HDBSCAN: Advantages and Disadvantages in BERTopic

44251017 Huang Jiahui

1 Basic Idea and Algorithm of HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that identifies clusters based on the density distribution of data points. Unlike traditional algorithms such as k-means, which require specifying the number of clusters in advance, HDBSCAN automatically determines clusters by analyzing the data's intrinsic structure[1][2][3].

2 Algorithm

- **Calculating Core Distance:** For each data point, the core distance is calculated based on its neighbors. The core distance is the radius of the smallest circle encompassing at least $min\_samples$ points around a given point[1].
- **Mutual Reachability Distance:** Next, the mutual reachability distance between two points is defined as the maximum value among their core distances and their original distance. This approach ensures that density changes are smoothly represented in the clustering process[1][4].
- **Minimum Spanning Tree (MST) Construction:** Using mutual reachability distances, a Minimum Spanning Tree (MST) is constructed. MST connects all points with minimal total distance, reflecting the underlying data structure and highlighting density variations[5][6].

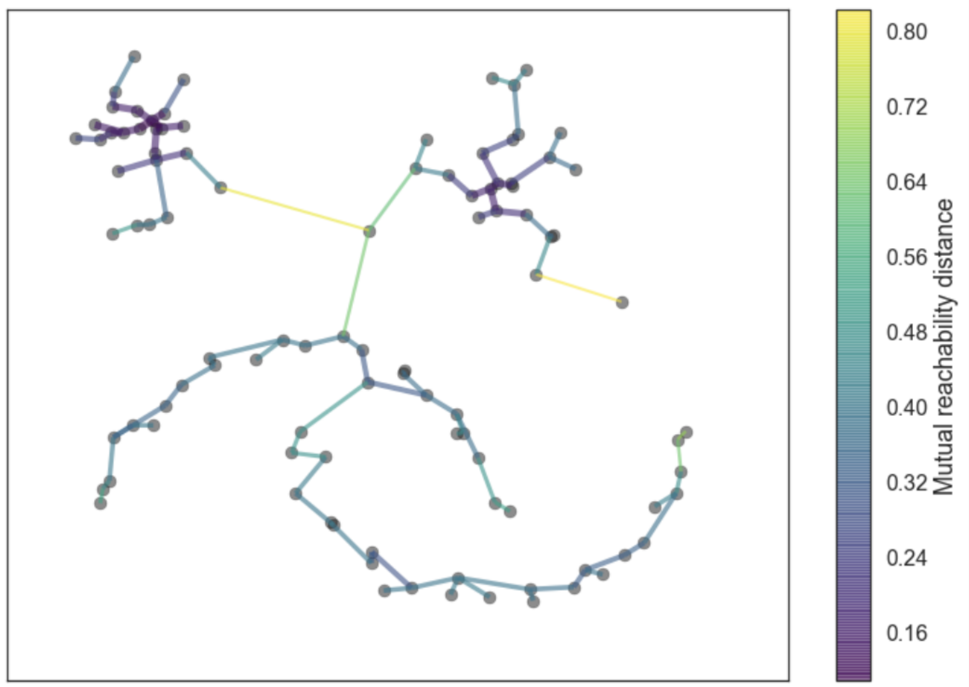

Fig.1 Build the minimum spanning tree

- **Creating the Cluster Hierarchy:** By sorting edges in the MST from smallest to largest mutual reachability distance, clusters gradually form and split, creating a hierarchical structure.
- **Condensing the Cluster Tree:** The hierarchical cluster tree is simplified by removing unstable or insignificant clusters based on parameters like $min\_cluster\_size$.
- **Extracting Clusters:** Finally, the algorithm selects clusters based on their stability (how persistent clusters are across various density thresholds) and identifies noise points—those not fitting well into any cluster.

3 Advantages and Disadvantages of HDBSCAN in BERTopic

3.1 Advantages

- **Automatic Cluster Determination:** HDBSCAN does not require specifying the number of clusters beforehand, which is beneficial for topic modeling tasks where the appropriate number of topics is unknown[7].
- **Robust to Noise:** It effectively identifies noise, thus filtering irrelevant or outlier documents and providing cleaner topic clusters.
- **Handles Variable Density:** Unlike DBSCAN, HDBSCAN can detect clusters of varying densities simultaneously. This capability is essential in natural language data, where topic clusters might vary significantly in size and density.
- **Hierarchical Structure:** The hierarchical nature of HDBSCAN allows exploration of topic clusters at various granularity levels, providing deeper insights into topic relations and subtopic structures.

3.2 Disadvantages

- **Computational Complexity:** HDBSCAN can be computationally intensive, especially for high-dimensional embedding vectors common in BERTopic. As data size and dimensionality increase, the clustering process may become prohibitively slow.
- **Parameter Sensitivity:** Parameters such as min_cluster_size and min_samples significantly affect the results. Incorrect parameter settings can either excessively fragment clusters or merge distinct topics.
- **Interpretation Complexity:** The hierarchical output can be complicated to interpret directly in the context of topic modeling, requiring additional tools or steps to simplify and present results meaningfully.

4 Conclusion

HDBSCAN is an effective clustering method for topic modeling tasks such as BERTopic, particularly due to its adaptability in automatically determining the number of clusters and managing noise points. Its hierarchical structure offers flexibility in exploring data. Nevertheless, careful consideration of parameter settings and computational resources is crucial to fully leverage its advantages.

References

[1] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[2] Geoffrey Stewart and Mahmood Al-Khassaweneh. An implementation of the hdbscan* clustering algorithm. *Applied Sciences*, 12(5):2405, 2022.

[3] Michael Strobl, Jörg Sander, Ricardo JGB Campello, and Osmar Zaïane. Model-based clustering with hdbscan. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 364–379. Springer, 2020.

[4] David Hanny and Bernd Resch. Clustering-based joint topic-sentiment modeling of social media data: a neural networks approach. *Information*, 15(4):200, 2024.

[5] Hongchen Li, Xinyi Lu, Yujia Wu, and Jie Luo. Research on a data mining algorithm based on bertopic for medication rules in traditional chinese medicine prescriptions. *Medicine Advances*, 1(4):353–360, 2023.

[6] Jiang Ningpeng, Han Tian, Wang Haibo, Xu Ruzhi, and Ma Shiyu. A study on structured text parsing for policies based on bertopic. In *2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 6, pages 16–22. IEEE, 2024.

[7] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.