International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119 Vol. 3 No. 1 June 2016, pp. 57-64

http://ijpsat.ijsht-journals.org

A Comparative Study - Selection of Optimal Traffic Signal using Graph Theory and Data Mining

Sonali Kashyap, Sumit Dugar

VIT University
Vellore- Tamil Nadu, India

ABSTRACT: Methods from graph theory have made an impact in Machine Learning recently through many avenues. It arises when we view the data samples as the vertices of the graph with the similarity between the data points encoded by the weights on the edges. This view of the data can be used to motivate a number of techniques, including spectral clustering, visualization and supervised classification. In this paper we basically focus on predicting the crowd based on the weight of the edges for a particular set of sample data points. We then classify these crowded places in any one of the genres like school, residential college, marketplace etc. This use of graph representations in machine learning is useful for detecting the strength & density of a particular place which might become a key aspect of development.

KEYWORDS: Vertex, Edge, Clustering, Data mining, Classification.

1. INTRODUCTION

Data mining problems are common in research fields, for example for identifying disease clusters and multidimensional patterns within large databases, e.g., socioeconomic differentials in health. Although numerous data mining methods have been developed, currently available methods are not designed to handle complex pattern searching queries and no satisfactory methods are available for this purpose. The aim of this paper is to test graph-theoretical methods for data mining in traffic based databases to identify areas of high deprivation that are surrounded by affluent areas and thronged areas surrounded by public regions like schools, colleges and marketplaces. Graph-theory plays a major role in this context. Most of the data we deal with in data science can be modelled into graphs. These graphs can be mined using famous techniques and algorithms in graph theory to get desired results.

2. BASIC DEFINITIONS

VERTEX: IN GEOMETRY, A VERTEX IS A POINT WHERE TWO OR MORE CURVES, LINES, OR EDGES MEET. AS A CONSEQUENCE OF THIS DEFINITION, THE POINT WHERE TWO LINES MEET TO FORM AN ANGLE AND THE CORNERS OF POLYGONS AND POLYHEDRAL ARE VERTICES.

EDGE: IN GEOMETRY, AN EDGE IS A PARTICULAR TYPE OF LINE SEGMENT JOINING TWO VERTICES IN A POLYGON, POLYHEDRON, OR HIGHER-DIMENSIONAL POLYTOPE. IN A POLYGON, AN EDGE IS A LINE SEGMENT ON THE BOUNDARY, AND IS OFTEN CALLED A SIDE. IN A POLYHEDRON OR MORE GENERALLY A POLYTOPE, AN EDGE IS A LINE SEGMENT WHERE TWO FACES MEET.

CLUSTERING: CLUSTER ANALYSIS OR CLUSTERING IS THE TASK OF GROUPING A SET OF OBJECTS IN SUCH A WAY THAT OBJECTS IN THE SAME GROUP (CALLED A CLUSTER) ARE MORE SIMILAR (IN SOME SENSE OR ANOTHER) TO EACH OTHER THAN TO THOSE IN OTHER GROUPS (CLUSTERS).IT IS A COMMON

Corresponding Author: Sonali Kashyap

TECHNIQUE FOR STATISTICAL DATA ANALYSIS, WHICH IS USED IN MANY FIELDS, INCLUDING MACHINE LEARNING, DATA MINING, PATTERN RECOGNITION, IMAGE ANALYSIS AND BIOINFORMATICS

Data mining: Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Classification: is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

3. BACKGROUND

Most of the data we deal with in data science can be modelled into graphs. These graphs can be mined using famous techniques and algorithms in graph theory. Eg: Social Networks are the easiest thing to comprehend among numerous things in data science related to graph theory. In social networks taking the people/web pages are nodes and their connections as edges can be modelled as a graph. Over this various mining operations can be performed better by using graph theory related algorithms. Erdos-Renyi Random Graphs are one such thing from graph theory which can simulate the evolution of social networks. Graphs are underlying in most of the data we deal. So graph theory is mostly used in mining information from such kind of data.

