Data Stream clustering based on Spectral Graph Theory

Jesús David Yanquén Correa

Universidad Nacional de Colombia
Engineering School
Bogotá, Colombia
2018

Data Stream clustering based on Spectral Graph Theory

Jesús David Yanquén Correa

Submitted to the Engineering School of the Universidad Nacional de Colombia, in partial fulfillment of the requirements for the degree of

Master of Science in Systems and Computer Engineering

Under the guidance of

German Hernandez, Ph.D.

Associate Professor
Engineering School

Universidad Nacional de Colombia
Engineering School
Bogotá, Colombia
2018

Acknowledgments

I express my sincere gratitude to my thesis advisor, Professor German Hernandez for his help, support, guidance and patience in this process.

Also, I am grateful to my family who have always given me their unconditional love.

Finally, but not least, thank God for giving me the necessary understanding to carry out this work.

Abstract

Abstract goes here

Contents

List of Figures

List of Tables

Part I.

Preliminaries

1. Introduction

In the last years clustering data stream ...

1.1. Problem Statement

1.1.1. Challenges

1.1.2. Thesis Goals

Main Goal

To propose a model to identify clusters of a data stream based on spectral graph theory. The following are the general purposes of this research:

Specific Objetives

- To propose an approach of data stream clustering based on spectral graph theory.
- To design a model to manage the evolution on the clustering result of previous approach.
- To implement the model and to evaluate its effectiveness and efficiency.

1.2. Thesis Scope

The scope of this work

1.3. Main contributions

The following is the outline of the main contributions of this work:

1.4. Outline

The remainder of this document is organized as follows. Next chapter gives a review of how data stream clustering has been envolved. Chapter 3 shows an overall background of how Spectral Graph Theory can be applied as a factible solution in the clustering problem. In Chapter 4 a semi-supervised learning framework based on direct eigenspace mapping is presented. Chapter 5 shows how to use the previous learning framework to handle datastreams. Chapter 6 presents the evaluations of the proposed model. Chapter 7 shows a summary that discusses the main aspects of this research, presents the conclusions and the future work.

2. A Review of Data Stream Clustering

Data Stream clustering has been an attractive research field in the last years due its several applications in different contexts like network traffic analysis, log record analysis, online anomaly detection ...

2.1. Previous work

Clustering without proper feature selection may not produce desirable data clusters. PCA is a technique to represent the data capturing its maximum variant dimensions A set of basis vectors that are the dominant eigenvectors of the covariance matrix

[1] Traditional clustering algorithms are based on is applied to optimize problem by minimizing certain quality metrics such as squared error. The result of clusters is convex shape in d- dimensional space. Therefore, for clustering with arbitrary shape, this strategy does not work properly. We can see arbitrary shape in area of science such as weather satellite, image segmentation spatial data which is collected from geographic information systems (GIS) and even sensor data that is posed as arbitrary shape. These applications generate enormous volume of data. Those set of algorithm which find irregular shaped cluster are called as 'shape-base clustering algorithms' [4].

In this Proposal, our focus is on the arbitrary-shape clustering task. We use the term shape-based clustering for all algorithmic techniques that capture clusters with arbitrary shapes, varying densities and sizes. Various algorithms have been applied for finding Shape-based clustering. These algorithms are categorized into: Hierarchical clustering, prototype-base algorithm, spectral-base clustering, kernel-base method, density-base method and manifold-base clustering.

Among these algorithms, prototype algorithm can only model cluster with specific shape, therefore cannot be applied for arbitrary shape. In the density method like DBScan, finding appropriate parameters is difficult and this difficulty makes use of these algorithm limited. Spectral algorithm for choosing a good similarity graph is not trivial and can be quite unstable under different choices of the parameters for neighbouring graphs [5]. Kernel-base clustering which has been used for to distinguish non-linear clustering boundaries, closely related to spectral algorithm. Kernel clustering has some major problems; its computational is expensive due to difficulty in scaling large-scale dataset because of the number of parameters is quadric respect to the number of examples. Second, it is sensitive to choose of kernel functions. Third, it needs data pre-processing to ensure that any clustering boundary will pass through the origins which makes it unstable for clustering unbalanced dataset. Spectral, density-based (DBSCAN), and nearest-neighbour graph based (Chameleon) approaches are the most successful among the many shape-based clustering methods. But when they turns to the complexity time and space limitation in data stream, aforementioned algorithms have high computational in comparison to linear distance-base clustering [5]. Some works have been done in the area of data stream clustering for arbitrary shape clustering. This continues effort has one common goal to achieve and produce an accurate data stream clustering method. Despite these studies still cannot lead the users to obtain correct result.

Part II.

Methods

3. Spectral Graph Theory

When we talk about graph theory, maybe we remember data structures like trees and related topics such as find the shortest path, breadth-first search or depth-first search, but when we said *spectral* probably we are not clear about what it refers. In short, Spectral Graph Theory studies the properties of a special graph called *Laplacian* through its eigenvalues and eigenvectors, however, there are much more than this.

Riemannian manifold Laplace Beltrami Operator Dynamic Laplacian

Assumption that the data lies in a lower dimensional manifold in a high dimensional space

Eigenfunction of the Laplacian is very usefull in the analysis on Riemannian manifold. It is the key to carry an analog of Fourier series on manifolds

we may interpret the eigenfunction φ as the probability density of a quantum particle in the energy state λ. That is, the probability that a particle in the state φ belong to the set A

a version of the dynamic isoperimetric problem, generalised to the situation where the dynamics need not be volume preserving A manifold M is splited by a compact hypersurface Γ in two submanifolds M_1 and M_2. Then is necesary transform each manifold M_1 and M_2 to N_1 and N_2 respectively

the initial manifold M(0) is transformed under the smooth flow maps

"Descriptors are some set of numbers that are produced to describe a given shape. The shape may not be entirely reconstructable from the descriptors, but the descriptors for different shapes should be different enough that the shapes can be discriminated"

4. Proposed Model

Assumption that the data lies in a lower dimensional manifold in a high dimensional space

5. Conclusions and Future Work

A. Algorithm

B. Algorithm2

Bibliography

[1] Rabbert Klein. Black holes and their relation to hiding eggs. *Theoretical Easter Physics*, 2010. (to appear).