Research Analysis on Clustering Ensemble Methods

Sonia Yadav[1], Sachin Sharma[2]

[1] Research Scholar, School of Computer Applications, ManavRachna International Institute of Research and Studies (MRIIRS), Faridabad, India.
Department of Computer Science, Deshbandhu College, University of Delhi, New Delhi.
[2] Associate Professor, School of Computer Applications, ManavRachna International Institute of Research and Studies (MRIIRS), Faridabad, India.

Received: 18.04.2024 Revised : 27.05.2024 Accepted: 29.05.2024

ABSTRACT
Clustering ensemble methods aim to improve the robustness, stability, and accuracy of clustering results by combining multiple individual clustering solutions. The idea is to leverage diverse clustering algorithms or variations of the same algorithm to capture different aspects of the underlying data structure. Ensemble methods can be particularly effective when dealing with complex datasets, noisy data, or when individual clustering algorithms are sensitive to specific initialization conditions.

Keywords: Clustering ensemble methods, Clustering algorithms, Dataset, Supervised and unsupervised.

1. INTRODUCTION

Ensemble clustering is the process of combining the multiple clustering results of a set of objects into a single improved clustering. It is sometimes referred to as the Consensus solution or Clustering Aggregation. In recent years, various studies have been conducted to develop clustering ensemble methods inspired by the success of the ensemble method in the supervised learning field. However, compared to the research on classification ensemble methods, building a clustering ensemble is not straightforward, and further work is required in this field. There are several reasons that make the task of building a clustering ensemble more challenging than that of classification. One is that clustering is unsupervised learning in which the data are unlabelled, so there is no prior knowledge with which the algorithm can discover the true cluster structure, and there is no "ground truth" to validate the clustering result. Moreover, no cross-validation technique can be carried out to tune the clustering algorithm's parameters, thus there are no guidelines with which the user can select the most appropriate clustering algorithm for a given dataset. Another challenge is that the number of clusters produced may differ among the generated solutions by different clustering algorithms. In addition, the number of clusters in the final solution is unknown in advance. The final solution is obtained by accessing a set of base solutions, which in fact are cluster labels, and not the original data used. Ghosh and Acharya pointed out that there are several motivations for using clustering ensembles, and that these are much broader than those for using classification ensemble, where the main motivation of the latter is to improve the classification accuracy.
These reasons include:

- To improve the quality of the clustering results compared to those produced by single clustering algorithms.
- To reuse existing clustering (knowledge reuse): in some applications a variety of partitions may exist, so they can be combined to obtain a final clustering result. This delivers a more consolidated clustering result; several examples are provided in.
- To generate robust clustering results across different types of datasets. It is widely known that the popular clustering algorithms often fail to produce a good clustering result when the data do not match with their assumptions. Among these objectives, the first point is the most widely accepted one. The cluster quality is usually measured with a numerical measurement to assess different aspects of cluster validation.

2. Motivation

Clustering ensemble methods are motivated by the recognition that different clustering algorithms or parameter settings may yield varying results on the same dataset.
They aim to overcome the limitations of individual algorithms by combining their outputs to achieve more reliable and robust clustering solutions.

Diversity
Ensemble methods benefit from the diversity of individual clustering solutions. Diversity is essential because it allows the ensemble to capture different perspectives on the data and reduce the impact of algorithmic biases or sensitivity to initialization.

Aggregation Techniques
Clustering ensemble methods use various aggregation techniques to combine individual clustering. Common approaches include voting mechanisms, averaging, or more sophisticated meta-models that learn to combine the outputs of base clustering models.
Bagging involves generating multiple bootstrap samples of the original dataset, performing clustering on each sample independently, and aggregating the results.
Boosting focuses on improving the performance of weak clustering algorithms by giving more weight to misclassified instances in successive iterations.
Voting-based methods combine the results of multiple clustering algorithms through a voting mechanism.
Stacking uses a meta-model to combine the predictions of multiple base clustering models.
Hierarchical clustering ensemble combines results from different hierarchical clustering algorithms or applies hierarchical clustering to the outputs of other clustering algorithms.
Self-Organizing Maps (SOM) ensemble combines the results of multiple SOMs to enhance clustering performance.
Consensus clustering generates multiple clustering solutions and finds a consensus among them to identify stable clusters.

3. Clustering Ensemble Methods
Bagging (Bootstrap Aggregating)
Description: Bagging involves creating multiple bootstrap samples (random samples with replacement) from the original dataset. Each sample is used to independently perform clustering. The final clustering result is obtained by aggregating the individual results, often through a majority vote or averaging.
Example: Iterative Bagging of K-Means (IBK-Means) involves applying K-Means clustering to different bootstrap samples iteratively.

