PARAMETER-FREE INCREMENTAL CO- CLUSTERING ON OPTIMIZED FEATURES

S. R. Sandhan[1] & Prof. S. S. Banait[2]
[1]M.E. Student, [2]Assistant Professor
Department of Computer Engineering,
K. K. Wagh Institute of Engineering Education & Research, Nashik,
Savitribai Phule Pune University, Maharashtra, India

Received: December 05, 2018 Accepted: January 16, 2019

ABSTRACT: *To improve effectiveness and efficiency for clustering dynamic multi-modal data in cyber-physical-social systems, several co-clustering approaches have been proposed with the rapid growth of Cyber-Physical-Social Systems. Large amounts of dynamic multi-modal data are being generated and collected, so clustering Multi modal dynamic data is challenging. In high- dimensional dynamic data, classic metrics fail in identifying real similarities between objects. Moreover, the huge number of features makes the cluster interpretation hard. In the proposed method, the single-modality similarity measure is extended to multiple modalities and three operations, those are cluster creating, cluster merging, and instance partitioning, are defined to incrementally integrate new arriving objects to current clustering patterns without introducing additive parameters. Moreover, an adaptive weight scheme is designed to measure the importance of feature modalities based on the intra-cluster scatters and also add feature selection method for reducing time. Extensive experiments on various multimodal dynamic datasets demonstrate the effectiveness of the proposed approach.*

Key Words: *Parameter-Free learning; Multi modal data; Clustering; Feature selection.*

I. Introduction

Clustering is a popular data mining technique that enables to partition data into groups (clusters) in such a way that objects inside a group are similar to each other, and objects belonging to different groups are dissimilar. When data are represented in a high-dimensional space, traditional clustering algorithms fail in finding an optimal partitioning because of the problem known as curse of dimensionality. Even though some distance metrics have been proposed to deal with high- dimensional data (e.g., cosine similarity) and feature selection tries to solve the problem by a reduction in the number of features, novel approaches emerged in the last years. Clustering is a fundamental tool in unsupervised learning that is used to group together similar objects [14], and has practical importance in a wide variety of applications such as text, web-log and market- basket data analysis. Typically, the data that arises in these applications is arranged as a contingency or co-occurrence table, such as, word-document co-occurrence table or web page-user browsing data.

Most clustering algorithms focus on one-way clustering, i.e., cluster one dimension of the table based on similarities along the second dimension. For example, documents may be clustered based upon their word distributions or words may be clustered based upon their distribution amongst documents. we propose a parameter-free incremental co-clustering (PFICC) algorithm for multi-modal data in cyber-physical-social systems, which can deal with the multimodality, high-volume and dynamical-evolving data without introducing additive parameters. The contributions are fourfold: (1) The single-modality similarity measure is extended and a new weighted method is designed to calculate the similarity between multi-modal objects. (2) Three clustering operations, cluster creating, cluster merging and instance partitioning, are defined to incrementally integrate new arriving objects and adjust clustering patterns dynamically. (3) An adaptive weight scheme is designed to incrementally measure the importance of feature modalities based on the intra-cluster scatters. (4) No parameter setting is needed for the proposed method in the co-clustering processes.

II. LITURATURE WORK

For capturing conceptual factors from multi-view data, a multi-view Concept Learning (MCL) was proposed which is a novel nonnegative latent representation learning algorithm. Both multi-view information and label information is exploited by MCL in [1]. A Tucker deep computation model was proposed by using the

Tucker decomposition to compress the weight tensors in the full-connected layers for multimedia feature learning. Tucker deep computation model was used to tackle the problem of training a deep computation model with millions of parameters needs high-performance servers with large-scale memory and powerful computing units, limiting the growth of the model size for multimedia feature learning on common devices such as portable CPUs and conventional desktops. To train the parameters of the Tucker deep computation model, a learning algorithm based on the back-propagation strategy is devised in [2]. A new robust large-scale multi-view clustering method [3] was proposed to integrate heterogeneous representations of large-scale data. The proposed new methods were evaluated by six benchmark data sets and compared the performance with several commonly used clustering approaches as well as the baseline multi-view clustering methods. A new framework is presented in [4] for efficient analysis of high-dimensional economic big data based on innovative distributed feature selection. The functionality rests on three pillars: (i) novel data pre-processing techniques to prepare high-quality economic data, (ii) an innovative distributed feature identification solution to locate important and representative economic indicators from multidimensional data sets, and (iii) new econometric models to capture the hidden patterns for economic development. To offer a neat categorization and organization, the multi-view learning methods are described in terms of three classes. Representative algorithms and newly proposed algorithms are presented for each category. An comprehensive introduction was provided for the recent developments of multi-view learning methods on the basis of coherence with early methods in [5].

