Master's Thesis Pre-Proposal: A Framework for Evolutionary Clustering on Multi-type Relational Data

Amit Salunke

February 11, 2011

Rapid development in data acquisition technology has resulted in generating large amount of raw data, providing significant potential for the development of automatic data analysis, classification, and retrieval techniques. Data in many applications such as social networks, blogs, geosciences, and biomedicine, demonstrates an evolving nature. That is, the similarity between the data instances and/or the number of instances might change at different timesteps. Moreover, apart from being large-scale, these datasets can be multi-type (e.g. genes, proteins, samples, in bioinformatics). So, knowledge discovery in such evolving multi-type relational data is not a trival problem.

The recent active topic of evolutionary clustering opens up many new research avenues into the mining of temporally evolving data. Clustering of evolving data (a.k.a. Evolutionary clustering) has been a relatively new topic and was first formulated by Chakrabarti et al. [1]. They proposed heuristic solutions to evolutionary hierarchical clustering problems and evolutionary k-means clustering problems. Chi et al. [2] extended this work by proposing two evolutionary spectral clustering algorithms by incorporating a measure of temporal smoothness in the overall clustering quality. Recently, Wang et al. [5] propose the evolutionary clustering using kernel function framework (ECKF) that clusters large evolutionary datasets by the amalgamation of low-rank matrix approximation methods and matrix factorization based clustering.

Heterogeneous data co-clustering has attracted more and more attention in recent years due to its high impact on various important applications, such as web mining, e-commerce and bioinformatics. Long et al. [3] proposed the spectral relational clustering, to cluster multi-type interrelated data objects or relational data with various structures simultaneously. The algorithm iteratively embeds each type of data objects into low dimensional spaces and benefits from the interactions among the hidden structures of different types of data objects.

Low-rank matrix approximation methods have found profound application in diverse areas due to their ability to extract correlations and remove noise from matrix-structured data [4].

However, in spite of these efforts, the problem of clustering multi-type evolving data has not been addressed. Since, the current methodology is unable to cluster more than one data type simultaneously, the ability to use other data mining algorithms on multi-type evolving data is desirable. To address some of these challenges, I will propose an algorithm to perform co-clustering on high-order heterogeneous evolving data, by utilizing low-rank matrix approximation and matrix factorization based clustering techniques.

To validate the proposed approach, extensive experiments on synthetic data as well as publicly available datasets, Pubmed [1] and Enron Email Corpus dataset [2] will be performed, and will be tested against results published in ECKF Framework [5].

References

[1] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 554–560, New York, NY, USA, 2006. ACM.

[2] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 153–162, New York, NY, USA, 2007. ACM.

[3] Bo Long, Zhongfei (Mark) Zhang, Xiaoyun Wú, and Philip S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 585–592, New York, NY, USA, 2006. ACM.

[4] Hanghang Tong, Spiros Papadimitriou, Jimeng Sun, Philip S. Yu, and Christos Faloutsos. Colibri: fast mining of large static and dynamic graphs. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 686–694, New York, NY, USA, 2008. ACM.

[5] Lijun Wang, Manjeet Rege, Ming Dong, and Yongsheng Ding. Low-rank kernel matrix factorization for large scale evolutionary clustering. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2010.

[1] http://www.ncbi.nlm.nih.gov/pubmed/

[2] http://www.cs.cmu.edu/∼enron/