A binned technique for scalable model-based clustering on huge datasets

Filippo Antonazzo, Christophe Biernacki, Christine Keribin

A binned technique for scalable model-based clustering on huge datasets

*Filippo Antonazzo, *Christophe Biernacki, †Christine Keribin

*Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France

†Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay 91405 Orsay, France

Abstract

Clustering is impacted by the regular increase of sample sizes which provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to the need of many computational resources and also to high energy consumption. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the quality of the related estimation. After a brief review of existing methods, a first application in the context of univariate model-based clustering is provided, with a numerical illustration of its advantages. The issues of a trivial multivariate extension are discussed and a marginal-binned strategy is proposed to estimate bivariate Gaussian diagonal mixtures.

Keywords: Big Data, clustering, binned data, green computing.

1 Scalable clustering for huge datasets

Today, thanks to the technological development of the last decades, it is common to work on *huge datasets*, which are large collections of data whose volume (both of observations and attributes) is still growing. But, despite the enormous statistical information conveyed, any statistical analysis, such as clustering, conducted with classical methods is difficult because it requests too much time, too much memory and too much energy. This is also in contrast with the current eco-friendly policies of many national governments and industries which are searching for methods able to do suitable statistical analysis without employing complex and wasteful technologies. We want to satisfy this need, proposing a method capable to analyse big data employing limited computational resources, like those of a standard laptop.

For the same reasons, scalable clustering algorithms for huge datasets flourished in literature during the last two decades. Some algorithms employ data-reduction techniques, like random subsampling [9] or data-compression through the use of sufficient statistics [14]. Other authors transform the space of analysis [11] or examine dense data units built imposing a grid on the original data [1]. It is also possible to reduce the number of operations, adopting particular data structure, such as trees [14] or graphs [9], or imposing some criteria [1] to prune irrelevant clusters that, thus, exit from the computational process. In addition, the problem of dimensionality is usually tackled down by performing clustering in subspaces of lower dimension [2].

The objective of the paper is to introduce scalability in model-based clustering [8], a statistical approach well appreciated because it enables a theoretically well-posed framework where formal criteria to assess the quality of the clustering are available. It is in this context that we will propose our novel method based on binned data, which, assuming observations with values belonging to a real space $\mathcal{X}$, correspond to a reduced dataset only containing the counts of observations in given regions of $\mathcal{X}$. In practice they usually appear as soon as it is impossible to collect data with infinite precision, like in [7] and [3], but we will use them with a different point of view. The key idea we defend is to group original data in order to obtain *artificially* binned ones and reduce the dimensionality of the problem working with them. We first consider the univariate case (where $\mathcal{X} = \mathbb{R}$) to introduce the notation and highlight, through a numerical example, how much promising is our method. Finally, we discuss how to extend it to the multivariate context, pointing out possible issues of trivial generalizations and presenting a new marginal-binned methodology able to cope with them in a restrictive bivariate diagonal scenario, as a final simulation shows.

2 Binned model-based clustering approach: univariate case

Let $\mathbf{x} = (x_1, \ldots, x_n)$, with $x_i \in \mathcal{X} = \mathbb{R}$, a raw sample of n observations arising from a univariate K-Gaussian mixture of density

$$f(x; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2) \quad \pi_k > 0, \ \ \sum_{k=1}^{K} \pi_k = 1, \tag{1}$$

in which μ_k denotes the mean of the k-th component, σ_k^2 is its variance and $\boldsymbol{\theta}$ is the vector that contains all the parameters, thus $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. The key-idea is to build a grid G made of $R << n$ cut points $(a_1, \ldots, a_R)$ that divides the real space $\mathbb{R}$ into $R+1$ intervals $[a_{r-1}, a_r[$, $r = 1, \ldots, R+1$, setting $a_0 = -\infty$ and $a_{R+1} = \infty$. In this way, binned data are stored in a vector $\mathbf{y} = (y_1, \ldots, y_{R+1})$, where each component is defined as

$$y_r = \#\{x_i : a_{r-1} \leq x_i < a_r\}. \tag{2}$$

As $R \ll n$, working with binned data instead of raw ones reduces the dimensionality of the problem and also proposes interesting theoretical questions.

