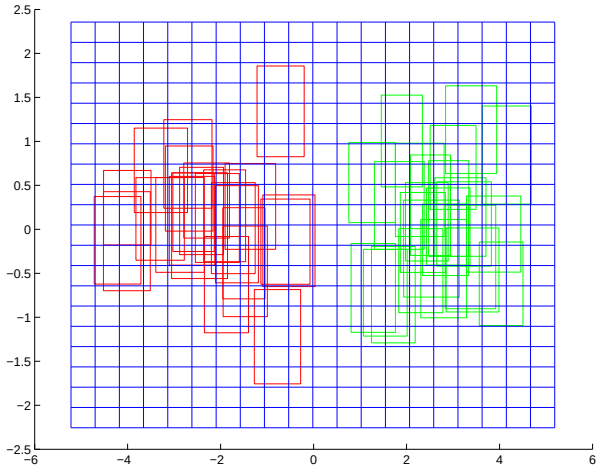

Fig. 5. Rectangular uncertainty zones obtained from a mixture of two Gaussian components of $\mathbb{R}^2$.

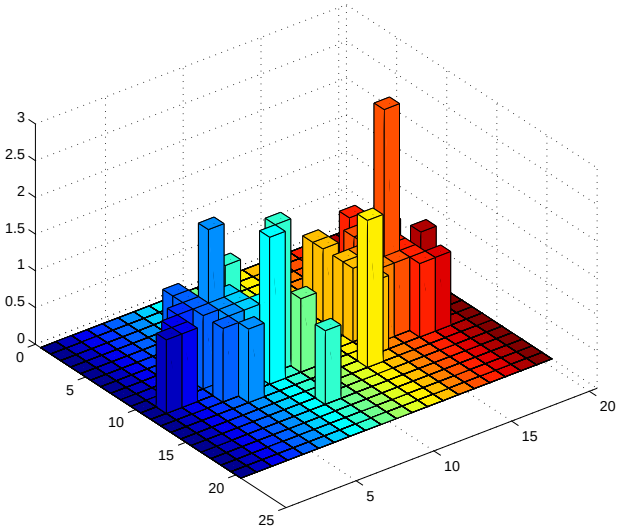

Fig. 6. Binned data obtained by the classical discretization.

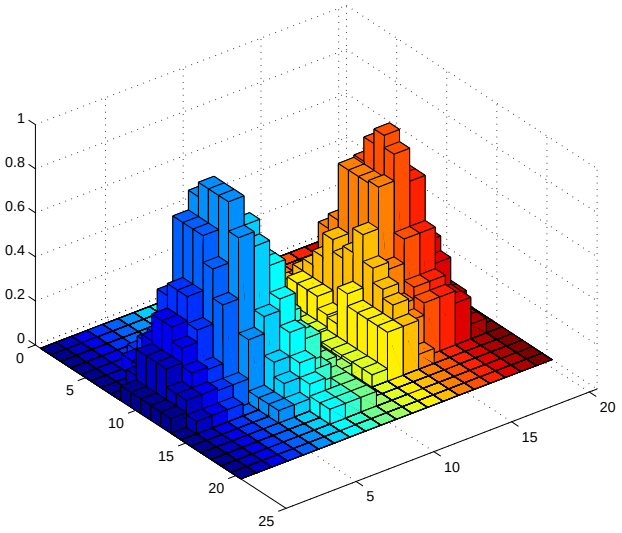

Fig. 7. Binned data obtained by the proposed discretization of uncertainty zones.

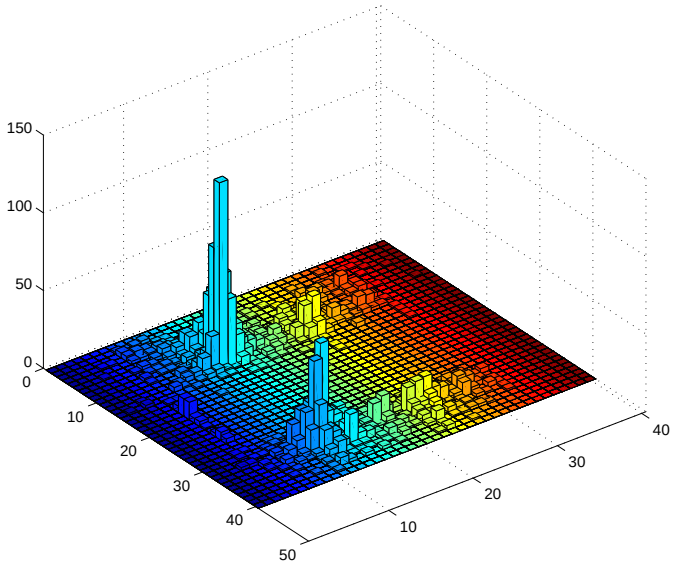

Fig. 8. Binned uncertain data corresponding to uncertain data of Figure 3.

events appear on the pressure equipment surface. Data at hand are acoustic emission event locations, in a rectangle of $\mathbb{R}^2$ representing the unfolded surface of the cylindrical pressure equipment, with the corresponding degree of localization imprecision for each acoustic emission event. The flaw diagnosis is achieved in two steps:

(1) Identification of spatial concentrations (sources) of acoustic emission events (clustering step) which is the object of the present study.
(2) Classification of identified sources in danger classes: minor, active or critical (discrimination step).

