ADAPTIVE CONSISTENCY PROPAGATION METHOD FOR GRAPH CLUSTERING

L.Karthikeyan[1], S.Mahalakshmi[2]

[1]*Dept of Computer Science, School of Arts and Science, Vinayaka Mission's Research Foundation, Chennai, India.*
Email : karthik29071999@gmail.com

[2]*Assistant Professor, Dept of Computer Science, School of Arts and Science, Vinayaka Mission's Research Foundation, Chennai, India.*

Email: maha.cs@avit.ac.in

ABSTRACT

Graph clustering plays an important role in data mining. Based on an input data graph, data points are partitioned into clusters. However, most existing methods keep the data graph fixed during the clustering procedure, so they are limited to exploit the implied data manifold and highly dependent on the initial graph construction. Inspired by the recent development on manifold learning, this paper proposes an Adaptive Consistency Propagation (ACP) method for graph clustering. In order to utilize the features captured from different perspectives, we further put forward the Multi-view version of the ACP model (MACP). The main contributions are threefold: (1) the manifold structure of input data is sufficiently exploited by propagating the topological connectivity's between data points from near to far; (2) the optimal graph for clustering is learned by taking graph learning as a part of the optimization procedure; (3) the negotiation among the heterogeneous features is captured by the multi-view clustering model. Extensive experiments on real-world datasets validate the effectiveness of the proposed methods on both single- and multi-view clustering, and show their superior performance over the state-of-the-arts.

Keywords: ***Clustering, Manifold Learning, Graph Learning, Consistency PropagationI.***

INTRODUCTION

Clustering is a fundamental task in the field of data mining with various applications, and has attracted many researchers in the past several decades. The objective of clustering is to divide the data points into different clusters. To achieve this goal, plenty of methods have been proposed, including k-means clustering, hierarchical clustering, spectral clustering, spectral embedded clustering , maximum margin clustering, support vector clustering, normalized cut, multi-view clustering, Nonnegative Matrix Factorization , etc. Among the existing approaches, graph-based clustering methods (e.g., Ratiocut, Normalized-cut) have achieved relatively good performance because of the utilization of manifold information, and been widely used in various applications, such as image segmentation and protein sequence clustering .

Semi-supervised learning is expected that, when labeled samples are limited, the performance can be improved by the massive and easily obtained unlabeled samples when compared to supervised algorithms that only use a small number of labeled samples for training. However, it has been found that the performances of current semi-supervised learning approaches may be seriously deteriorated because the noise in misclassified samples is added into the iteration. Thus, how to avoid the adverse impact of label noise is a key issue for LPA to spread label information among samples. Recently, a novel method combined with particle competition and cooperation (PCC) was presented. Simply, this approach propagates the labels to the whole network by the random-greedy walk of particles based on PCC mechanism. Hence, aiming at the

issue of misclassified label noise in label propagation, the particle competition and cooperation mechanism is adopted in LPA.

Related Works

"Spectral k-way ratio-cut partitioning and clustering,"

Recent research on partitioning has focussed on the ratio-cut cost metric which maintains a balance between the sizes of the edges cut and the sizes of the partitions without fixing the size of the partitions a priori. Iterative approaches and spectral approaches to two-way ratio-cut partitioning have yielded higher quality partitioning results. In this paper we develop a spectral approach to multiway ratio-cut partitioning which provides a generalization of the ratio-cut cost metric to k-way partitioning and a lower bound on this cost metric. Our approach involves finding the k smallest eigenvalue/eigenvector pairs of the Laplacian of the graph. The eigenvectors provide an embedding of the graph's n vertices into a k-dimensional subspace. We devise a time and space efficient clustering heuristic to coerce the points in the embedding into k partitions. Advancement over the current work is evidenced by the results of experiments on the standard benchmarks.

"Support vector clustering,"

We present a novel clustering method using the approach of support vector machines. Data points are mapped by means of a Gaussian kernel to a high dimensional feature space, where we search for the minimal enclosing sphere. This sphere, when mapped back to data space, can separate into several components, each enclosing a separate cluster of points. We present a simple algorithm for identifying these clusters. The width of the Gaussian kernel controls the scale at which the data is probed while the soft margin constant helps coping with outliers and overlapping clusters. The structure of a dataset is explored by varying the two parameters, maintaining a minimal number of support vectors to assure smooth cluster boundaries. We demonstrate the performance of our algorithm on several datasets.

"The relationships among various nonnegative matrix factorization methods for clustering,"

The nonnegative matrix factorization (NMF) has been shown recently to be useful for clustering and various extensions and variations of NMF have been proposed recently. Despite significant research progress in this area, few attempts have been made to establish the connections between various factorization methods while highlighting their differences. In this paper we aim to provide a comprehensive study on matrix factorization for clustering. In particular, we present an overview and summary on various matrix factorization algorithms and theoretically analyze the relationships among them. Experiments are also conducted to empirically evaluate and compare various factorization methods. In addition, our study also answers several previously unaddressed yet important questions for matrix factorizations including the interpretation and normalization of cluster posterior and the benefits and evaluation of simultaneous clustering. We expect our study would provide good insights on matrix factorization research for clustering.

"Maximum margin clustering,"

We propose a new method for clustering based on finding maximum mar-gin hyperplanes through data. By reformulating the problem in terms of the implied equivalence relation matrix, we can pose the problem as a convex integer program. Although this still yields a difficult computational problem, the hard-clustering constraints can be relaxed to a

soft-clustering formulation which can be feasibly solved with a semi definite program. Since our clustering technique only depends on the data through the kernel matrix, we can easily achieve nonlinear clusterings in the same manner as spectral clustering. Experimental results show that our maximum margin clustering technique often obtains more accurate results than conventional clustering methods. The real benefit of our approach, however, is that it leads naturally to a semi-supervised training method for support vector machines. By maximizing the margin simultaneously on labeled and unlabeled training data, we achieve state of the art performance by using a single, integrated learning principle.

