INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

LOW-RANK APPROXIMATION-BASED SPECTRAL CLUSTERING FOR LARGE DATASETS

Rachana Jakkula[1], Jyothi.P[2]

[1] Asst.Professor, Department of Computer Science &Engineering,
TeegalaKrishna Reddy Engineering College, Hyderabad, T.S-500 097, India
rachana.friend@gmail.com
[2]Asst.Professor, Department of Computer Science & Engineering, TeegalaKrishna Reddy Engineering College,
Hyderabad, T.S-500 097,India
jyothi.p20@gmail.com

Abstract: -Spectral clustering is a well-known graph-theoretic approach of finding natural groupings in a given dataset. Nowadays, digital data are accumulated at a faster than ever speed in various fields, such as the Web, science, engineering, biomedicine, and real-world sensing. It is not uncommon for a dataset to contain tens of thousands of samples and/or features. Spectral clustering generally becomes infeasible for analysing these big data. spectral clustering has become one of the most popular modern clustering algorithms. It is simple to implement, can be solved efficiently by standard linear algebra software, and very often outperforms traditional clustering algorithms such as the k-means algorithm. A particular class of graph clustering algorithms is known as spectral clustering algorithms. These algorithms are mostly based on the Eigen-decomposition of Laplacian matrices of either weighted or unweighted graphs. This survey presents different graph clustering formulations, most of which based on graph cut and partitioning problems, and describes the main spectral clustering algorithms found in literature that solve these problems.In this paper, we propose a Low-rank Approximation-based Spectral (LAS) clustering for big data analytics. By integrating low-rank matrix approximations, i.e., the approximations to the affinity matrix and its subspace, as well as those for the Laplacian matrix and the Laplacian subspace, LAS gains great computational and spatial efficiency for processing big data. In addition, we propose various fast sampling strategies to efficiently select data samples. From a theoretical perspective, we mathematically prove the correctness of LAS, and provide the analysis of its approximation error, and computational complexity.

Keywords: Introduction, Clustering Big Data, K-means, Spectral Clustering

Introduction

BIG Data refers to a very strong growth of heterogeneous data flows due to the increased use of new technologies. In fact, with the growth of the web, the use of social networks, mobile, connected and communicating objects,

information is now more abundant than ever and it is growing faster every day. Some studies argue that handling and using intelligently this huge data could become a new pillar of economics as well as scientific research, experimentation and simulation. Indeed, many opportunities of Big Data appear in different areas such as health (enhancing the efficiency of some treatments), bio-medical, marketing (increasing sales), transportation (reducing costs), business, finance (minimizing risks), management (decision making with high efficiency and speed), social, media, and government services. Clustering is the Key to Big Data Problem. Clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, and biology to social sciences or psychology. In virtually every scientific field dealing with empirical data, people attempt to get a first impression on their data by trying to identify groups of "similar behavior" in their data. In this article we would like to introduce the reader to the family of spectral clustering algorithms. Compared to the "traditional algorithms" such as k-means or single linkage, spectral clustering has many fundamental advantages. Results obtained by spectral clustering often outperform the traditional approaches, spectral clustering is very simple to implement and can be solved efficiently by standard linear algebra methods. We derive spectral clustering from scratch and present different points of view to why spectral clustering works. Apart from basic linear algebra, no particular mathematical background is required.

Clustering Big Data:

Clustering is an essential data mining and tool for analysing big data. There are difficulties for applying clustering techniques to big data duo to new challenges that are raised with big data. As Big Data is referring to terabytes and petabytes of data and clustering algorithms are come with high computational costs, the question is how to cope with this problem and how to deploy clustering techniques to big data and get the results in a reasonable time. This study is aimed to review the trend and progress of clustering algorithms to cope with big data challenges from very first proposed algorithms until today's novel solutions.

Clustering methods

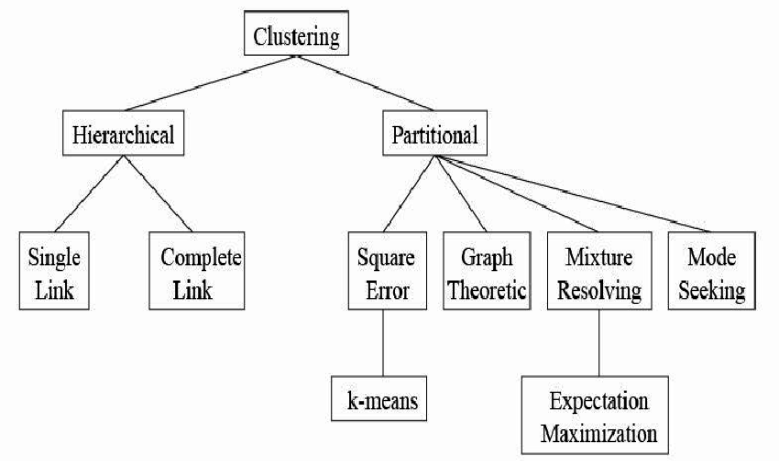


Fig-1

K-means

The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.