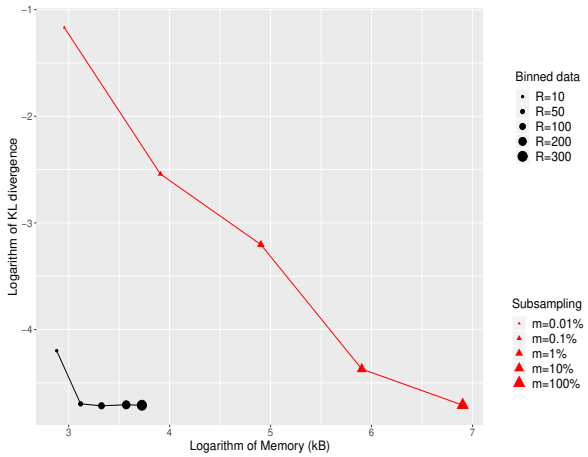

Figure 1: Binned estimation of a simulated 3-class mixture: logarithm of Kullback-Leibler divergence between the true mixture distribution and the estimated one for different values of R and m in function of the required computer memory (logarithmic scale).

In fact, the binned statistical model is a multinomial one $M(n, p(\boldsymbol{\theta}))$ with $p(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), \ldots, p_{R+1}(\boldsymbol{\theta}))$, where $p_r(\boldsymbol{\theta}) = \int_{a_{r-1}}^{a_r} f(x; \boldsymbol{\theta}) dx$. It could be proved (result not provided here) that this model remains identifiable under certain (and weak) conditions on the grid G.

Here is a numerical example to motivate the fundamental interest of our proposed "binned" method, which is compared to the subsampling strategy (depending on the subsample percentage m) on a simulation sample of $n = 10^6$ raw data i.i.d. arising from a univariate Gaussian mixture with three components. Binned data are created through a grid with the tuning parameter R. An EM algorithm [4] is performed respectively with different values of R and m (thus different candidate subsample and binned datasets). In Figure 1 it is possible to note that the loss of information (measured by the Kullback-Leibler divergence) induced by binning is much lower than that obtained with subsampling, even negligible if we use a grid moderately dense. This is in addition accompanied by an evident gain in terms of computer memory. Such promising results could be also obtained (but not displayed here) concerning gain in terms of algorithm running time or model selection behaviour.

3 Issues of a trivial multivariate extension

Once analyzed the univariate case, extending the method to a D-variate situation seems straightforward. Let $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}^D$, a sample arising from a multivariate K-Gaussian mixture of density

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}; \mu_k, \Sigma_k) \quad \pi_k > 0, \ \sum_{k=1}^{K} \pi_k = 1, \tag{3}$$

where, for each component $k = 1, \ldots, K$, $\mu_k = (\mu_{k1}, \ldots, \mu_{kD})$ is the vector of means and Σ_k is the variance-covariance matrix, with diagonal $(\sigma^2_{k1}, \ldots, \sigma^2_{kD})$. It is immediate to define a multivariate grid G building it as a Cartesian product between D one-dimensional grids. It means that $G = G_1 \times \ldots \times G_D$, where each grid G_d has R_d cut points $(a_{d1}, \ldots, a_{dR_d})$. Assuming that $R_d = R$, for $d = 1, \ldots, D$, we can define a $(R+1)^D$-dimensional binned vector $\mathbf{y} = (y_1, \ldots, y_{(R+1)^D})$, where, for $r = 1, \ldots, (R+1)^D$:

$$y_r = \#\{\boldsymbol{x}_i : 1 + z_{i1} + z_{i2}(R+1) + z_{i3}(R+1)^2 \ldots + z_{iD}(R+1)^{D-1} = r\},$$
$$\text{with} \quad z_{id} = l \text{ if } a_{dl} \leq x_{id} < a_{d(l+1)},\ l = 0, \ldots, R, \ \ \forall d = 1, \ldots, D,$$

where $a_{d0} = -\infty$ and $a_{d(R+1)} = \infty$ for each $d = 1, \ldots, D$.