In this paper, we alleviate the traffic problems in advance and propose a system to cluster and visualize geographic data based on a simple classification algorithm which depends on decision tree. This paper describes research that is devising a procedure for developing, implementing and monitoring traffic signal timing plans using available data mining tools. The hypothesis premise of the research is that the data collected by Google API systems can be used to improve system design and operations for the current methods of traffic control by providing them with the corresponding regions of the crowd ahead. The data-mining tool serves as the foundation in this research for signal plan development in hierarchical cluster analysis, while classification may be used for monitoring plan effectiveness.

The study of this paper can be split into the following three parts

- Visualization of data from Maps instantly and converting them to clusters using APIs.
- Applying Density based Outlier spectral analysis to find density among various clusters.
- Classifying these clusters crowds based on an algorithm into various regions.

The proposed method shall be explained in detail based on these three steps.

3.1. STEP 1: DATA VISUALIZATION

A primary goal of data visualization [1] is to communicate information clearly and efficiently to users via the statistical graphics, plots, information graphics, tables, and charts selected. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines, circles or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the important step in data analysis or data science process. Data visualization [2] can take in many different forms. We will be using Google API in extracting the data from Maps using a JavaScript Library. The Google Maps JavaScript API uses libraries to provide supplemental features: the visualization library contains a number of classes that turn raw data into beautiful visualizations. The visualization library includes the following classes:

- The Heat map Layer class visualizes data intensity at geographical points. Our earthquake mapping tutorial uses the Heat Map Layer class to plot earthquake locations and intensity and walks you through each step of the code. The Visualization classes are a self-contained library, separate from the main Maps API JavaScript code. To use the

functionality contained within this library, you must first load it using the libraries parameter in the Maps API bootstrap URL:

```
<script type="text/javascript"src="https://maps.googleapis.com/maps/api/js?libraries=visualization"></script>
```

The great thing about data visualization is that the nature of the data is secondary. Data visualization fills that role in making large amounts of data understandable and actionable, provided that the visualization suits the data.

For example, geographical data (such as college places, marketplaces, or residential areas etc.) can be accessed through Google JS API. This type of data would have more impact as a bar chart or list than it would being represented on a map using a tool like Instant Atlas. The actual figures overlaid onto the map would provide too much information to be useful, so you could use the features in this tool to show quantities as colors to represent intensity. The underlying concept is that the data is the raw material, and there are a range of tools available for visualizing data. Since simple markers give little detail about that particular place, users are then shown how to represent each marker with a circle size or heatmap that is determined by the magnitude of the crowd present at that place at a particular instant.

Output of data visualization of data provides a graphical display which should

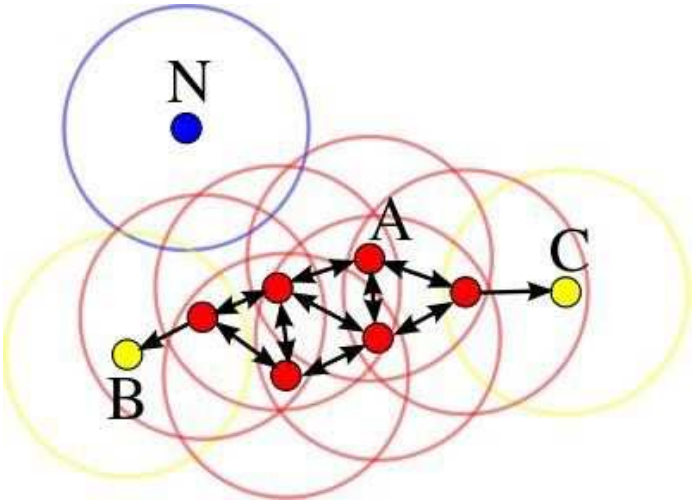
- Make large data sets coherent
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to the fine structure
- Be closely integrated with the statistical and verbal descriptions of a data set.