A novel incomplete multi-view clustering method was developed in [6], which projects all incomplete multi-view data to a complete and unified representation in a common subspace. A deep incomplete multi-view clustering (DIMC) incorporating with the constraint of intrinsic geometric structure is proposed here to couple incomplete multi-view samples, to identify the common subspace. Two high-order possibilistic c-means algorithms was proposed in [7] based on the canonical polyadic decomposition (CP-HOPCM) and the tensor-train network (TT-HOPCM) for clustering big data. Canonical polyadic decomposition and the tensor-train network was used to compress the attributes of each big data sample. Constructing Principal Components Analysis (PCA) or random projections using multiple views of the data, via Canonical Correlation Analysis (CCA) was considered in [8]. The problem of modeling Internet images and associated text or tags was investigated for tasks such as image-to-image search, tag-to-image search, and image-to-tag search (image annotation). The problem of modeling Internet images and associated text or tags for tasks such as image-to-image search, tag-to-image search, and image-totag search is presented in [9]. A popular and successful approach for mapping visual and textual features to the same latent space, and incorporate a third view capturing high-level image semantics, represented either by a single category or multiple non-mutually- exclusive concepts. A large-margin learning framework to discover a predictive latent subspace representation shared by mul- tiple views is presented in [10]. This approach is based on an undirected latent space Markov network that fulfills a weak conditional independence assumption that multi-view observations and response variables are independent given a set of latent variables.

To organize and highlight similarities and differences between the variety of multi-view learning approaches, a number of representative multi-view learning algorithms are available in different areas and they are classified into three groups: 1) co- training, 2) multiple kernel learning, and 3) subspace learning. Co-training style algorithms train alternately to maximize the mutual agreement on two distinct views of the data in [11]. A spectral clustering algorithm for the multi-view setting is proposed in [12] where the access is available to multiple views of the data, each of which can be independently used for clustering. A spectral clustering algorithm has a flavour of co-training, which is already a widely used idea in semi-supervised learning. This approach is only search for the clusterings that agree across the views. A new algorithm which learns discriminative subspaces in an unsupervised fashion is proposed in [13] to solve the problem of unsupervised clustering with multi-view data of high dimensionality. This algorithm works based upon the assumption that a reliable clustering should assign same-class samples to the same cluster in each view. The framework combines the simplicity of k-means clustering and Linear Discriminant Analysis (LDA) within a co-training scheme which exploits labels learned automatically in one view to learn discriminative subspaces in another. Composite Kernel Learning method is proposed in [14] to address the situation where distinct components give rise to a group structure among kernels. Multiple Kernel Learning enables to learn the kernel, from an ensemble of basis kernels, whose combination is optimized in the learning process. To combine multiple kernels instead of using a single one, several methods have been proposed in [15]. These different kernels may correspond to using different notions of similarity or may be using information coming from multiple sources (different representations or different feature subsets). An incremental clustering algorithm by fast finding and searching of density peaks based on k-mediods is proposed in [16]. In this two cluster operations, namely cluster creating and cluster merging, are defined to integrate the current pattern into the previous one for the final clustering result, and k-mediods is employed to modify the clustering centers according to the new arriving objects. To handle large dynamic data, incremental clustering approaches are proposed in [17].

A new incremental clustering approach called incremental multiple medoids-based fuzzy clustering (IMMFC) is proposed to handle complex patterns that are not compact and well separated. Affinity Propagation (AP) clustering has been successfully used in a lot of clustering problems. However, most of the applications deal with static data. The difficulties in Incremental Affinity Propagation (IAP) clustering, and then propose two strategies to solve them. Correspondingly, two IAP clustering algorithms are proposed in [18]. The incremental behaviours of Density based clustering is described in [19]. It specially focuses on the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and its incremental approach. DBSCAN relies on a density based notion of clusters. It discovers clusters of arbitrary shapes in spatial databases with noise. In incremental approach, the DBSCAN algorithm is applied to a dynamic database where the data may be frequently updated. After insertions or deletions to the dynamic database, the clustering discovered by DBSCAN has to be updated. And we measure the new cluster by directly compute the new data entering into the existing clusters instead of rerunning the algorithm. Incremental clustering approaches have been proposed in [20] for handling large data when given data set is too large to be stored. The key idea of these approaches is to find representatives to represent each cluster in each data chunk and final data analysis is carried out based on those identified representatives from all the chunks.

