Consensus Big Data Clustering for Bayesian Mixture Models

Subjects: Statistics & Probability

Contributor: Christos Karras, Aristeidis Karras, Konstantinos C. Giotopoulos, Markos Avlonitis, Spyros Sioutas

In the context of big-data analysis, the clustering technique holds significant importance for the effective categorization and organization of extensive datasets. However, pinpointing the ideal number of clusters and handling high-dimensional data can be challenging. To tackle these issues, several strategies have been suggested, such as a consensus clustering ensemble that yields more significant outcomes compared to individual models. Another valuable technique for cluster analysis is Bayesian mixture modelling, which is known for its adaptability in determining cluster numbers.

Keywords: cluster analysis ; Bayesian mixture modelling ; consensus clustering

1. Introduction

Clustering is a key technique in unsupervised learning and is employed across various domains such as computer vision, natural language processing, and bioinformatics. Its primary objective is to assemble related items and disclose hidden patterns within data. Confronting complex datasets, however, can prove challenging, as conventional clustering approaches may not be effective. In response to this issue, Bayesian nonparametric methods have gained popularity in recent years as a potent means of organising large datasets. These approaches offer a versatile and potent solution for managing the data's unpredictability and complexity, making them a crucial tool in the field of clustering. Clustering is crucial in the fields of information science and big-data management for organizing and handling huge volumes of data. In recent years, exponential data proliferation has increased the demand for efficient and effective solutions to handle, manage, and analyse enormous data volumes. Clustering can accomplish this by grouping comparable data points together, hence lowering the dataset's size and making it simpler to examine. Apart from traditional techniques, there are much more promising ones. The product partition model (PPM) is one of the most widely used Bayesian nonparametric clustering algorithms. PPMs are a class of models that classify data into clusters and assign a set of parameters to each cluster. They use a prior over the parameters to draw conclusions about the clusters. Despite the efficacy of PPMs, a single clustering solution may not be enough for complicated datasets, resulting in the development of consensus clustering. Consensus clustering is a kind of ensemble clustering that produces a final grouping by combining the results of numerous clustering methods [1][2].

The motivation behind this work lies in addressing the challenges associated with clustering complex datasets, which is crucial for efficient big-data management and analysis. The determination of the number of clusters and handling of high-dimensional data are significant challenges that arise while dealing with these complex datasets.

2. Cluster Analysis

Cluster analysis has been utilised extensively in numerous disciplines to identify patterns and structures within data. Caruso et al. [3] applied cluster analysis to an actual mixed-type dataset and reported their findings. Meanwhile, Absalom et al. [4] provided a comprehensive survey of clustering algorithms, discussing the state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Jiang et al. [5] conducted a survey of cluster analysis for gene expression data. Furthermore, Huang et al. [6] proposed a locally weighted ensemble clustering method that assigns weights to individual partitions based on local information. These studies demonstrate the diversity of clustering methods and their applications, emphasizing the importance of choosing the appropriate method for specific datasets.

Consensus clustering utilises W runs of a base model or learner (such as K-means clustering) and combines the W suggested partitions into a consensus matrix, where the (i,j)-th entries reflect the percentage of model runs in which the ith and jth individuals co-cluster. This ratio indicates the degree of confidence in the co-clustering of any two elements. Moreover, ensembles may reduce computational execution time. This occurs because individual learners may be weaker (and hence consume less of the available data or stop before complete convergence), and the learners in the vast

majority of ensemble techniques are independent of one another, enabling the use of a parallel environment for each of the faster model runs [7].

Bayesian clustering is a popular machine-learning technique for grouping data points into clusters based on their probability distributions. Hidden Markov models (HMM) [8] have been used to model the underlying probabilistic structure of data in Bayesian clustering. Accelerating hyperparameters via Bayesian optimizations can also help in building automated machine learning (AutoML) schemes [9], while such optimizations can also be applied in Tiny Machine Learning (TinyML) environments wherein devices can be trained to fulfil ML tasks [10]. Ensemble Bayesian Clustering [11] is a variation of Bayesian clustering that combines multiple models to produce more robust results, while cluster analysis [12] extends traditional clustering methods by considering the uncertainty in the data, which leads to more accurate results.

