Adaptive Graphs for Multi-View Subspace Clustering

Subjects: Computer Science, Interdisciplinary Applications
Contributor: Qiliang Liu, Weihua Huan, Min Deng

Clustering of multi-source geospatial big data provides opportunities to comprehensively describe urban structures. Most existing studies focus only on the clustering of a single type of geospatial big data, which leads to biased results. Although multi-view subspace clustering methods are advantageous for fusing multi-source geospatial big data, exploiting a robust shared subspace in high-dimensional, non-uniform, and noisy geospatial big data remains a challenge.

Keywords: multi-view subspace clustering ; geospatial big data ; shared nearest neighbor graph

1. Introduction

Multi-source geospatial big data have become increasingly available in the current era of big data, such as taxi GPS trajectories [1], smart card transactions [2], mobile phone data [3], social media check-in records [4], and points of interests (POIs) [5]. Geospatial big data provides a new opportunity for understanding the “human-earth” relationship [6]. Clustering geospatial big data are vital for describing urban structures and understanding the organization of cities [7]. For example, remote sensing techniques have been widely used for uncovering urban land use information based on physical characteristics of ground components (e.g., spectral, shape, and texture) [8]; however, remote sensing techniques are hard to capture the socioeconomic attributes and human dynamics that are highly related to urban land use [3]. In contrast, clustering of human mobility data can help understand urban land use information from the perspective of social function which is an important complement of remote sensing [6]. Clustering of geospatial big data are also useful for identifying urban functional structures and human activity patterns, which are useful for human-centric urban planning [9][10][11]. For example, the actual functions of a region may be inconsistent with the original zoning scheme designed by urban planners [12]. Clusters discovered from geospatial big data can reveal the urban function zones naturally formulated according to human activities, which may provide useful calibration for urban planners [9]. Clusters discovered from social media check-in records are also useful for identifying emergency events in a city, which are helpful for maintaining public safety [4].

Although clustering of geospatial big data has received attention in recent years, most existing studies focus on a single type of geospatial big data [11][13]. Owing to the bias of each type of geospatial big data, the clustering results obtained from single-source geospatial big data cannot provide a comprehensive view of urban structures [14]. A few studies have used a weighted average strategy to fuse multi-source geospatial big data [15][16]. Multi-source geospatial big data usually reflect different or overlapping dimensions of human activities. Without considering the shared and complementary information among different types of geospatial big data, the weighted average strategy may introduce unpredictable errors [17]. Multi-view subspace clustering has the potential to fuse the underlying complementary information of multi-source geospatial big data [18][19]; however, high-dimensional, non-uniform, and noisy geospatial big data bring two challenges [20][21][22][23]: (1) the quality of the low-dimensional subspace is substantially influenced by the redundant features and noise in the original data; and (2) neighboring relationships of data points in high-dimensional and non-uniform original data space are difficult to preserve in a low-dimensional subspace. Therefore, existing multi-view subspace clustering methods are highly likely to generate an inaccurate subspace, which degrades the clustering performance. To overcome the above challenges, this developed a method with adaptive graphs to constrain multi-view subspace clustering of geospatial big data from multiple sources (agc2msc).

2. Multi-source Geospatial Big Data

Most existing studies mainly focus on the clustering of a single type of geospatial big data, e.g., taxi GPS trajectories [1], social media check-in records [4], POIs [13], and mobile phone data [24]. After extracting clustering features from a certain type of geospatial big data, traditional clustering methods such as k-means [25], spectral clustering [26], and DBSCAN [27] are used to identify clusters. To consider the dynamic characteristic of geospatial big data, some online and incremental clustering methods are also currently available [4]; these methods are useful for understanding the organizations of cities from the perspective of social functions [7]. Despite these fruitful results, the bias of a single type of geospatial big data

hinders the comprehensive understanding of urban structures [11][17]. To overcome this limitation, clustering of multi-source geospatial big data has received increasing attention in recent years. For example, some scholars [28] first combined the taxi trajectory data and public transit records to reveal human mobility patterns, then used POI features as prior knowledge to extract features of human mobility patterns, and finally performed k-means on the extracted features. To consider the contributions of different types of geospatial big data, the weighted average strategy was employed to fuse the features of multi-source geospatial big data. The weights of different types of geospatial big data can be determined based on the proportions of total bus and cab ridership [15] or the entropy weight approach [16]. The weighted average methods can fuse the information of multi-source geospatial big data to a certain extent; however, they cannot incorporate complex interactions and correlations among multi-source geospatial big data. Researchers can assume that the cone reflects the socioeconomic information that comprehensively describe the urban structures (i.e., the underlying structure of multi-source geospatial big data). In practice, this socioeconomic information is often embedded in different types of geospatial data (e.g., triangle and circle). Different types of geospatial data can be regarded as different views to observe socioeconomic information. The weighted average strategy does not capture the complementarity of multi-source geospatial big data. Therefore, the result of the weighted average strategy may be only a simple superposition of multiple features. Therefore, the underlying structure of multi-source geospatial big data cannot be reconstructed by using the weighted average strategy.