The difficulties Using K-Mean are:
1) Difficult to predict K-Value.
2) With global cluster, it didn't work well.
3) Different initial partitions can result in different final clusters.
4) It does not work well with clusters (in the original data) of Different size and Different density.

Spectral or Subspace Clustering

In spectral clustering the data set is represented as a graph. Each data point is represented as a vertex in the graph. spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset. Spectral clustering cares about connectivity instead of geometrical proximity. So if your data isn't well geometrically separated, but clusters aren't connected, spectral clustering will work well. Spectral clustering refers to a class of techniques which rely on the Eigen- structure of a similarity matrix to partition points into disjoint clusters with points in the same cluster having high similarity and points in different clusters having low similarity. The goal of spectral clustering is to cluster data that is connected but not necessarily compact or clustered within convex boundaries.

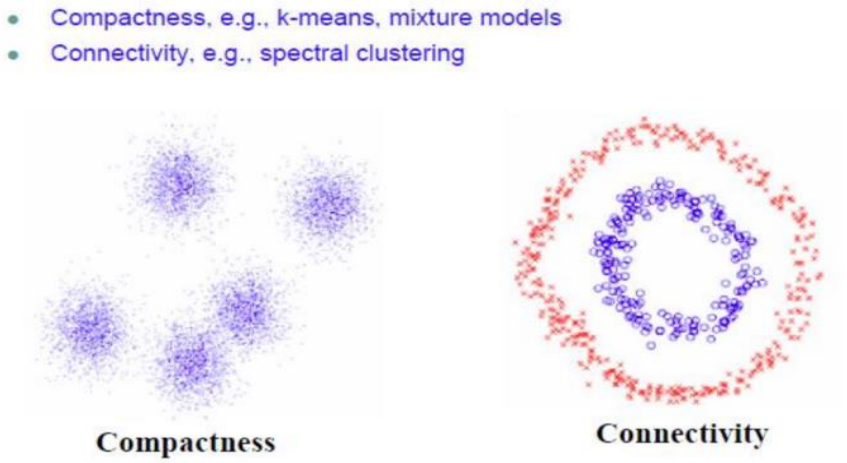


Fig-2

The basic idea:

1. project your data into R^n
2. define an A*ffinity* matrix A, using a Gaussian Kernel K or say just an Adjacency matrix (i.e. $A_{i,j} = \delta_{i,j}$)
3. construct the Graph Laplacian from A(i.e. decide on a normalization)
4. solve an Eigenvalue problem , such as $Lv = \lambda v$ (or a Generalized Eigenvalue problem $Lv = \lambda Dv$)
5. select k eigenvectors $\{v_i, i = 1, k\}$ corresponding to the k lowest (or highest) eigenvalues $\{\lambda_i, i = 1, k\}$, to define a k-dimensional subspace $P^t LP$
6. form clusters in this subspace using, say, k-means

The Cluster Eigenspace Problem

If good clusters can be identified, then the Laplacian L is approximately block-diagonal, with each block defining a cluster. So, if we have 3 major clusters $(C1, C2, C3)$, Where $L_{C1,C1}$ corresponds to subblock for cluster $C1$, etc. These blocks let us identify clusters with non-convex boundaries, as shown below:

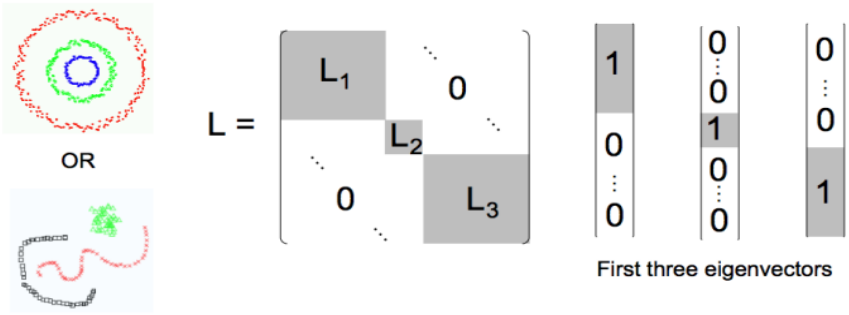


Fig-3

We also expect that the 3 lowest eigenvalues & eigenvectors (λ_i, v_i) of L each correspond to a different cluster. This occurs when each eigenvector corresponds to the lowest eigenvector of some sub block of $L_{C,C}$. That is, if