Traditional clustering algorithms require a preset selection of the number of clusters *K*, which can be challenging as it plagues many investigations, with researchers often depending on certain rules to choose a final model. Various selections of *K* are compared, for instance, using an evaluation metric for *K*. Techniques for selecting *K* using the consensus matrix are offered in [13]; however, this implies that any uncertainty over *K* is not reflected in the final clustering, and each model run utilises the same, fixed number of clusters. An alternative clustering technique incorporates cluster analysis within a statistical framework [14], which implies that models may be formally compared and issues such as choosing *K* can be represented as a model-selection problem using relevant tools.

In recent years, various clustering techniques have been developed to address the challenges associated with traditional clustering methods. Locally weighted ensemble clustering [6] leverages the advantages of ensemble clustering while accounting for the local structure of the data, leading to more accurate and robust results. Consensus clustering, a type of ensemble clustering, combines multiple runs of a base model into a consensus matrix to increase confidence in co-clustering [13]. Enhanced ensemble clustering via fast propagation of cluster-wise similarities [15][16] improves the efficiency and effectiveness of clustering by propagating cluster-wise similarities more rapidly. Real-world applications of these clustering techniques can be found in various domains, such as gene expression analysis, cell classification in flow cytometry experiments, and protein localization estimation [17][18][19].

Recent advancements in ensemble clustering have addressed various challenges posed by high-dimensional data and complex structures. Yan and Liu [20] proposed a consensus clustering approach specifically designed for high-dimensional data, while Niu et al. [21] developed a multi-view ensemble clustering approach using a joint affinity matrix to improve the quality of clustering. Huang et al. [22] introduced an ensemble hierarchical clustering algorithm that considers merits at both cluster and partition levels. In addition, Zhou et al. [23] presented a clustering ensemble method based on structured hypergraph learning, and Zamora and Sublime [24] proposed an ensemble and multi-view clustering method based on Kolmogorov complexity. Huang et al. [25] tackled the challenge of high-dimensional data by developing a multidiversified ensemble clustering approach, focusing on various aspects such as subspaces, metrics, and more. Huang et al. [26] also proposed an ultra-scalable spectral clustering and ensemble clustering technique. Wang et al. [27] developed a Markov clustering ensemble method, and Huang et al. [28] presented a fast multi-view clustering approach via ensembles for scalability, superiority, and simplicity. These studies showcase the diverse range of ensemble clustering techniques developed to address complex data challenges and improve the performance of clustering algorithms.

Clustering ensemble techniques have been developed and applied across various domains, addressing the challenges and limitations of traditional clustering methods. Nie et al. [29] concentrated on the analysis of scRNA-seq data, discussing the methods, applications, and difficulties associated with ensemble clustering in this field. Boongoen and Iam-On [30] presented an exhaustive review of cluster ensembles, highlighting recent extensions and applications. Troyanovsky [31] examined the ensemble of specialised cadherin clusters in adherens junctions, demonstrating the versatility of ensemble clustering methods. Zhang and Zhu [32] introduced Ensemble Clustering based on Bayesian Network (ECBN) inference for single-cell RNA-seq data analysis, offering a novel method for addressing the difficulties inherent to this data format. Hu et al. [33] proposed an ultra-scalable ensemble clustering method for cell-type recognition using scRNA-seq data of Alzheimer's disease. Bian et al. [34] developed an ensemble consensus clustering method, scEFSC, for accurate single-cell RNA-seq data analysis based on multiple feature selections. Wang and Pan [35] introduced a semi-supervised consensus clustering method for gene expression data analysis, while Yu et al. [36] explored knowledge-based cluster ensemble approaches for cancer discovery from biomolecular data. Finally, Yang et al. [37] proposed a consensus clustering approach using a constrained self-organizing map and an improved Cop-Kmeans ensemble for intelligent decision support systems, showcasing the broad applicability of ensemble clustering techniques in various fields.

Bayesian mixture models, with their adaptable densities, are highly attractive for data analysis across various types. The number of clusters *K* can be inferred directly from the data as a random variable, resulting in joint modelling of *K* and the

clustering process [38][39][40][41][42][43]. Inference of the number of clusters can be achieved through methods such as the Dirichlet process [38], finite mixture models [39][40], or over-fitting mixture models [41]. These models have found success in a wide range of biological applications, including gene expression profiles [17], cell classification in flow cytometry experiments [18][44] and scRNAseq experiments [45], as well as protein localization estimation [19]. Bayesian mixture models can also be extended to jointly cluster multiple datasets [46][47].

MCMC techniques are the most-used method for executing Bayesian inference, and they are used to build a chain of clusterings. The convergence of the chain is evaluated to see if its behaviour conforms to the asymptotic theory predicted. However, despite the ergodicity of MCMC approaches, individual chains often fail to investigate the complete support of the posterior distribution and have lengthy runtimes.