Boosting
Description: Boosting methods aim to improve the performance of weak clustering algorithms by iteratively assigning higher weights to misclassified instances. The idea is to focus on instances that are difficult to cluster correctly.
Example: AdaBoost clustering assigns higher weights to misclassified data points in each iteration, encouraging the algorithm to give more attention to challenging instances.

Voting-Based Methods
Description: In voting-based ensemble methods, the results of multiple clustering algorithms are combined through a voting mechanism. This can involve a majority vote or a weighted vote based on the confidence of each algorithm's clustering decisions.
Example: K-Modes Voting combines the results of multiple K-Modes clustering algorithms through a voting process.

Stacking
Description: Stacking involves training multiple base clustering models, and then using a meta-model to combine their predictions. The meta-model is trained on the outputs of the base models to improve overall clustering performance.
Example: Stacked Ensemble Clustering may use algorithms like K-Means, DBSCAN, and Agglomerative Clustering as base models, and a meta-model like Random Forest to combine their outputs.

Hierarchical Clustering Ensemble
Description: Hierarchical Clustering Ensemble involves combining the results of different hierarchical clustering algorithms or applying hierarchical clustering to the outputs of other clustering algorithms.
Example: Hierarchical Clustering Ensemble (HCE) might utilize results from algorithms like K-Means, DBSCAN, and Agglomerative Clustering, and then perform hierarchical clustering on the combined results.

Self-Organizing Maps (SOM) Ensemble
Description: SOM Ensemble combines the results of multiple Self-Organizing Maps, which are neural network-based clustering algorithms. The ensemble approach helps capture different aspects of the data.
Example: Train multiple SOMs on different subsets of the data and combine their cluster assignments.

Consensus Clustering
Description: Consensus Clustering involves generating multiple clustering solutions using various algorithms or parameters and finding a consensus among them. This helps identify robust clusters that are stable across different runs.
Example: Consensus K-Means generates multiple K-Means clusterings with different initializations and combines them to find a stable consensus.
These ensemble methods offer diverse strategies for combining the results of individual clustering algorithms, enhancing the robustness and accuracy of the overall clustering process. The choice of a specific method depends on the characteristics of the data and the goals of the clustering task.

4. Challenges in Clustering Ensemble
Unsupervised Nature: Clustering is unsupervised, making it challenging as there's no labeled data to guide the algorithm.
Lack of Ground Truth: Without labeled data, there's no "ground truth" to validate the clustering results.
No Parameter Tuning via Cross-Validation: The absence of labeled data also means no cross-validation for parameter tuning.
Differing Number of Clusters: Different clustering algorithms may produce varying numbers of clusters.
Unknown Number of Clusters in Advance: The final number of clusters is unknown before clustering.

5. Advantages of Clustering Ensemble
Improved Robustness: Ensembles are less sensitive to variations in data and initialization conditions.
Increased Stability: Consistent clustering results across multiple runs or algorithms enhances stability.
Enhanced Accuracy: Combining diverse perspectives can lead to more accurate representations of complex data structures.

6. CONCLUSION
In summary, clustering ensemble methods provide a powerful approach to enhance the quality and reliability of clustering results by leveraging the strengths of multiple algorithms. The choice of a specific ensemble method depends on the characteristics of the data and the goals of the clustering task. Ensemble clustering addresses the challenges posed by unsupervised learning, offering a way to improve clustering quality, reuse existing knowledge, and enhance robustness across different datasets. Further research and development in this field are essential to refine and expand the applicability of clustering ensemble methods.

REFERENCES
[1] Shi Y , Yu Z , Chen C L P, You J, Wong H S, Wang Y D, Zhang J. Transfer clustering ensemble selection. IEEE Transactions on Cybernetics, 2018, PP(99): 1–14
[2] Nozza D, Fersini E, Messina E. Deep learning and ensemble methods for domain adaptation. In: Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence (ICTAI). 2016, 184– 189 180.
[3] Liu X, Liu Z, Wang G, Cai Z, Zhang H. Ensemble transfer learning algorithm. IEEE Access, 2018, 6: 2389–2396 181. Brys T, Harutyunyan A, Vrancx P, Nowé A, Taylor M E. Multiobjectivization and ensembles of shapings in reinforcement learning.
[4] Neurocomputing, 2017, 263: 48–59 182. Chen X L, Cao L, Li C X, Xu Z X, Lai J. Ensemble network architecture for deep reinforcement learning. Mathematical Problems in Engineering, 2018, 2018: 1–6
[5] Yin Z, Zhao M, Wang Y, Yang J, Zhang J. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer Methods and Programs in Biomedicine, 2017, 140: 93–110 176.
[6] Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE Journal of Biomedical and Health Informatics, 2017, 21(1): 31–40
[7] Wei S, Li Z, Zhang C. Combined constraint-based with metric-based in semi-supervised clustering ensemble. International Journal of Machine Learning and Cybernetics, 2018, 9(7): 1085–1100