In this section, we present the results obtained by our approach applied to an acoustic emission data set of 2466 acoustic emission events. Then, we compare the partitions obtained by the algorithms *CEM*, *int-EM-CEM*, *bin-EM-CEM* and *int-bin-EM-CEM* (*bin-EM-CEM* algorithm applied to binned uncertain data). Our objective is not to move away too much from the partition provided by the *CEM* algorithm. Indeed, on the present data set, the *CEM* algorithm provides a rather acceptable result. For example, Figure 9 presents the partition obtained by the *CEM* algorithm and Figure 10 presents the final result, obtained using this clustering by *CEM*, and then Bayesian discrimination with Gaussian classes, on the present acoustic emission data set. In Figure 10, the vertically lengthened cluster corresponds to a true defect. Figure 11 shows that this defect is lengthened along the weldings of the pressure equipment, which constitute potential critical zones and, since in general they are lenghtened horizontally and vertically, imply a strong possibility to have horizontal or vertical classes of defect. This result may be considered as a validation, by the practice, of the choice of diagonal Gaussian mixture model for acoustic emission control of pressure equipment. In the rest of this section, we evaluate the precision of the algorithms *int-EM-CEM*, *bin-EM-CEM* and *int-bin-EM-CEM*, compared to that of *CEM*.

We applied, on the basis of the same initialization (provided by *CEM* applied to 'raw' acoustic emission measure-

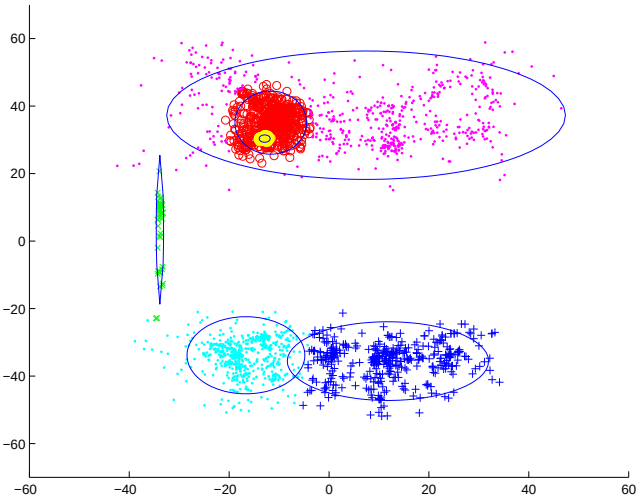

Fig. 9. Partition obtained by the *CEM* algorithm.

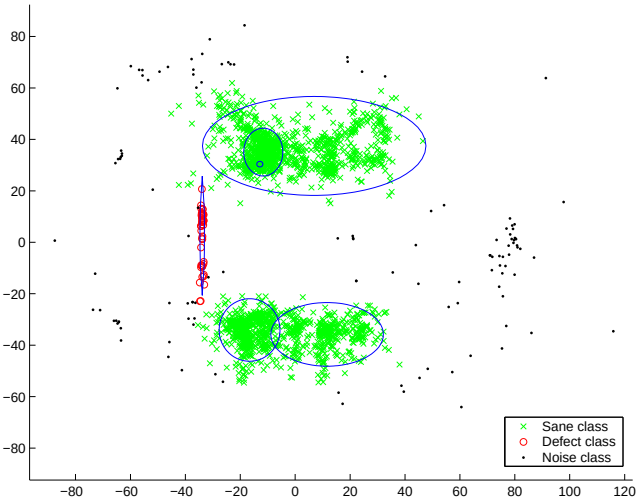

Fig. 10. Results obtained by the classical bayesian discrimination step done after the clustering step performed by CEM algorithm.