Compared with the weighted average strategy, multi-view subspace clustering has the potential to reconstruct the underlying structure of multi-source geospatial data [18][19]. Multi-view subspace clustering assumes that multi-view data points are drawn from a shared low-dimensional subspace, rather than being uniformly distributed in the original space [29]. The features of each type of geospatial big data can be reconstructed from the shared subspace. In theory, multi-view subspace clustering can fuse the shared and complementary information among different types of geospatial big data. Existing multi-view subspace clustering methods are mainly extensions of self-representation-based subspace clustering methods [30][31]. Self-representation-based subspace clustering assumes that each point xi

can be represented by a linear combination of other points xj ($j \neq i$) [32][33][34][35]. Previous multi-view subspace clustering methods first calculate a subspace representation for each type of data and then combine the multiple subspace representations for clustering [21][29][36][37][38]. Although these methods can consider the shared and/or specific information of multi-source data, the subspaces reconstructed using the original data are not robust to redundant features and noise in the original data [39]. To address this limitation, latent multi-view subspace clustering methods have recently been developed [18][22]; these methods first use dimension reduction techniques to project the original data features into a latent representation, and then use the latent representation for subspace clustering. Although latent multi-view subspace clustering methods can boost the clustering performance of multi-source geospatial big data, two challenges should be further addressed: (1) Existing method usually used a linear projection to transform the original data features into a latent representation [22][39][40]; however, the relationship between each type of data and its latent representation is usually non-linear [18][41]. Therefore, the inaccurate latent representations obtained by existing methods may degrade the clustering performance. (2) The neighboring relationships of data points in high-dimensional, non-uniform, and noisy original data are difficult to preserve in the shared subspace [34][42]. Some scholars have used neighbor graphs as constraints to preserve the neighboring relationships of data points in multi-view subspace clustering [21][43][44]; however, the neighbor graphs defined based on Euclidean distance and k-nearest neighbor cannot construct appropriate neighboring relationships for high-dimensional and non-uniform geospatial big data [45][46]. Therefore, existing methods are highly likely to generate an inaccurate subspace, which will reduce the clustering quality [47].

References

1. Pan, G.; Qi, G.; Wu, Z.; Zhang, D.; Li, S. Land-use classification using taxi gps traces. IEEE Trans. Intell. Transp. Syst. 2013, 14, 113–123.
2. Long, Y.; Shen, Z. Discovering functional zones using bus smart card data and points of interest in Beijing. In Geospatial Analysis to Support Urban Planning in Beijing; Long, Y., Shen, Z., Eds.; Springer: Berlin, Germany, 2015; Volume 116, pp. 193–217.
3. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.L.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. Int. J. Geogr. Inf. Sci. 2014, 28, 1988–2007.
4. Comito, C.; Pizzuti, C.; Procopio, N. Online clustering for topic detection in social data streams. In Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence, San Jose, CA, USA, 6–8 November 2016; pp. 362–369.

5. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. Int. J. Geogr. Inf. Sci. 2017, 31, 825–848.
6. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.L. Limits of predictability in urban mobility. Science 2010, 327, 1018–1021.
7. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Shi, L. Social Sensing: A new approach to Understanding Our Socioeconomic Environments. Ann. Assoc. Am. Geogr. 2015, 105, 512–530.
8. Yin, J.; Dong, J.; Hamm, N.; Li, Z.; Wang, J.; Xing, H.; Fu, P. Integrating remote sensing and geospatial big data for urban land use mapping: A review. Int. J. Appl. Earth. Obs. 2021, 103, 102514.
9. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.
10. Song, C.; Pei, T.; Ma, T.; Du, Y.; Shu, H.; Guo, S.; Fan, Z. Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization. Int. J. Geogr. Inf. Sci. 2019, 33, 134–154.
11. Zhang, X.; Xu, Y.; Tu, W.; Ratti, C. Do different datasets tell the same story about urban mobility—A comparative study of public transit and taxi usage. J. Transp. Geogr. 2018, 70, 78–90.
12. Zhai, W.; Bai, X.; Shi, Y.; Han, Y.; Peng, Z.R.; Gu, C. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. Comput. Environ. Urban Syst. 2019, 74, 1–12.
13. Hu, S.; He, Z.; Wu, L.; Yin, L.; Xu, Y.; Cui, H. A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. Comput. Environ. Urban Syst. 2020, 80, 101442.
14. Ye, C.; Zhang, F.; Mu, L.; Gao, Y.; Liu, Y. Urban function recognition by integrating social media and street-level imagery. Environ. Plan. B-Urban Anal. City Sci. 2021, 48, 1430–1444.
15. Yue, M.; Kang, C.; Andris, C.; Qin, K.; Liu, Y.; Meng, Q. Understanding the interplay between bus, metro, and cab ridership dynamics in Shenzhen, China. Trans. GIS 2018, 22, 855–871.
16. Tu, W.; Zhu, T.; Xia, J.; Zhou, Y.; Lai, Y.; Jiang, J.; Li, Q. Portraying the spatial dynamics of urban vibrancy using multi-source urban big data. Comput. Environ. Urban. Syst. 2020, 80, 101428.
17. Liu, J.; Li, J.; Li, W.; Wu, J. Rethinking big data: A review on the data quality and usage issues. ISPRS-J. Photogramm. Remote Sens. 2016, 115, 134–142.
18. Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; Xu, D. Generalized Latent Multi-View Subspace Clustering. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 42, 86–99.
19. Liu, Q.; Huan, W.; Deng, M.; Zheng, X.; Yuan, H. Inferring Urban Land Use from Multi-Source Urban Mobility Data Using Latent Multi-View Subspace Clustering. ISPRS Int. J. Geo-Inf. 2021, 10, 274.
20. Sagiroglu, S.; Sinanc, D. Big Data: A Review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
21. Fan, Y.; He, R.; Hu, B.G. Global and local consistent multi-view subspace clustering. In Proceedings of the Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 564–568.
22. Zhou, T.; Zhang, C.; Peng, X.; Bhaskar, H.; Yang, J. Dual Shared-Specific Multi-view Subspace Clustering. IEEE T. Cybern. 2019, 50, 3517–3530.
23. Zheng, Q.; Zhu, J.; Ma, Y.; Li, Z.; Tian, Z. Multi-view subspace clustering networks with local and global graph information. Neurocomputing 2021, 449, 15–23.
24. Toole, J.L.; Ulm, M.; González, M.C.; Bauer, D. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China, 12–16 August 2012; pp. 1–8.
25. Krishna, K.; Murty, M. Genetic K-means algorithm. IEEE Trans. Syst. Man Cybern. 1999, 29, 433–439.
26. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems; MIT Press: Vancouver, BC, Canada, 2002; Volume 14, pp. 849–856.
27. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 1996; Volume 96, pp. 226–231.
28. Yuan, N.J.; Zheng, Y.; Xie, X.; Wang, Y.; Zheng, K.; Xiong, H. Discovering urban functional zones using latent activity trajectories. IEEE Trans. Knowl. Data Eng. 2015, 27, 712–725.
29. Gao, H.; Nie, F.; Li, X.; Huang, H. Multi-view subspace clustering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 4238–4246.

