Mixture model clustering of binned uncertain data: the classification approach

Hani Hamdan
STMS, UMR CNRS 9912, IRCAM
1, place Igor Stravinsky
75004 Paris, France
Hani.Hamdan@ircam.fr

Abstract— Basing cluster analysis on Gaussian mixture models is a powerful approach. In this context, two commonly used maximum likelihood approaches have been proposed: the mixture approach and the classification approach. Loosely speaking, the mixture approach aims to maximize the likelihood over the mixture parameters, whereas the classification approach aims to maximize the likelihood over the mixture parameters and over the identifying labels of the mixture component origin for each point. 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 algorithms, in the context of binned bivariate measurements of acoustic emission event localization.

I. 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 [1]. In image retrieval, many studies rely heavily on the use of color and feature histograms (see [2], [3] for instance). A number of techniques in approximate querying of databases and in data mining of massive data sets also use histogram representations (see [4], [5], [6] 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 [7] applied using a diagonal Gaussian mixture model [8], 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 the 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, we proposed in [9] 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 \( \mathbb{R}^p \)) with the bins. The preliminary results of this method when it was associated with the binned-EM algorithm [10] (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 [11] 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 II presents the mixture model in two different contexts: that of imprecise data and that of uncertain data. Section III explains the classification approach in imprecise data mixture model clustering. In this section, the bin-EM-CEM algorithm [11], [12] is described and summarized. In Section IV, we present the principle of fitting the binning data procedure to imprecise data, and we define the imprecise data concept. Section V describes briefly an application of this approach to flaw diagnosis, on pressure equipments, using acoustic emission. The binning 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 [13], [14], are also done.

II. THE MODEL

In this work, we suppose that \( ((x_1, z_1), \ldots, (x_n, z_n)) \) is a sample resulting from the couple of random vectors \((X, Z)\) associated to a distributions mixture defined on \( \mathbb{R}^p \) by:

\[
f(x; \Phi) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)
\]

where \( \Phi = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K) \) and \( \pi_1, \ldots, \pi_K \) are the proportions of the mixture and \( \theta_1, \ldots, \theta_K \) the parameters of each component density ; \( z_i (1 \leq i \leq n) \) indicates the origin component of \( x_i (1 \leq i \leq n) \) and we note \( z = (z_1, \ldots, z_n) \).

We consider that every point \( x_i \) belongs to an uncertainty zone \( \mathcal{R}_i \) of \( \mathbb{R}^p \) and we suppose that the only available knowledge of the sample \( ((x_1, z_1), \ldots, (x_n, z_n)) \) is the set of uncertainty zones \( \mathcal{R}_i (1 \leq i \leq n) \) as \( x_i \in \mathcal{R}_i \) (\( 1 \leq i \leq n \)).

We note \( \mathbf{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_n) \) the vector of these zones. On the other hand, we consider a partition \( \mathcal{H}_1, \ldots, \mathcal{H}_n \) 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 \( ((x_1, z_1), \ldots, (x_n, z_n)) \) is the set of the frequencies \( n_i \) of \( x_i \) belonging to \( \mathcal{H}_v \). We note \( n = (n_1, \ldots, n_v) \) the vector of these frequencies \((\sum_{v=1}^{v} n_v = n)\). Since there is no available information about the exact positions of the points \( x_i \) into the bins \( \mathcal{H}_1, \ldots, \mathcal{H}_v \), 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 \( z_i (1 \leq i \leq n) \) can be replaced by the labels \( z_r (1 \leq r \leq v) \) of the bins if the bin membership of each \( x_i (1 \leq i \leq n) \) is known, and \( z_r (1 \leq r \leq v) \) is coded in the form \( z_r = (z_{r1}, \ldots, z_{rv}) \) where \( z_{rk} \) is equal to 1 if \( \mathcal{H}_v \) is resulting from the component \( k \) and is equal to 0 elsewhere.

III. THE CLASSIFICATION APPROACH IN BINNED DATA MIXTURE MODEL CLUSTERING

In this paragraph, we describe the classification approach [11] for simultaneous estimation of the mixture parameter \( \Phi \)
and the partition \((z_1, \ldots, z_v)\) maximizing the log-likelihood criterion

\[
L(\Phi; a, z) = L(\Phi; n_1, \ldots, n_v, z_1, \ldots, z_v) = \log P(a; \Phi) = \sum_{r=1}^n \log \left( z_{r, z_r} \int \phi_{z_r}(x; \theta_{z_r}) dx \right) + \log \prod_{v=1}^v \text{Pr}(z_v). \tag{2}
\]

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

1) computation of

\[
z^{(q)} = \arg\max_z L(\Phi^{(q)}, a, z).
\]

2) computation of

\[
\Phi^{(q+1)} = \arg\max_\Phi L(\Phi; a, z^{(q)}).
\]

The development of these two steps leads to the bin-EM-CEM algorithm [11], [12] which we summarized in Algorithm 1, 

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

\[
q \leftarrow 0 \quad (q \text{ indicates the current iteration})
\]