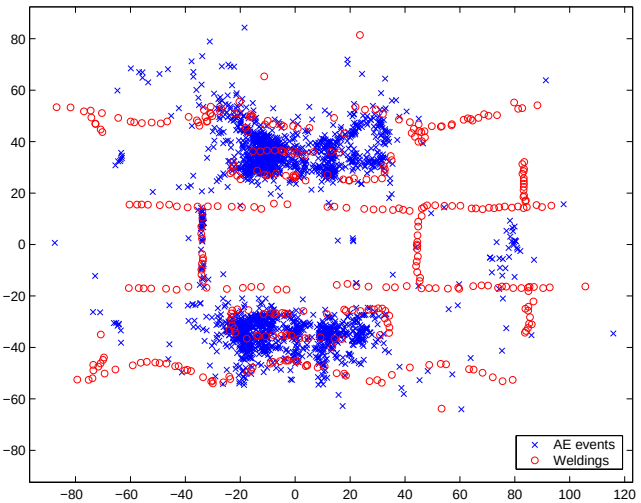

Fig. 11. Acoustic emission data and weldings on the unfolded surface of the cylindrical pressure equipment.

ments), the algorithms *int-EM-CEM*, *bin-EM-CEM* and *int-bin-EM-CEM* (see Figures 12, 13 and 14). The number of clusters, for all these algorithms, was fixed to six. For the two algorithms treating binned data (*bin-EM-CEM* and *int-bin-EM-CEM*), we considered 40 bins per dimension.

When studying the partitions obtained in Figures 9, 12, 13 and 14, we noticed that the partitions obtained by the algorithms *CEM*, *int-EM-CEM* and *bin-EM-CEM* (Figures 9, 12 and 13) are approximately identical. We have six clusters: two horizontal elliptic clusters which are well separated, a vertical cluster, and three clusters which are overlapped. Since the vertical cluster corresponding to the defect, has been identified by these algorithms, it will be discriminated as defect class by the Bayesian discrimination step. However, there is a difference between the partition obtained by the *int-bin-EM-CEM* algorithm (Figure 14), and the one given by the *CEM* algorithm. Nevertheless, the cluster lengthened vertically, is identified. If we compare this partition with the one obtained by the *bin-EM-CEM* algorithm (Figure 13), we first note that there are more non empty bins in the first partition than in the second one. This is due to the fact that when we discretize an uncertainty zone, it will be distributed, according to its surface, on several bins, and thus we get more non empty bins than when we discretize imprecise points in the classical case. On the other hand, we also note that the obtained clusters are flattened. This is due to the fact that the frequencies of the uncertainty zones when we applied the *int-bin-EM-CEM* algorithm, were distributed in a more homogeneous manner than the frequencies of the points in the *bin-EM-CEM* strategy. This homogeneous distribution (smoothing) does so that the elevated frequencies (for example the frequency of the bin belonging to the smallest cluster in Figure 13 ; see also Figures 2 and 8) are distributed between several bins while providing a tendency to flatten the clusters. Therefore, the two smaller overlapped clusters of Figure 13 have been identified by the algorithm *int-bin-EM-CEM* (see Figure 14) in one cluster having approximately the size of the biggest cluster with little flatness.

Before concluding, we insist on the fact that all the tested algorithms could identify the cluster lengthened vertically. In the decision (discrimination) step (classical bayesian discrimination with Gaussian classes) following the clustering step in our flaw diagnosis strategy, the cluster lengthened vertically, has been found to be a flaw (defect) cluster. Notice that this result is consistent with the existence of a flaw.

6. CONCLUSION

In this paper, we presented the principle of fitting the binning data procedure to imprecise data, and we defined the binned uncertain data concept. We applied the binning uncertain data procedure with the classification approach in binned data mixture model clustering, to acoustic emission data in a flaw diagnosis application. In this framework, the binning uncertain data procedure was compared to the classical binning data procedure and also to the *CEM* and *int-EM-CEM* algorithms, in the context of imprecise bivariate measurements of acoustic emission event localization. Experimentations show that

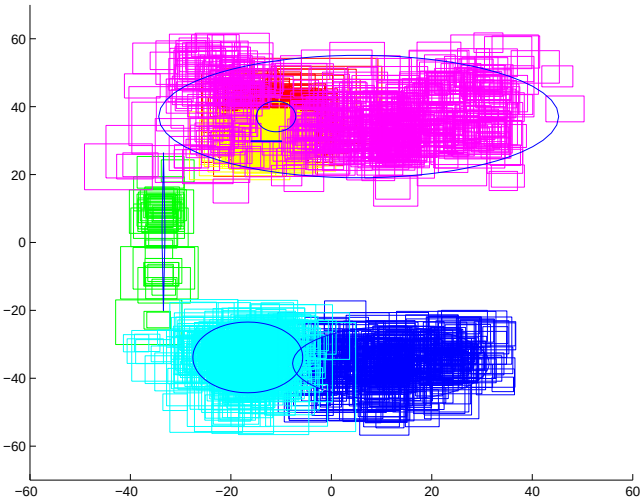

Fig. 12. Partition obtained by the *int-EM-CEM* algorithm.