3.2. DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (DBSCAN) DBSCAN

is a data clustering algorithm : given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbours), marking as outliers points that lie alone in low-density regions (whose nearest neighbours are too far away).Consider a set of points in some space to be clustered. For the purpose of DBSCAN[3] clustering, the points are classified as core points, (density) reachable points and outliers, as follows:

- A point p is a core point if at least min Pts points are within distance ε of it, and those points are said to be directly reachable from p. No points are directly reachable from a non-core point.
- A point q is reachable from p if there is a path p1, … , pn with p1 = p and pn = q, where each pi+1 is directly reachable from pi (so all the points on the path must be core points, with the possible exception of q).
- All points not reachable from any other point are outliers.

Now if p is a core point, then it forms a cluster together with all points (core or non-core) that are reachable from it. Each cluster contains at least one core point; non-core points can be part of a cluster, but they form its "edge", since they cannot be used to reach more points.



In this diagram, min Pts = 3. Point A and the other red points are core points, because at least three points surround it in an ε radius. Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor density-reachable.

Reachability is not a symmetric relation since, by definition, no point may be reachable from a non-core point, regardless of distance (so a non-core point may be reachable, but nothing can be reached from it). Therefore a further notion of connectedness is needed to formally define the extent of the clusters found by DBSCAN. Two points p and q are density-connected if there is a point o such that both p and q are density reachable from o. Density-connectedness is symmetric.

A cluster then satisfies two properties:

1. All points within the cluster are mutually density-connected.
2. If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

3.2.1. ADVANTAGES

1. DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means.
2. DBSCAN can find arbitrarily shaped clusters. It can even find a cluster completely surrounded by (but not connected to) a different cluster. Due to the Min Pts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.
3. DBSCAN has a notion of noise, and is robust to outliers.
4. DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database. (However, points sitting on the edge of two different clusters might swap cluster membership if the ordering of the points is changed, and the cluster assignment is unique only up to isomorphism.)
5. DBSCAN is designed for use with databases that can accelerate region queries, e.g. using an R* tree.
6. The parameters min Pts and ε can be set by a domain expert, if the data is well understood.

3.2.2. DISADVANTAGES

1. DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data is processed. Fortunately, this situation does not arise often, and has little impact on the clustering[4][5][6] result: both on core points and noise points, DBSCAN is deterministic. DBSCAN* is a variation that treats border points as noise, and this way achieves a fully deterministic result as well as a more consistent statistical interpretation of density-connected components.
2. The quality of DBSCAN depends on the distance measure used in the function region Query (P,ε). The most common distance metric used is Euclidean distance. Especially for high-dimensional data, this metric can be rendered almost useless due to the so-called "Curse of dimensionality", making it difficult to find an appropriate value for ε. This effect, however, is also present in any other algorithm based on Euclidean distance.
3. DBSCAN cannot cluster data sets well with large differences in densities, since the min Pts-ε combination cannot then be chosen appropriately for all clusters.
4. If the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult.

5. Outlier[9]: A data object that deviates significantly from the normal objects as if it were generated by a different mechanism.
6. Outliers are different from the noise data.
7. Noise is random error or variance in a measured variable.
8. Noise should be removed before outlier detection.

Applications:

- Credit card fraud detection
- Telecom fraud detection
- Customer segmentation
- Medical analysis

3.2.3. TYPES OF OUTLIERS:

Three kinds: global, contextual and collective outliers

Global outlier (or point anomaly)

Object is Og if it significantly deviates from the rest of the data set.

Contextual outlier (or conditional outlier)

Object is Oc if it deviates significantly based on a selected context.

3.2.4. COLLECTIVE OUTLIERS

A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers.