References

1. Coleman, S.; Kirk, P.D.; Wallace, C. Consensus clustering for Bayesian mixture models. BMC Bioinform. 2022, 23, 1–21.
2. Lock, E.F.; Dunson, D.B. Bayesian consensus clustering. Bioinformatics 2013, 29, 2610–2616.
3. Caruso, G.; Gattone, S.A.; Balzanella, A.; Di Battista, T. Cluster Analysis: An Application to a Real Mixed-Type Data Set. In Models and Theories in Social Systems; Springer International Publishing: Cham, Switzerland, 2019; pp. 525–533.
4. Ezugwu, A.E.; Ikotun, A.M.; Oyelade, O.O.; Abualigah, L.; Agushaka, J.O.; Eke, C.I.; Akinyelu, A.A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Eng. Appl. Artif. Intell. 2022, 110, 104743.
5. Jiang, D.; Tang, C.; Zhang, A. Cluster analysis for gene expression data: A survey. IEEE Trans. Knowl. Data Eng. 2004, 16, 1370–1386.
6. Huang, D.; Wang, C.D.; Lai, J.H. Locally weighted ensemble clustering. IEEE Trans. Cybern. 2017, 48, 1460–1473.
7. Ghaemi, R.; Sulaiman, M.N.; Ibrahim, H.; Mustapha, N. A survey: Clustering ensembles techniques. Int. J. Comput. Inf. Eng. 2009, 3, 365–374.
8. Can, C.E.; Ergun, G.; Soyer, R. Bayesian analysis of proportions via a hidden Markov model. Methodol. Comput. Appl. Probab. 2022, 24, 3121–3139.
9. Karras, A.; Karras, C.; Schizas, N.; Avlonitis, M.; Sioutas, S. AutoML with Bayesian Optimizations for Big Data Management. Information 2023, 14, 223.
10. Schizas, N.; Karras, A.; Karras, C.; Sioutas, S. TinyML for Ultra-Low Power AI and Large Scale IoT Deployments: A Systematic Review. Future Internet 2022, 14, 363.
11. Zhu, Z.; Xu, M.; Ke, J.; Yang, H.; Chen, X.M. A Bayesian clustering ensemble Gaussian process model for network-wide traffic flow clustering and prediction. Transp. Res. Part Emerg. Technol. 2023, 148, 104032.
12. Greve, J.; Grün, B.; Malsiner-Walli, G.; Frühwirth-Schnatter, S. Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. Aust. N. Z. J. Stat. 2022, 64, 205–229.
13. Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. 2003, 52, 91–118.
14. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 2002, 97, 611–631.
15. Huang, D.; Wang, C.D.; Peng, H.; Lai, J.; Kwoh, C.K. Enhanced ensemble clustering via fast propagation of cluster-wise similarities. IEEE Trans. Syst. Man Cybern. Syst. 2018, 51, 508–520.
16. Cai, X.; Huang, D. Link-Based Consensus Clustering with Random Walk Propagation. In Proceedings of the Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, 8–12 December 2021; Proceedings, Part V 28. Springer: Berlin/Heidelberg, Germany, 2021; pp. 693–700.
17. Medvedovic, M.; Sivaganesan, S. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 2002, 18, 1194–1206.
18. Chan, C.; Feng, F.; Ottinger, J.; Foster, D.; West, M.; Kepler, T.B. Statistical mixture modeling for cell subtype identification in flow cytometry. Cytom. Part A J. Int. Soc. Anal. Cytol. 2008, 73, 693–701.