[8] Yu Z, Zhang Y, Chen C L P, You J, Wong H S, Dai D, Wu S, Zhang J. Multiobjectivesemisupervised classifier ensemble. IEEE Transactions on Cybernetics, 2019, 49(6): 2280–2293
[9] Yu Z, Zhang Y, You J, Chen C P, Wong H S, Han G, Zhang J. Adaptive semi-supervised classifier ensemble for high dimensional data clasfication. IEEE Transactions on Cybernetics, 2019, 49(2): 366–379
[10] Li Q, Li G, Niu W, Cao Y, Chang L, Tan J, Guo L. Boosting imbalanced data learning with wiener process oversampling. Frontiers of Computer Science, 2017, 11(5): 836–851
[11] Zhou Z H, Feng J. Deep forest: towards an alternative to deep neural networks. 2017, arXiv preprint arXiv:1702.08835
[12] Soares R G F, Chen H, Yao X. A cluster-based semi-supervised ensemble for multiclass classification. IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1(6): 408–420
[13] Alves M, Bazzan A L C, Recamonde-Mendoza M. Social-training: ensemble learning with voting aggregation for semi-supervised classification tasks. In: Proceedings of 2017 Brazilian Conference on Intelligent Systems (BRACIS). 2017, 7–12
[14] Wang Y, Li T. Improving semi-supervised co-forest algorithm in evolving data streams. Applied Intelligence, 2018, 48(10): 3248– 3262
[15] Abdelgayed T S, Morsi W G, Sidhu T S. Fault detection and classification based on co-training of semi-supervised machine learning. IEEE Transactions on Industrial Electronics, 2018, 65(2): 1595–1605
[16] Seyed Saeed Hamidi, Ebrahim Akbari, HomayunMotameni, Consensus clustering algorithm based on the automatic partitioning similarity graph, Data & Knowledge Engineering, Volume 124,2019,101754,ISSN 0169-023X,https://doi.org/10.1016/j.datak.2019.101754.
[17] QinghuaGu, Yan Wang, Peipei Wang, Xuexian Li, Lu Chen, Neal N. Xiong, Di Liu, An improved weighted ensemble clustering based on two-tier uncertainty measurement, Expert Systems with Applications,Volume 238, Part A,2024,121672,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2023.121672.
[18] Guangyan Ji, Gui-Fu Lu, Bing Cai,Scalable incomplete multi-view clustering via tensor Schatten p-norm and tensorized bipartite graph,Engineering Applications of Artificial Intelligence,Volume 123, Part B,2023,106379,ISSN 0952-1976, https://doi.org/10.1016/j.engappai.
[19] BaichengLyu, Da Li, Wenhua Wu, Hui Li,A ReCon-BCALoD clustering algorithm for field monitoring data of marine structures,Applied Ocean Research,Volume 138,2023,103644,ISSN 0141-1187,https://doi.org/10.1016/j.apor.2023.103644.
[20] Ignacio Pérez-Martínez, MaríaMartínez-Rojas, Jose Manuel Soto-Hidalgo,A methodology for urban planning generation: A novel approach based on generative design, Engineering Applications of Artificial Intelligence, Volume 124,2023, 106609, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2023.106609.
[21] Yubo Wang, Shelesh Krishna Saraswat, IrajElyasiKomari, Big data analysis using a parallel ensemble clustering architecture and an unsupervised feature selection approach, Journal of King Saud University - Computer and Information Sciences, Volume 35, Issue 1,2023, Pages 270-282,ISSN 1319-1578,https://doi.org/10.1016/j.jksuci.2022.11.016
[22] K. Berahmand, M. Mohammadi, A. Faroughi, R.P. MohammadianiA novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrixClust. Comput., 25 (2) (2022), pp. 869-888
[23] C. Bian, X. Wang, Y. Su, Y. Wang, K.C. Wong, X. LiscEFSC: Accurate single-cell RNA-seq data analysis via ensemble consensus clustering based on multiple feature selectionsComput. Struct. Biotechnol. J., 20 (2022), pp. 2181-2197
[24] Yadav, Sonia & Sharma, Sachin. (2024). Study Of Existing Methods & Techniques Of K-Means Clustering. Educational Administration: Theory and Practice. 30. 10.53555/kuey.v30i4.1755.
[25] Yadav, Sonia. (2023). Egyptian Vulture Optimization Algorithm-Based Data Hiding Technique. DongbeiDaxueXuebao/Journal of Northeastern University. 25.
[26] Yadav, Sonia & Sharma, Sachin. (2023). Article Info Page Number. 72. 593-599.
[27] Niranjan, Keshav& Yadav, Sonia. (2021). Natural Language Interface for Data Base: A Case of Hindi Language. International Journal of Computer Trends and Technology. 8. 92-94.
[28] Niranjan, Keshav& Yadav, Sonia & Lal, Heera. (2021). Intervention of IT in Indian health care system.