Note: A data set may have multiple types of outlier. One object may belong to more than one type of outlier

3.2.5.CHALLENGES OF OUTLIER DETECTION

Modelling normal objects and outliers properly

Hard to enumerate all possible normal behaviours in an application

The border between normal and outlier objects is often a grey area

3.2.6. APPLICATION-SPECIFIC OUTLIER DETECTION

Choice of distance measure among objects and the model of relationship among objects are often application-dependent

E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations

Handling noise in outlier detection:

Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection.

3.2.7. PROXIMITY-BASED APPROACHES: DISTANCE-BASED VS DENSITY-BASED OUTLIER DETECTION

Objects that are far away from the others are outliers[7].

Assumption of proximity-based approach: The proximity of an outlier deviates significantly from that of most of the others in the data set

Two types of proximity-based outlier detection methods

Distance-based outlier detection: An object o is an outlier if its neighbourhood does not have enough other points

Density-based outlier detection: An object o is an outlier if its density is relatively much lower than that of its neighbours

3.2.8. DISTANCE-BASED OUTLIER DETECTION

For each object o, examine the number of other objects in the r-neighbourhood of o, where is a user-specified distance threshold. An object o is an outlier if most (taking pas a fraction threshold) of the objects in D are far away from o, i.e., not in

the r-neighbourhood of o. An object o is a DB(r, p) outlier if equivalently, one can check the distance between o and its k-th nearest neighbour ok, where o is an outlier if dist(o, ok) > r

Efficient computation: Nested loop algorithm

For any object oi, calculate its distance from other objects, and count the number of other objects in the r-neighbourhood. If p n other objects are within r distance, terminate the inner loop

Otherwise, oi is a DB(r, p) outlier Efficiency: Actually CPU time is not O(n2) but linear to the data set size since for most non-outlier objects, the inner loop terminates early.

3.2.9. DENSITY-BASED OUTLIER DETECTION

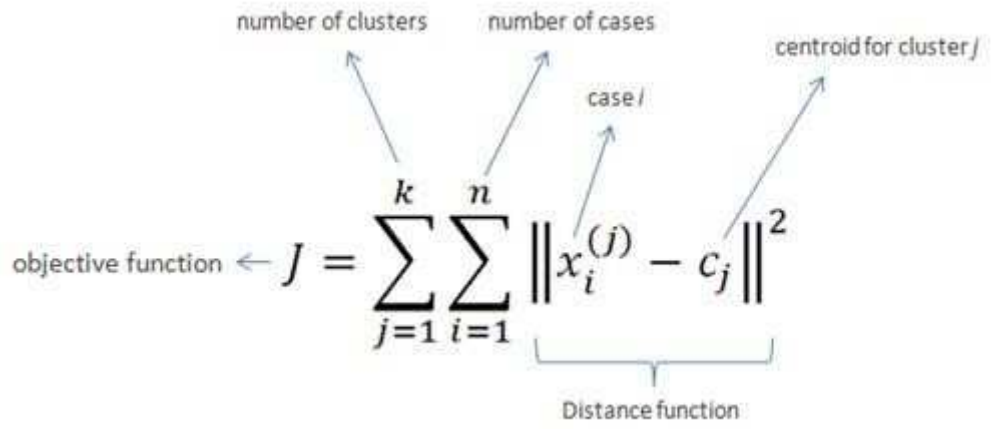
Local outliers: Outliers comparing to their local neighbourhood, instead of the global data distribution.

The density around an outlier object is significantly different from the density around its neighbours.