Despite the relatively simple formalization, using such a grid is not feasible. Indeed, the following issues arise:

- It is impossible to obtain a manageable amount of binned data because the number of non-empty bins increases exponentially increasing the number of variables (proof not provided here).
- The related EM algorithm employs several multidimensional numerical integrations. Thus, it would become too complex in terms of computing time.

Consequently, we propose below a specific alternative strategy (called "marginal-binned") to estimate multivariate diagonal mixtures not affected by these problems. For simplicity, we will illustrate it in a restrictive bivariate scenario, where $\mathcal{X} = \mathbb{R}^2$, even if the proposal is more general.

4 A marginal-binned strategy for bivariate diagonal mixtures

Let consider a bivariate ($D = 2$) diagonal Gaussian mixture with K components. Thus, the variances Σ_k in (3) are diagonal and the vector of parameters is simply:

$$\boldsymbol{\theta} = (\underbrace{\pi_1, \ldots, \pi_K}_{\pi}, \underbrace{\mu_{11}, \ldots, \mu_{K1}, \sigma^2_{11}, \ldots, \sigma^2_{K1}}_{\boldsymbol{\alpha_1}}, \underbrace{\mu_{12}, \ldots, \mu_{K2}, \sigma^2_{12}, \ldots, \sigma^2_{K2}}_{\boldsymbol{\alpha_2}}).$$

Denoting with $\mathbf{x}_1$ and $\mathbf{x}_2$ the first and the second component of a sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{x}_i \in \mathbb{R}^2$, and adopting a square grid $G = G_1 \times G_2$ with $R_1 = R_2 = R$, we define:

- $\mathbf{y}_1$: binned data vector of $\mathbf{x}_1$ under G_1;
- $\mathbf{y}_2$: binned data vector of $\mathbf{x}_2$ under G_2.

It means that, for each $d = 1, 2$, $\mathbf{y}_d = (y_{d1}, \ldots, y_{d(R+1)})$, where each component is defined as $y_{dr} = \#\{x_{di} : a_{d(r-1)} \leq x_{di} < a_{dr}\}$. We name $\mathbf{y}_1$ and $\mathbf{y}_2$ as

the *marginal counts* of $\mathbf{y}$. By construction, they are equivalent to the counts obtained by binning the univariate marginals of the joint distribution. It can be observed that each of them is a binned data vector arising from a univariate mixture with density $f_d(x_d; \boldsymbol{\theta}_d) = \sum_{k=1}^{K} \pi_k \phi(x_d; \mu_{kd}, \sigma^2_{kd})$, with parameter $\boldsymbol{\theta}_d = (\pi, \boldsymbol{\alpha_d})$.

Given the one-dimensional binned log-likelihoods $\ell_1(\boldsymbol{\theta}_1; \mathbf{y}_1)$ and $\ell_2(\boldsymbol{\theta}_2; \mathbf{y}_2)$, it is possible to obtain an estimate of $\boldsymbol{\theta}$ maximizing their sum $cl(\theta; \mathbf{y}_1, \mathbf{y}_2) = \ell_1(\boldsymbol{\theta}_1; \mathbf{y}_1) + \ell_2(\boldsymbol{\theta}_2; \mathbf{y}_2)$. This method is not new in literature: in fact, it is known as *composite likelihood estimation*, firstly introduced in [6], who also gives interesting theoretical properties of the estimators obtaining by maximizing the *composite likelihood* $cl(\theta; \mathbf{y}_1, \mathbf{y}_2)$, like consistency and asymptotic distribution. Important contributions are given in [5] and [12], who furnished, in a composite likelihood framework, a specific formulation of the EM algorithm and an application with binned data, respectively. In a mixture model context, a similar approach is followed by [10], but in a problem involving discrete data, with a more complex formulation and without taking into account the computational and memory issues mentioned in Section 3.