19. Crook, O.M.; Mulvey, C.M.; Kirk, P.D.; Lilley, K.S.; Gatto, L. A Bayesian mixture modelling approach for spatial proteomics. PLoS Comput. Biol. 2018, 14, e1006516.
20. Yan, J.; Liu, W. An ensemble clustering approach (consensus clustering) for high-dimensional data. Secur. Commun. Netw. 2022, 2022, 5629710.
21. Niu, X.; Zhang, C.; Zhao, X.; Hu, L.; Zhang, J. A multi-view ensemble clustering approach using joint affinity matrix. Expert Syst. Appl. 2023, 216, 119484.
22. Huang, Q.; Gao, R.; Akhavan, H. An ensemble hierarchical clustering algorithm based on merits at cluster and partition levels. Pattern Recognit. 2023, 136, 109255.
23. Zhou, P.; Wang, X.; Du, L.; Li, X. Clustering ensemble via structured hypergraph learning. Inf. Fusion 2022, 78, 171–179.
24. Zamora, J.; Sublime, J. An Ensemble and Multi-View Clustering Method Based on Kolmogorov Complexity. Entropy 2023, 25, 371.
25. Huang, D.; Wang, C.D.; Lai, J.H.; Kwoh, C.K. Toward Multidiversified Ensemble Clustering of High-Dimensional Data: From Subspaces to Metrics and Beyond. IEEE Trans. Cybern. 2022, 52, 12231–12244.
26. Huang, D.; Wang, C.D.; Wu, J.S.; Lai, J.H.; Kwoh, C.K. Ultra-Scalable Spectral Clustering and Ensemble Clustering. IEEE Trans. Knowl. Data Eng. 2020, 32, 1212–1226.
27. Wang, L.; Luo, J.; Wang, H.; Li, T. Markov clustering ensemble. Knowl.-Based Syst. 2022, 251, 109196.
28. Huang, D.; Wang, C.D.; Lai, J.H. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. IEEE Trans. Knowl. Data Eng. 2023.
29. Nie, X.; Qin, D.; Zhou, X.; Duo, H.; Hao, Y.; Li, B.; Liang, G. Clustering ensemble in scRNA-seq data analysis: Methods, applications and challenges. Comput. Biol. Med. 2023, 106939.
30. Boongoen, T.; Iam-On, N. Cluster ensembles: A survey of approaches with recent extensions and applications. Comput. Sci. Rev. 2018, 28, 1–25.
31. Troyanovsky, S.M. Adherens junction: The ensemble of specialized cadherin clusters. Trends Cell Biol. 2022, 33, 374–387.
32. Zhang, D.; Zhu, Y. ECBN: Ensemble Clustering based on Bayesian Network inference for Single-cell RNA-seq Data. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 5884–5888.
33. Hu, L.; Zhou, J.; Qiu, Y.; Li, X. An Ultra-Scalable Ensemble Clustering Method for Cell Type Recognition Based on scRNA-seq Data of Alzheimer's Disease. In Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers, Dalian, China, 14–16 April 2022; pp. 275–280.
34. Bian, C.; Wang, X.; Su, Y.; Wang, Y.; Wong, K.C.; Li, X. scEFSC: Accurate single-cell RNA-seq data analysis via ensemble consensus clustering based on multiple feature selections. Comput. Struct. Biotechnol. J. 2022, 20, 2181–2197.
35. Wang, Y.; Pan, Y. Semi-supervised consensus clustering for gene expression data analysis. BioData Min. 2014, 7, 1–13.
36. Yu, Z.; Wongb, H.S.; You, J.; Yang, Q.; Liao, H. Knowledge based cluster ensemble for cancer discovery from biomolecular data. IEEE Trans. Nanobiosci. 2011, 10, 76–85.
37. Yang, Y.; Tan, W.; Li, T.; Ruan, D. Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems. Knowl.-Based Syst. 2012, 32, 101–115.
38. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. Ann. Stat. 1973, 1, 209–230.
39. Miller, J.W.; Harrison, M.T. Mixture models with a prior on the number of components. J. Am. Stat. Assoc. 2018, 113, 340–356.
40. Richardson, S.; Green, P.J. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 1997, 59, 731–792.
41. Rousseau, J.; Mengersen, K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. J. R. Stat. Soc. Ser. B Stat. Methodol. 2011, 73, 689–710.
42. Law, M.; Jain, A.; Figueiredo, M. Feature selection in mixture-based clustering. In Advances in Neural Information Processing Systems; MIT Press: Cambridge, MA, USA, 2002; Volume 15.
43. Scrucca, L.; Fop, M.; Murphy, T.B.; Raftery, A.E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. 2016, 8, 289.

44. Hejblum, B.P.; Alkhassim, C.; Gottardo, R.; Caron, F.; Thiébaut, R. Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. Ann. Appl. Stat. 2019, 13, 638–660.
45. Prabhakaran, S.; Azizi, E.; Carr, A.; Pe’er, D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; PMLR: Baltimore, MD, USA, 2016; pp. 1070–1079.
46. Gabasova, E.; Reid, J.; Wernisch, L. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS Comput. Biol. 2017, 13, e1005781.
47. Kirk, P.; Griffin, J.E.; Savage, R.S.; Ghahramani, Z.; Wild, D.L. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 2012, 28, 3290–3297.