Method: Use the relative density of an object against its neighbours as the indicator of the degree of the object being outliers.k-distance of an object o, distk(o): distance between o and its k-th NNk-distance neighbourhood of o, Nk(o) = {o'| o' in D, dist(o, o') =distk(o)}

Nk(o) could be bigger than k since multiple objects may have identical distance to o. Whereask-means clustering

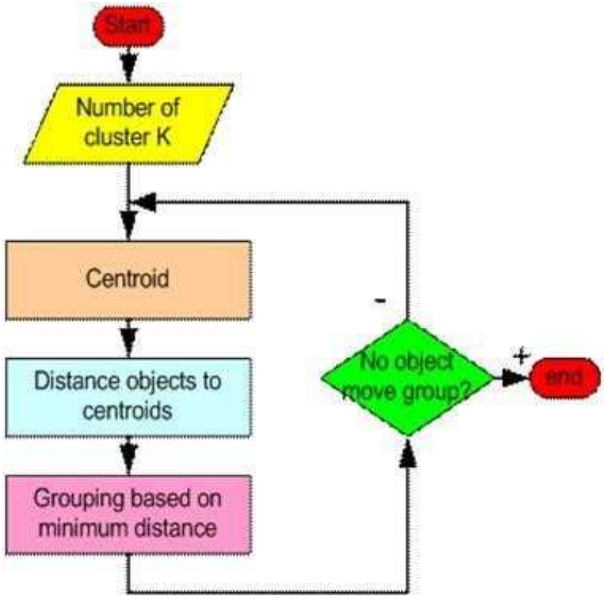
K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



The basic step of k-means clustering [8][10] is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence Iterate until stable (= no object move group)

- Determine the centroid coordinate
- Determine the distance of each object to the centroids
- Group the object based on minimum distance



Traffic-Classification Algorithm:

1. Check for the regions score (weight of edges) in each and every clusters by getting the K-distance for each outliers.
2. Formulate all the scores under a group-by theory and remove the duplicates and unnecessary clusters from it.
3. Using the shortest distance algorithm, project only those distances first which are or will be useful to the object.
4. Classify those several clusters weights into different regions.
5. If the score>0(greater than Zero) advise it to go positive and correct.
6. If the score<0(less than Zero) neglect it.
7. If the score lies in between:-
   - 30-50: (check Step 4 & 5)Respond it to be a normal free Traffic.
   - 50-70: (check Step 4 & 5) Respond it to have regions like schools, colleges, residential areas or some offices.
   - 70-100 :(check Step 4 & 5) Respond it to be a marketplace region with a heavy traffic.
8. Provided if, the above three are varying , then we can classify them based on the past results.(Externally given from ML algorithms)
9. Also if, there lies changing of scores rapidly then, it can predict that it is heavy moving traffic.
10. Based on above classifications, the driver can control & proceed its vehicle accordingly in order to avoid traffic.
11. If any of these is violated by the driver, then the system will give an strong alarm on doing so and ask it to take actions appropriately.

4. APPLICATIONS

- Used in condensing the large heavy traffic and making required changes effectively.
- Maps classification gets more synthesized based on its correct output.
- Reduce accidents to limit, if taken care by drivers and responding to situations holistically.
- Classification can help government, private bodies or construction authorities to plan and devise their ideas accordingly.

REFERENCES

[1] [Online] Available: http://www.cse.gatech.edu/content/data-visualization
[2] [Online] Available: https://developers.google.com/maps/documentation/javascript/visualization
[3] [Online] Available: https://en.wikipedia.org/wiki/DBSCAN
[4] [Online] Available: http://scikit-learn.org/stable/modules/clustering.html
[5] [Online] Available: https://charlesmartin14.wordpress.com/2012/10/09/spectral-clustering/
[6] [Online] Available: https://apandre.wordpress.com/visible-data/cluster-analysis/
[7] Yunxin Tao ; Coll. of Inf. Sci. & Technol., Nanjing Univ. of Aeronaut. & Astronaut., Nanjing ; Dechang Pi," Unifying Density-Based Clustering and Outlier Detection," Knowledge Discovery and Data Mining, 2009. WKDD 2009, 23-25 Jan. 2009 Moscow, pp. 644-647.
[8] [Online] Available: http://www.edureka.co/blog/k-means-clustering/
[9] [Online] Available: https://www.google.co.in/?gws_rd=ssl#q=outlier+analysis+in+data+mining
[10] [Online] Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html