**Initialization** of the proportions, centers and variances to an arbitrary value \(\Phi^{(0)}\)

repeat

\{E-step: computation of probabilities \(p^{(q)}_r = P(x \in \mathcal{H}_r | \Phi^{(q)})\) and \(p^{(q)}_{k/r} = P(z_k = 1 | x \in \mathcal{H}_r, \Phi^{(q)})\) \}

for \(r = 1 \) to \(v \) do

\[
p^{(q)}_r = \sum_{k=1}^{K} \text{Pr}(z_k = 1 | x \in \mathcal{H}_r, \theta_{k}) dx
\]

for \(k = 1 \) to \(K \) do

\[
p^{(q)}_{k/r} = \frac{\text{Pr}(z_k = 1 | x \in \mathcal{H}_r, \theta_{k}) dx}{\text{Pr}(z_k = 1 | x \in \mathcal{H}_r, \theta_{k}) dx}
\]

\{C-step: computation of the partition \(z^{(q)} \) by MAP (Maximum A Posteriori)\}

for \(r = 1 \) to \(v \) do

\[
z^{(q)}_r \leftarrow \arg\max_k p^{(q)}_{k/r}
\]

\{M-step: computation of the parameter \(\Phi^{(q+1)}\)\}

\[
\Phi^{(q+1)} = \Phi^{(q)} \quad (q \text{ indicates the current iteration of the internal binned-EM algorithm})
\]

repeat

\{Internal E-step: computation of probabilities \(p^{(q)}_{r/k} = P(x \in \mathcal{H}_r | z = k, \Phi^{(q)})\)\}

for \(r = 1 \) to \(v \), \(k = 1 \) to \(K \) do

\[
p^{(q)}_{r/k} \leftarrow \text{Pr}(z_k = 1 | x \in \mathcal{H}_r, \theta_{k}) dx
\]

\{Internal M-step: computation of the parameter \(\Phi^{(q+1)}\)\}

for \(k = 1 \) to \(K \) do

\[
\mu_{k}^{(q+1)} \leftarrow \frac{\sum_{r=1}^{v} n_{r,k}(z^{(q)}) \int \phi_{z_r}(x; \theta_{z_r}) dx}{\sum_{r=1}^{v} n_{r,k}(z^{(q)})}
\]

\[
\Sigma_{k}^{(q+1)} \leftarrow \text{diag} \left( \sum_{r=1}^{v} n_{r,k}(z^{(q)}) \int (x - \mu_{k}^{(q+1)}(z^{(q)})) (x - \mu_{k}^{(q+1)}(z^{(q)}))^T \phi_{z_r}(x; \theta_{z_r}) dx \right)
\]

\[
\phi^{(q+1)} \leftarrow \text{argmax} \left( \frac{L(\Phi^{(q+1)}; a, z^{(q)}) - L(\Phi^{(q)}; a, z^{(q)})}{L(\Phi^{(q+1)}; a, z^{(q)})} \right) < \epsilon \quad \text{or} \quad q = q_{\text{max}}
\]

until \(q = q_{\text{max}}\)

\[\Phi^{(q+1)} \leftarrow \Phi^{(q)}
\]

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 multivariate uncertainty zones \(\mathcal{R}_1, \ldots, \mathcal{R}_m\) (cf. Section II) according to the principle we defined above. The discretization we propose, is carried out by sharing each uncertainty zone \(\mathcal{R}_j\) between several bins, i.e., by incrementing the frequencies of the bins by proportions proportional to the volumes of intersections (overlapping) between the uncertainty zone \(\mathcal{R}_j\) and the bins \(\mathcal{H}_1, \ldots, \mathcal{H}_m\).
to need the uncertain data of zone overlapping only on the cation for a 100 50 V.

At the Our work binned uncertain, to only can the corresponding data, each individual in the cluster maximizing its membership degree (MAP: Maximum A Posteriori).

V. 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 events appear on the pressure equipment surface. Data at hand are acoustic emission event locations, in a rectangle of $R^2$ 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 6 presents the final result, obtained using clustering by CEM, and then Bayesian discrimination with Gaussian classes, on the present acoustic emission data set. In this Figure, the vertically lengthened cluster corresponds to a true defect. Figure 7 shows that this defect is lengthened along the weldings of the pressure equipment, which constitute potential critical zones and, since in general they are lengthened 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 measurements), the algorithms int-EM-CEM, bin-EM-CEM and int-bin-EM-CEM (see Figures 9, 10 and 11). 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 8, 9, 10 and 11, we noticed that the partitions obtained by the algorithms CEM, int-EM-CEM and bin-EM-CEM (Figures 8, 9 and 10) 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 11), 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 10), we first note that there are more non empty bins in the first.

Fig. 4. Principle of discretization of uncertain data.

Fig. 5. Binned uncertain data corresponding to uncertain data of Figure 3.

Figure 5 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 $R^2$) with the bins $H_1, \ldots, H_6$, can be considered as membership degrees of $R_i$ 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).
Fig. 6. Results obtained by the classical bayesian discrimination step done after the clustering step performed by CEM algorithm.

Fig. 7. Acoustic emission data and weldings on the unfolded surface of the cylindrical pressure equipment.

Fig. 8. Partition obtained by the CEM algorithm.

Fig. 9. Partition obtained by the int-EM-CEM algorithm.

Fig. 10. Partition obtained by the bin-EM-CEM algorithm.

Fig. 11. Partition obtained by the int-bin-EM-CEM algorithm.
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 10 ; see also Figures 2 and 5) are distributed between several bins while providing a tendency to flatten the clusters. Therefore, the two smaller overlapped clusters of Figure 10 have been identified by the algorithm int-bin-EM-CEM (see Figure 11) 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.

VI. 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 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.

ACKNOWLEDGMENT

This work is done during the stay of the author at the U.M.R. CNRS 6599 Heudiasyc of the Université de Technologie de Compiègne (UTC).

REFERENCES