[ht]

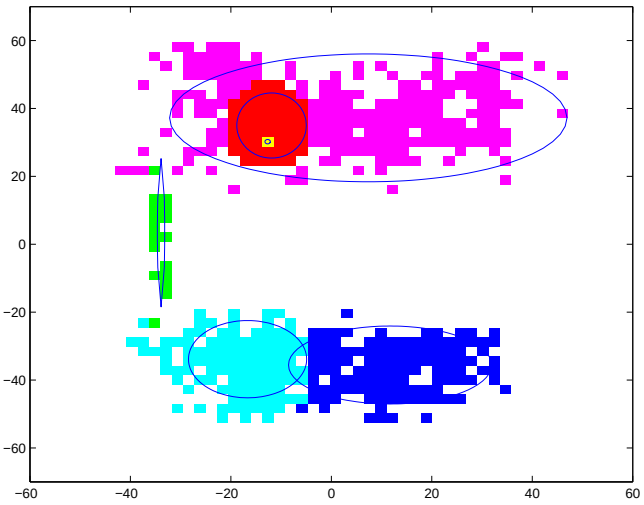

Fig. 13. Partition obtained by the *bin-EM-CEM* algorithm.

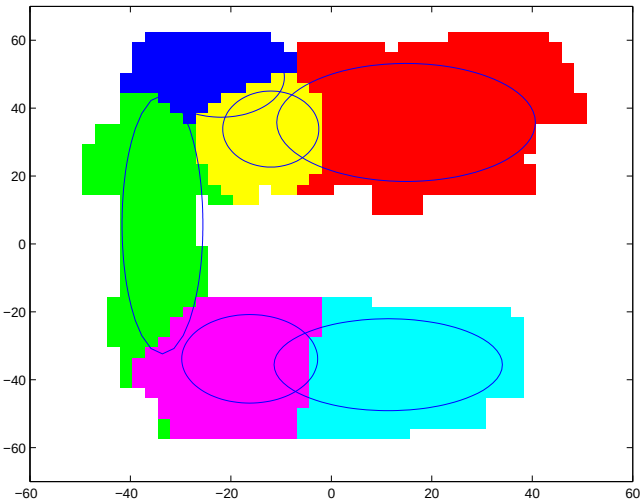

Fig. 14. Partition obtained by the *int-bin-EM-CEM* algorithm.

our method produces satisfactory results and that the use of diagonal Gaussian mixture model is well adapted to acoustic emission events clustering. The prospects of this work would be to extend the proposed method to spherical pressure equipments.

REFERENCES

Cadez, I.V., Smyth, P., McLachlan, G.J., and McLaren, C.E. (2002). Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, 47(1), 7–34.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3), 315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5), 781–793.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: The QBIC system. *Computer*, 28(9), 23–32.

Hamdan, H. and Govaert, G. (2004a). CEM algorithm for imprecise data. Application to flaw diagnosis using acoustic emission. In *IEEE International Conference on Systems, Man and Cybernetics*, 4774–4779. The Hague, The Netherlands.

Hamdan, H. and Govaert, G. (2004b). The fitting of binned data clustering to imprecise data. In *IEEE International Conference on Information & Communication Technologies: from Theory to Applications*, 1–6. Damascus, Syria.

Hamdan, H. and Govaert, G. (2004c). Int-EM-CEM algorithm for imprecise data. Comparison with the CEM algorithm using monte carlo simulations. In *IEEE International Conference on Cybernetics and Intelligent Systems*, 410–415. Singapore.

Lee, J.H., Kim, D.H., and Chung, C.W. (1999). Multi-dimensional selectivity estimation using compressed histogram information. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 205–214. ACM Press, Philadelphia, Pennsylvania, United States.

Matias, Y., Vitter, J.S., and Wang, M. (1998). Wavelet-based histograms for selectivity estimation. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 448–459. ACM Press, Seattle, Washington, United States.

Maybury, M.T. (ed.) (1997). *Intelligent Multimedia Information Retrieval*. AAAI Press and MIT Press, Cambridge, MA.

Poosala, V. (1997). *Histogram-based estimation techniques in database systems*. Ph.D. thesis, University of Wisconsin at Madison.

Samé, A. (2004). *Modèles de mélange et classification de données acoustiques en temps réel*. Ph.D. thesis, Université de Technologie de Compiègne.

Samé, A., Ambroise, C., and Govaert, G. (2006). A classification EM algorithm for binned data. *Computational Statistics and Data Analysis*, 51(2), 466–480.

Swain, M.J. and Ballard, D.H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1), 11–32.