III. CONCLUSION

A parameter-free incremental co-clustering method is proposed aiming at clustering multi-modal CPSS data effectively and efficiently. PFICC extends the single modality similarity measure to multiple modalities so that PFICC can be applied to the integration of data including different types of feature sets. an adaptive weight scheme is designed to measure the importance of feature modalities based on the intra-cluster scatters.

IV. Acknowledgment

Authors would like to thank Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

References

1. Z. Guan, L. Zhang, J. Peng, and J. Fan, "Multi-view concept learning for data representation," IEEE Trans. Knowl. Data Eng., vol. 27, no. 11, pp. 3016–3028, Nov. 2015.
2. Q. Zhang, L. T. Yang, X. Liu, Z. Chen, and P. Li, "A tucker deep computation model for mobile multimedia feature learning," ACM Trans. Multimedia Comput., Commun., Appl., vol. 13, no. 3s, p. 39, 2017.
3. X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in Proc. IJCAI, 2013, pp. 2598–2604.
4. L. Zhao, Z. Chen, Y. Hu, G. Min, and Z. Jiang, "Distributed feature selection for efficient economic big data analysis," IEEE Trans. Big Data, to be published, doi: 10.1109/TBDATA.2016.2601934.
5. J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," Inf. Fusion, vol. 38, pp. 43–54, Nov. 2017.
6. L. Zhao, Z. Chen, Y. Yang, Z. J. Wang, and V. C. M. Leung, "Incomplete multi-view clustering via deep semantic mapping," Neurocomputing, to be published, doi: https://doi.org/10.1016/j.neucom.2017.07.016.
7. Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "High-order possibilistic cmeans algorithms based on tensor decompositions for big data in IoT," Inf. Fusion, vol. 39, pp. 72–80, Jan. 2018.
8. K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in Proc. 26th Annu. Int. Conf. Mach. Learn., 2009, pp. 129–136.
9. Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," Int. J. Comput. Vis., vol. 106, no. 2, pp. 210–233, 2014.
10. N. Chen, J. Zhu, and E. P. Xing, "Predictive subspace learning for multiview data: A large margin approach," in Proc. Adv. Neural Inf. Process. Syst., 2010, pp. 361–369.
11. C. Xu, D. Tao, and C. Xu. (2013). "A survey on multi-view learning." [Online]. Available: https://arxiv.org/abs/1304.5634
12. A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in Proc. 28th Int. Conf. Mach. Learn. (ICML), 2011, pp. 393–400.
13. X. Zhao, N. Evans, and J.-L. Dugelay, "A subspace co-training framework for multi-view clustering," Pattern Recognit. Lett., vol. 41, pp. 73–82, May 2014.

14. M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy, ''Composite kernel learning,'' Mach. Learn., vol. 79, nos. 1–2, pp. 73–103, 2010.
15. M. Gönen and E Alpaydın, ''Multiple kernel learning algorithms,'' J. Mach. Learn. Res., vol. 12, pp. 2211–2268, Jul. 2011.
16. Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, ''An incremental CFS algorithm for clustering large data in industrial Internet of Things,'' IEEE Trans. Ind. Informat., vol. 13, no. 3, pp. 1193–1201, Jun. 2017.
17. Y. Wang, L. Chen, and J.-P. Mei, ''Incremental fuzzy clustering with multiple medoids for large data,'' IEEE Trans. Fuzzy Syst., vol. 22, no. 6, pp. 1557–1568, Dec. 2014.
18. L. Sun and C. Guo, ''Incremental affinity propagation clustering based on message passing,'' IEEE Trans. Knowl. Data Eng., vol. 26, no. 11, pp. 2731–2744, Nov. 2014.
19. S. Chakraborty and N. K. Nagwani. (2014). ''Analysis and study of incremental dbscan clustering algorithm.'' [Online]. Available: https://arxiv.org/abs/1406.4754
20. Y. Wang, L. Chen, and X. Li. (2016) ''Incremental minimax optimization based fuzzy clustering for large multi-view data.'' [Online]. Available: https://arxiv.org/abs/1608.07001.