$Lv_i = \lambda_i v_i$ are the lowest eigenvalue, eigenvector pairs in the full space, and

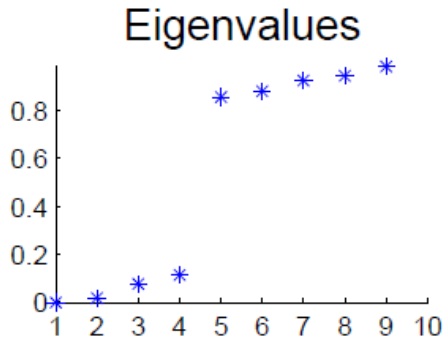
$L_{C1,C1}v_{C1} = \lambda_{C1}v_{C1}$ is the lowest eigenvalue , eigenvector pair for block C1,

then v_{C1} is a good approximation to one of the lowest 3 v_i. Likewise for subblocks C2 and C3.

More importantly, this also restricts the eigenvalue spectrum of L, so that the set lowest 3 full space eigenvalues consists of the lowest subblock eigenvalues

$\{\lambda_i, i = 1, 3\}$=$\{\lambda_{Ci}, i = 1, 3\}$

Also, to identify k clusters, the eigenvalue spectrum of L must have a gap, as shown below :



Frequently this gap is hard to find, and choosing the optimal k is called "rounding" Technically, running K-means in the subspace is not exactly the same as identifying each eigenvector with a specific cluster.Indeed, one might imagine using a slightly larger subspace than necessary, and only extracting the k clusters desired. The subtly here is getting choosing the right Affinity (matrix cutoff R and all), the right size of the subspace, and the right normalization (both of the Laplacian and the eigenvectors themselves, before or after diagonalization) You should refer to the original papers, the reviews, and whatever open source code you are using for very specific details.

CONCLUSION:

The power of Spectral Clustering is to identify *non-compact clusters* in a single data set. Spectral Clustering will only work on fairly uniform datasets–that is, data sets with N uniformly sized clusters.

REFERENCES

[1] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on PAMI, Vol. 22, No. 8, Aug 2000.
[2] Marina Meilă & Jianbo Shi, "Learning Segmentation by Random Walks", Neural Information Processing Systems 13 (NIPS 2000), 2001, pp. 873–879.
[3] Zare, Habil; P. Shooshtari; A. Gupta; R. Brinkman (2010). "Data reduction for spectral clustering to analyze high throughput flow cytometry data". BMC Bioinformatics. 11: 403. doi:10.1186/1471-2105-11-403. PMC 2923634. PMID 20667133.
[4] Arias-Castro, E. and Chen, G. and Lerman, G. (2011), "Spectral clustering based on local linear approximations.", Electronic Journal of Statistics, 5: 1537–1587, doi:10.1214/11-ejs651
[5] Free statistical software: https://github.com/ezahedi/Network-Clustering/tree/master/Spectral-clustering.
[6] Dhillon, I.S. and Guan, Y. and Kulis, B. (2004). "Kernel k-means: spectral clustering and normalized cuts". Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 551–556.
[7] Dhillon, Inderjit; Yuqiang Guan; Brian Kulis (November 2007). "Weighted Graph Cuts without Eigenvectors: A Multilevel Approach". IEEE Transactions on Pattern Analysis and Machine Intelligence. 29 (11): 1–14. doi:10.1109/tpami.2007.1115.
[8] Kannan, Ravi; Vempala, Santosh; Vetta, Adrian. "On Clusterings : Good. Bad and Spectral". Journal of the ACM. 51: 497–515,doi:10.1145/990308.990313

AUTHOR DETAILS:

Rachana Jakkula is Asst.Professor, Department of Computer Science &Engineering TeegalaKrishna Reddy Engineering College, Hyderabad, T.S-500 097, India

Jyothi.P is Asst.Professor, Department of Computer Science &Engineering TeegalaKrishna Reddy Engineering College, Hyderabad, T.S-500 097, India