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

Filippo Antonazzo, Christophe Biernacki, Christine Keribin

▶ **To cite this version:**

Filippo Antonazzo, Christophe Biernacki, Christine Keribin. A binned technique for scalable model-based clustering on huge datasets. MBC2 - Models and Learning for Clustering and Classification, Sep 2020, Catania, Italy. hal-03097284v1

HAL Id: hal-03097284
https://hal.science/hal-03097284v1

Submitted on 5 Jan 2021 (v1), last revised 5 Jan 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Filippo Antonazzo, Christophe Biernacki & Christine Keribin

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. Finally, an initial formalization of the multivariate extension is done, highlighting both issues and possible strategies.

Key words: Clustering, binned data, big data, green computing.

1 Scalable clustering for huge datasets

Today, thanks to the technological development of the last decades, it is very easy 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

Filippo Antonazzo
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France e-mail: filippo.antonazzo@inria.fr

Christophe Biernacki
Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France e-mail: christophe.biernacki@inria.fr

Christine Keribin
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay
91405 Orsay, France, e-mail: christine.keribin@universite-paris-saclay.fr

classical methods is difficult because it request too time, too memory and too 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 good 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 [7] or data-compression through the use of sufficient statistics [10]. Other authors transform the space of analysis [8] 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 tree [10] or graph [7], 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].

Each of these methods does not assume a statistical model behind the generation process of data. This one is on the contrary a distinct approach of what is known as model-based clustering [6], that 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 $\mathscr{X}$, correspond to a reduced dataset only containing the counts of observations in given regions of $\mathscr{X}$. In practice they usually appear as soon as it is impossible to collect data with infinite precision, but we will use binned data 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. The starting points for the use of binned data in model-based clustering are [5] and [3].

2 Binned model-based clustering approach: univariate case

Considering first the univariate case (where $\mathscr{X} = \mathbb{R}$) is necessary to introduce the notation and highlight how much promising is our method. Then we will discuss a possible extension to the multivariate context.

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

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

in which μ_k denotes the mean of the k-th component, σ_k^2 is its variance and θ is the vector that contains all the parameters, thus $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. The key-idea is to build a grid G made of $R << n$ points $(a_1, \ldots, a_R)$ that divides the real space $\mathbb{R}$ into $R+1$ intervals $[a_{j-1}, a_j[, \ \ j = 1, \ldots, R$, 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

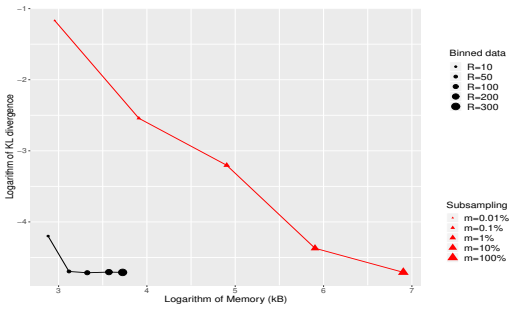

Fig. 1 Logarithm of Kullback-Leibler divergence from the true parameters for different values of R and m in function of the required computer memory (logarithmic scale).

each component is defined as

$$y_j = \#\{x_i : a_j \leq x_i < a_{j+1}\}. \tag{2}$$

As $R << n$, working with binned data instead of raw ones reduces the dimensionality of the problem and also proposes interesting theoretical questions. In fact, the binned statistical model is a multinomial one $M(n, p(\theta))$ with $p(\theta) = (p_1(\theta), \ldots, p_R(\theta))$, where $p_l(\theta) = \int_{a_{l-1}}^{a_l} f(x;\theta)dx$. It could be proved that this model remains identifiable under certain conditions on G.

Finally, here is a numerical example to motivate our proposed "binned" method, which was compared to the subsampling strategy (depending on the subsample percentage m) on a simulation sample of $n = 10^6$ raw data i.i.d. arises from a univariate Gaussian mixture with three components. Binned data are created through a grid with a tuning parameter R corresponding to the number of its points. An EM algorithm [4] was 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 Multivariate extension: issues and strategies

Once analyzed the univariate case, extending the method to a d-variate situation is straightforward. Given a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^d$, we can define a multivariate grid G building it as a cartesian product between d one-dimensional grid. It means that $G = G_1 \times \ldots \times G_d$, where each grid G_j has R_j points $(a_{j1}, \ldots, a_{jR_j})$. Assuming that $R_j = R$, for $j = 1, \ldots, d$, we can define a $(R+1)^d$-dimensional binned vector $\mathbf{y} = (y_1, \ldots, y_{(R+1)^d})$, where, for $k = 1, \ldots, (R+1)^d$:

$$y_k = \#\{x_i : 1 + z_i^1 + z_i^2(R+1) + z_i^3(R+1)^2 \ldots + z_i^d(R+1)^{d-1} = k\},$$
$$\text{with} \quad z_i^j = l \text{ if } a_{jl} \leq x_i < a_{j(l+1)},\ l = 0, \ldots, R, \ \ \forall j = 1, \ldots, d,$$

where $a_{j0} = -\infty$ and $a_{j(R+1)} = \infty$ for each $j = 1, \ldots, d$.

Despite the relatively simple formalization, using such a grid is not feasible. In fact, 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.
- The EM algorithm used employs several multidimensional numerical integrations. Thus, our algorithm would become too complex in terms of time.

In order to provide a solution to these issues, we developed some strategies:

1. Reducing the multidimensional problem to multiple one-dimensional ones, performing our univariate method on each dimension and combining the results.
2. Using simpler algorithms approximating EM, which avoid multidimensional integrations substituting them by combinations of one-dimensional ones.
3. Group or remove variables.
4. Imposing sparse grids (composed by two or three bins) on those variables that have low statistical information.

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. McLachlan, G. J. & Jones, P. N.: Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics, 571-578 (1988)
6. McLachlan, G. J., & Peel, D.: Finite mixture models. John Wiley & Sons (2004)
7. 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)
8. Sheikholeslami, G., Chatterjee, S. & Zhang, A.: Wavecluster: A multi-resolution clustering approach for very large spatial databases. In VLDB **98**, 428-439 (1998)
9. Zayani, A., N'Cir, C. E. B. & Essoussi, N.: Parallel clustering method for non-disjoint partitioning of large-scale data based on spark framework. In 2016 IEEE International Conference on Big Data (Big Data), 1064-1069 (2016)
10. 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)