30. Parsons, L.; Haque, E.; Liu, H. Subspace clustering for high dimensional data: A review. Acm Sigkdd Explor. Newsl. 2004, 6, 90–105.
31. Vidal, R. Subspace clustering. IEEE Signal. Process. Mag. 2011, 28, 52–68.
32. Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; Ma, Y. Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 35, 171–184.
33. Elhamifar, E.; Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 2765–2781.
34. Hu, H.; Lin, Z.; Feng, J.; Zhou, J. Smooth representation clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–17 June 2014; pp. 3834–3841.
35. Li, C. Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework. IEEE Trans. Image Process. 2017, 26, 2988–3001.
36. Cao, X.; Zhang, C.; Fu, H.; Liu, S.; Zhang, H. Diversity-induced multi-view subspace clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 586–594.
37. Luo, S.; Zhang, C.; Zhang, W.; Cao, X. Consistent and specific multi-view subspace clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 3730–3737.
38. Zhu, P.; Hui, B.; Zhang, C.; Du, D.; Wen, L.; Hu, Q. Multi-view Deep Subspace Clustering Networks. arXiv 2019, arXiv:1908.01978. 2019.
39. Zhang, C.; Hu, Q.; Fu, H.; Zhu, P.; Cao, X. Latent multi-view subspace clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4279–4287.
40. Yu, X.; Liu, H.; Wu, Y.; Zhang, C. Intrinsic self-representation for multi-view subspace clustering. Sci. China Inf. Sci. 2021, 51, 1625–1639.
41. Wang, X.; Liu, H.; Qian, X.; Jiang, Y.; Deng, Z.; Wang, S. Cascaded hidden space feature mapping, fuzzy clustering, and nonlinear switching regression on large datasets. IEEE Trans. Fuzzy Syst. 2018, 26, 640–655.
42. Wang, X.; Lei, Z.; Guo, X.; Zhang, C.; Shi, H.; Li, S.Z. Multi-view subspace clustering with intactness-aware similarity. Pattern Recognit. 2019, 88, 50–63.
43. Zhu, W.; Lu, J.; Zhou, J. Structured General and Specific Multi-view Subspace Clustering. Pattern Recognit. 2019, 93, 392–403.
44. Zheng, Q.; Zhu, J.; Li, Z.; Pang, S.; Wang, J.; Li, Y. Feature concatenation multi-view subspace clustering. Neurocomputing 2020, 379, 89–102.
45. Xia, S.; Xiong, Z.; Luo, Y.; Zhang, G. Effectiveness of the Euclidean distance in high dimensional spaces. Optik 2015, 126, 5614–5619.
46. Liu, Q.; Liu, W.; Deng, M.; Cai, J.; Liu, Y. An adaptive detection of multilevel co-location patterns based on natural neighborhoods. Int. J. Geogr. Inf. Sci. 2021, 35, 556–581.
47. Wang, Q.; Cheng, J.; Gao, Q.; Zhao, G.; Jiao, L. Deep multi-view subspace clustering with unified and discriminative learning. IEEE Trans. Multimed. 2020, 23, 3483–3493.