Combining the ideas contained in [5] and [12], we developed a new marginal-binned EM algorithm maximizing $cl(\theta; \mathbf{y}_1, \mathbf{y}_2)$ (details not displayed here) and we tried it on simulated data sets of size $n = 10^6$, generated by different bivariate diagonal mixture models with, for simplicity, two components. In particular, it is interesting to show results obtained in a difficult scenario, where the two components are not well separated: this is useful to illustrate the goodness of the proposed methodology. These ones are depicted in Figure 2, where the 0.95 density ellipses for the real and the estimated densities (with $R = 40$) of the two components are shown. It is possible to note that they are very close, as well as the respective means, denoting a good quality of estimation, despite the difficulty of the situation. The outcomes regarding time and memory performances confirm the results of the univariate simulation presented in Section 2, thus they are not displayed here.

5 Ongoing works

The depicted methodology has proved to be efficient both from the point of view of statistical quality and computational resources management. But, some problems remain open. Firstly, it is impossible to estimate non-diagonal mixtures using only marginal counts. However, we wonder if it is possible to recover an acceptable trade-off between computational savings and clustering quality using our marginal-binned strategy. In the section dedicated to the multivariate scenario we did not mention the problem of model selection: in [13] it is possible to find some choice criteria specific for composite likelihood estimations but their calculation could be too burdensome. So, it is important to find a criterion demanding a lighter computational effort. Finally, the crucial point of the work is grid selection. We aim to find a criterion able to select the grid providing an optimal estimation (in terms of statistical quality) without neglecting the main

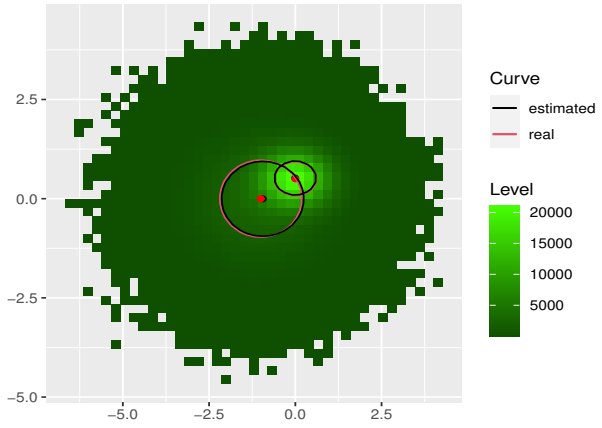

Figure 2: 0.95 density ellipses and means for the two components of the real density mixture (in red) and of the estimated one (in black). In background, the levelplot of the true density.

purpose of this methodology: saving energetic resources.

References

[1] Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, 94-105 (1998)

[2] Böhm, C., Kailing, K., Kröger, P. & Zimek, A.: Computing clusters of correlation connected objects. In Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 455-466 (2004)

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

[4] 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 (Methodological), **39**(1), 1-22 (1977)

[5] Gao, X. & Song, P. X. K.: Composite likelihood EM algorithm with applications to multivariate hidden Markov model. Statistica Sinica, 165-185 (2011)

[6] Lindsay, B. G.: Composite likelihood methods. Contemporary mathematics, **80**(1), 221-239 (1988)

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

[8] McLachlan, G. J., & Peel, D.: Finite mixture models. John Wiley & Sons (2004)

[9] Ng, R. T. & Han, J.: CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, **14**(5), 1003-1016 (2002)

[10] Ranalli, M., & Rocci, R.: Mixture models for ordinal data: a pairwise likelihood approach. Statistics and Computing, **26**(1-2), 529-547 (2016)

[11] Sheikholeslami, G., Chatterjee, S. & Zhang, A.: Wavecluster: A multi-resolution clustering approach for very large spatial databases. In VLDB **98**, 428-439 (1998)

[12] Whitaker, T., Beranger, B. & Sisson, S. A.: Composite likelihood methods for histogram-valued random variables. Statistics and Computing, 1-19 (2020)

[13] Varin, C., Reid, N. & Firth, D.: An overview of composite likelihood methods. Statistica Sinica, 5-42 (2011)

[14] Zhang, T., Ramakrishnan, R. & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, **1**(2), 141-182 (1997)