II. EXISTING SYSTEM

Existing graph-based clustering methods first construct a data graph according to the pair wise similarities of points, and then perform graph-theoretic optimization on the data graph. The two stage processing brings three major drawbacks. First, in the data graph, the similarity is large only for the neighbors. However, for data with manifold structure, the far away points may also keep high consistency if they are linked by consecutive neighbors. Therefore, these methods are limited to discover the underlying data structure. Second, once the data graph is constructed, they are fixed during the clustering. Then traditional methods are unable to learn the optimal graph for clustering, and tend to fail if the initial graph is constructed with low quality. Third, the graph-theoretic optimization cannot produce the clustering results directly, so a post-processing (e.g., k-means) has to be followed, which makes the result deviated from the optimal solution.

DISADVANTAGES

- In this paper, a new graph clustering method, namely Adaptive Consistency Propagation (ACP) is developed to tackle the above issues.
- During the optimization stage, they update the data graph adaptively. In this way, graph learning is successfully integrated into the clustering procedure.
- Benefited from graph learning, these methods are more robust to the initial graph quality. However, these methods still suffer from the first problem.

III. PROPOSED SYSTEM

Proposed to tackle the last two problems. During the optimization stage, they update the data graph adaptively. In this way, graph learning is successfully integrated into the clustering procedure. The topological consistency of points are fully captured to investigate the data manifold. By propagating the consistency through neighbors, the proposed method is suitable to handle data with manifold structures. A multi-view version of the proposed model is designed, which learns the correlation between the multi-view data and integrates them with the optimal combination.

ADVANTAGES

- The objective of clustering is to divide the data points into different clusters.
- Graph learning is jointly combined into the clustering framework. The data graph is optimized adaptively in the optimization stage, so the clustering is less affected by the quality of the initial graph.
- An multi-view version of the proposed model is designed, which learns the correlation between the multi-view data and integrates them with the optimal combination.

Fig.1. Overall architecture of Proposed System

IV Module Description

Clustering:

Clustering is a fundamental task in the field of data mining with various applications, and has attracted many researchers in the past several decades. The objective of clustering is to divide the data points into different clusters. To achieve this goal, plenty of methods have been proposed, including k-means clustering, hierarchical clustering, spectral clustering, spectral embedded clustering , maximum margin clustering, support vector clustering, normalized cut, multi-view clustering, Nonnegative Matrix Factorization , etc. Among the existing approaches, graph-based clustering methods (e.g., Ratiocut, Normalized-cut) have achieved relatively good performance because of the utilization of manifold information, and been widely used in various applications, such as image segmentation and protein sequence clustering .

Adaptive consistency propagation:

In this paper, a new graph clustering method, namely Adaptive Consistency Propagation (ACP) is developed to tackle the above issues. The multi-view version of the ACP method is also developed to deal with the data obtained from different feature extractors. The main contributions of this study are summarized as follows. The topological consistency of points are fully captured to investigate the data manifold. By propagating the consistency through neighbors, the proposed method is suitable to handle data with manifold structures. Graph learning is jointly combined into the clustering framework. The data graph is optimized adaptively in the optimization stage, so the clustering is less affected by the quality of the initial graph.

Propagation-Based Manifold Learning Method Revisited

The Propagation-Based Manifold Learning method (PBML) proposed by aims to learn the topological relationship of individuals in crowd scene. Given a similarity graph G 2 Rnn (n is the number of individuals) of individuals, PBML assumes that individuals with large similarity should share similar topological relevance to any other point.

Multi-view adaptive consistency propagation

In real world applications, objects could be represented from multiple views. For example, in computer vision, an image may be described by different features, such as SIFT, HOG and CENT. Each feature captures a specific statistical property, and it is necessary to integrate these heterogeneous features and utilize the complementary information. In this section, we propose the Multi-view version of the ACP model (MACP).

V. CONCLUSION

In this project, the Adaptive Consistency Propagation (ACP) and its multi-view version MACP are proposed for clustering. Most of the traditional methods only focus on the data points with neighboring relationship, and keep the data graph fixed during the optimization procedure. In our new methods, the local consistency is propagated adaptively from near to far, so the points from the same cluster can be all pulled together. In addition, with a reasonable constraint, ACP and MACP are able to learn the optimal graph

for clustering, and accomplish clustering simultaneously without any post-processing.

FUTURE WORK

Comprehensive experiments on single and multi-view clustering show the superior performance of our methods on various kinds of datasets.

REFERENCES

[1] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[2] F. J. Rohlf, "Adaptive hierarchical clustering schemes," Systematic Zoology, vol. 19, no. 1, pp. 58–82, 1970.

[3] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," TNNLS, vol. 30, no. 5, pp. 1581–1586, 2019.

[4] F. Nie, Z. Zeng, I. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," IEEE Transactions on Neural Networks, vol. 22, no. 11, pp. 1796–1808, 2011.

[5] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in Advances in Neural Information Processing Systems, 2004, pp. 1537–1544.

[6] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," Journal of Machine Learning Research, vol. 2, pp. 125–137, 2001.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis on Machine Intelligence, vol. 22, no. 8, pp. 888–905, 2000.

[8] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in AAAI Conference on Artificial Intelligence, 2017, pp. 4147–4153.

[9] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in IEEE International Conference on Data Mining, 2006, pp. 362–371.

[10] P. Chan, M. Schlag, and J. Zien, "Spectral k-way ratio-cut partitioning and clustering," IEEE Transactions on CAD of Integrated Circuits and Systems, vol. 13, no. 9, pp. 1088–1096, 1994.