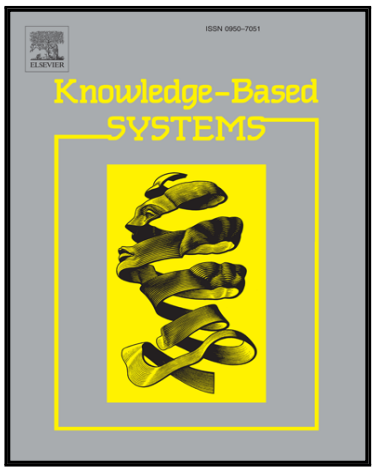

Local Graph Based Correlation Clustering

Divya Pandove, Rinkle Rani, Shivani Goel

PII: S0950-7051(17)30453-7
DOI: 10.1016/j.knosys.2017.09.034
Reference: KNOSYS 4055

To appear in: *Knowledge-Based Systems*

Received date: 5 February 2017
Revised date: 26 September 2017
Accepted date: 30 September 2017

Please cite this article as: Divya Pandove, Rinkle Rani, Shivani Goel, Local Graph Based Correlation Clustering, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.09.034

Local Graph Based Correlation Clustering

Divya Pandove[a,*], Rinkle Rani[b], Shivani Goel[c]

[a]*Research Lab, Computer Science and Engineering department, Thapar University, Patiala, Punjab, India-147004*
[b]*Computer Science and Engineering department, Thapar University, Patiala, Punjab, India-147004*
[c]*Department of Computer Science Engineering, School of Engineering and Applied Sciences, Bennett University, Greater Noida, U.P., India*

Abstract

In high-dimensional data, clusters often exist in the form complex hierarchical relationships. In order to explore these relationships, there is a need to integrate dimensionality reduction techniques with data mining approaches and graph theory. The correlations in data points emerge more clearly if this integration is flawless. We propose an approach called Local Graph Based Correlation Clustering (LGBACC). This approach merges hierarchical clustering, with PCA to uncover complex hierarchical relationships, and uses graph models to visualize the results. We propose a framework of this approach, that is divided into four phases. Each phase is flawlessly integrated with the next phase. Visualization of data after each phase is an important output and is knitted into the fabric of the framework. The focus of this technique remains on obtaining high quality clusters. The quality of the final clusters obtained is measured using standard indices. It is found that LGBACC is better to the existing hierarchical clustering approaches. We have used real world data sets to validate our framework. These datasets test the approach on low as well as high-dimensional data. It is found that LGBACC produces high quality clusters across a wide spectrum of dimensionality. Scalability test on synthetically produced high-dimensional, and large datasets show that the proposed approach runs efficiently. Hence, LGBACC is an efficient and scalable approach that produces high quality clusters in high-dimensional and large data spaces.

Keywords: Correlation clustering, Dimensionality reduction, Graph analysis, Hierarchical clustering, Cluster quality.

1. Introduction

Due to intensive data collection techniques used these days, most of the datasets are high-dimensional and large. The high number of features in these datasets might either be noisy, or exhibit false correlations among each other. In order to disregard irrelevant features, data mining techniques should be evolved enough to select relevant features. Many specialized clustering approaches have come up in the last few years, based on the classical clustering approaches. The most prominent challenge of clustering in high-dimensional data space is that relevancy of features is subjective to the clusters they belong to. In addition to this, correlations among different attributes of a data set may be relevant for different clusters. This phenomenon, that different correlations of features are relevant to varying clusters, is known as local feature selection [1]. A common practice to overcome this problem is to perform dimensionality reduction methods like principal component analysis (PCA), to convert the given data space into a data space with lower dimensions. This also helps to understand the underlying processes better, and in return improves the prediction and learning efficiency. The data clusters obtained are also more meaningful [2]. It is difficult to apply dimensionality reduction techniques to cluster high-dimensional data, as these techniques work on only one subspace of the original data space, making them global in nature. In contrast to this, the problem of "local feature relevance" requires to analyze multiple subspaces, as one cluster may be present in different subspaces [3]. Data clustering needs to be performed keeping in mind the problem of local feature relevance. In order to utilize the best features of both

*Corresponding author
Email addresses: dpandove@gmail.com (Divya Pandove), raggarwal@thapar.edu (Rinkle Rani), shigo108@yahoo.co.in (Shivani Goel)