The fitting of binned data clustering to imprecise data.

Hani HAMDAN (1,2) and Gérard GOVAERT (1).

(1) HEUDIASYC, UMR CNRS 6599, Université de Technologie de Compiègne
Centre de Recherches de Royallieu, BP 20529 - 60205 COMPIEGNE Cedex, FRANCE
Hani.Hamdan@utc.fr, Gerard.Govaert@utc.fr
(2) Centre Technique des Industries Mécaniques (CETIM)
52, Avenue Félix-Louat, BP 80067 - 60304 SENLIS Cedex, FRANCE
Hani.Hamdan@cetim.fr

KEYWORDS:
binned EM algorithm, interval EM algorithm, mixture model, clustering, fuzzy clustering, semi fuzzy clustering, imprecise data, interval data, binned data, KL distance, acoustic emission.

TOPICS:
Pattern recognition

EXTENDED ABSTRACT:

This paper addresses the problem of taking into account the data imprecision in the clustering of binned data using mixture models and binned EM algorithm [1]. Within the framework of a defects detection problem by acoustic emission control, we were brought to treat a set of points using the EM algorithm [2,3] applied to a diagonal Gaussian mixture model [4]. This one provides a satisfactory solution but the real time constraints imposed in our problem make its application impossible when the number of points becomes too big. As data sets become larger, data processing becomes increasingly complex and as a result, the data analysis will be expensive in computation time. The solution that we propose is to group data and available data will thus takes the form of a histogram. Such data are also called binned data. Binning data is common in data analysis and machine learning. An EM approach to solve this problem was already proposed [5] in dimension 1 and was then generalized to the multidimensional case [1] to deal with binned data. This EM approach requires the evaluation of multidimensional integrals over each bin at each iteration. Naive implementation of the procedure can lead to computationally inefficient results. To reduce the computational cost, a number of straightforward numerical techniques are proposed in [1]. In the diagonal Gaussian mixture model [4] case, these integrals can be computed by simply using the one-dimensional normal cumulative distribution function. In this paper, we fit the binning data procedure to imprecise data. We model imprecise data by multivariate uncertainty zones and we propose to assign each uncertainty zone to several bins with percentages proportional to its overlapping surfaces with the bins. The experimental results compare this binning procedure with the classical one (applied to imprecise points) and with the interval EM algorithm [6] considered here as a reference, using simulated data.

REFERENCES:

[1] I. V. Cadez, P. Smyth, G. J. McLachlan and C. E. McLaren. Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. Machine Learning 47, 7-34, 2001.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Pattern Recognition, 28(5), J. Royal Stat. Soc. B, 39(1):1-38, 1977.

[3] R. A. Redner and H. F. Walker. Mixtures Densities, Maximum Likelihood and the EM Algorithm. SIAM Review, 26 : 195-239, 1984.

[4] Celeux G. and G. Govaert. Gaussian parsimonious clustering models. Pattern Recognition, 28(5), 781-793, 1995.

[5] McLachlan G. J. and P. N. Jones. Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics, 44(2):571-578, 1988.

[6] H. Hamdan and G. Govaert. Clustering of imprecise data. Application to flaw diagnosis using acoustic emission. In IEEE International Conference on Control Applications, Taipei, Taiwan, 2-4 september 2004.