Team Rocket (150): PIKA – Program for Imagining a Klustering Algorithm

York Delloyd, Philip Edwards, Yongquan Tan, Hongji Wang, Bowen Yang

OMSCS, Georgia Institute of Technology

CSE 6242: Data Visual Analysis

Dr. Polo Chau

27 March 2020

PIKA – Program for Imagining a Klustering Algorithm

Introduction

PIKA, Program for Imagining a Klustering Algorithm, is designed to be an application where a user can visualize, on a map, the clusters of various major geographical areas and have the unsupervised learning model group locations. Since we know that urban areas can be successfully classified through different structural dimensions [8], performing the unsupervised version - clustering, should also be a feasible task. We will perform clustering so that a person can identify areas he/she is more familiar with. In addition, PIKA will also allow for users to customize what variables and factors they are interested in having considered in the algorithm.

The key feature of PIKA is the flexibility for the user to not only select what they care about, but to also have an input on how the data model handles their data to create a customized model just for the user.

Problem Definition

Currently, people who are looking into new areas to live are only able to see generic rankings for different cities. One issue that arises is that not everyone values different attributes of a city the same. For instance, an elderly couple would not be as interested as how good the school system is for an area or a younger couple would not be as invested in quality of health care as an elderly couple. Because of this lack of customization, users are gridlocked into the general top rankings. We want to provide an alternative that is personalized toward any and every user.

Survey

Since urban areas can be successfully classified through different structural dimensions [8], performing the unsupervised version - clustering, should also be a feasible task. We want to define similar cities without ranking and just grouping them together based on a customized attribute pool. We can do this because there exist similarities that describe such behavior in that… "[d]espite the great geographical distances, the rents and vacancies of some east and west coast cities tend to move together." [2]

Current approaches primarily focus on ranking cities and locations based off of similar cities and then adjusting based on a variety of factors [7]. One approach involved manually clustering cities based upon a limited number of attributes, such as sales price, liquidity, trading volume… etc. [5,6]. These clusters also demonstrate an inherent ranking of cities. In 1988, L.R Klein defined other attributes that determine desirability to live, such as location, education, and crime [15]. Clustering has a wide variety of uses as demonstrated in a 2017 study where it was used to define different urban morphological zones [16].

Another approach to consider k-means, where k centroids are initially randomly chosen for all nodes. After each iteration, the data points are shuffled between centroids until an optimum clustering is determined by minimizing sum of squared distances [10, 13]. However, a problem is that our dataset may contain categorical data, so K-Means may be problematic. An alternative is K-modes, which takes the mode of a cluster instead of mean [3]. This is useful in creating clusters from categorical data, where delta doesn't account for the difference in values [3]. Another approach for K-means on categorised variables by applying advanced different dissimilarity measures[11]. We are also investigating other clustering algorithms such as spectral clustering which treats data points as nodes on a graph [9, 10]. This could prove exceptionally important as we understand the "shape" of the clusters.

An alternative approach outside of clustering is to derive an index from our attributes, much like how a Constant-Utility Cost of Living index is derived by L.R Klein in 1947 [14]

Proposed Method

Our approach is to shed the usual approach of ranking cities and embracing the idea that cities are not better or worse, just similar and dissimilar. PIKA will utilize unsupervised learning, specifically clustering, to group cities together. This approach presents a number of new challenges such as explaining to the end user why groups of cities are similar and why they were grouped together.

Our first step is to create a dataset by combining together US census data and data from other federal agencies in order to describe each city. We plan on defining cities as US Metropolitan Statistical Areas (MSA) & will use the top 100 MSA by population size. Different data sources are aggregated at different levels, this introduces a lot of challenges when combining data sets. For example Unified Crime Reporting Statistics are aggregated at a police agency level. This requires us to use a mapping from police agency to MSA created by the Inter-university Consortium for Political and Social Research. Other census data is aggregated at a county level requiring us to utilize a mapping from counties to MSA.

We are creating a database to help us map together different datasets. When we have combined all of our data sources then we will decide whether we need to host the database in the backend or if we can use a flat file.

From here we will test out multiple different clustering algorithms such as k-means clustering, spectral clustering, & agglomerative clustering. We selected three clustering algorithms with different assumptions about the geometry of the data. K-means clustering uses the distance between points, spectral clustering uses graph distance, and agglomerative clustering uses pairwise distance.

While we cannot measure the validity of a cluster there are metrics we can use, such as Davies-Bouldin (DB) Index, Dunn Index, & Silhouette Coefficient to compare the relative quality of the clusters. The DB Index follows the intuition that good clusters are well-spaced and dense. Dunn Index follows a similar intuition as DB Index however focuses on the worst case, the clusters that are closest together and the least dense cluster. Silhouette Coefficient calculates how well assigned a point is to its cluster. Silhouette Coefficient is very expensive to compute. We hope that calculating 100 cities (100 points) is manageable however if it is too expensive we will either take a random sample of points, or choose to not use this metric.

We will use these metrics to decide upon a single clustering algorithm to productionalize & use in the hosted application. The challenge now shifts to visualizing the results to the end user and showing why cities are grouped together. We plan on experimenting with multi-dimensional visualization techniques to visualize the relationship across attributes that created clusters.

A core feature of PIKA is a user's ability to select the most important attributes to them. This requires PIKA to train new clusters based upon the user's selection. Online training presents technical requirements on our server. If necessary we will limit the number of cities and/or the number of attributes a user can select in order to limit training time and the user experience.

Regarding our user interface, we are making our web page easy to use with the user being only responsible for determining the weights of all available attributes. With it, the user can just select and submit and then wait for their results.

Experiments/Evaluation

PIKA has 4 main components: data gathering, model prediction, backend, and frontend. The data gathering will consist of obtaining data from the US census at https://www.census.gov/data/developers/data-sets.html as well as other other federal agencies such as the FBI.

Our approach to grouping similar cities is clustering based, therefore there is no idea of right versus wrong. We will analyze different clustering algorithms such as k means clustering, spectral clustering, & agglomerative clustering. As mentioned above, while measuring the validity is not possible, we will be taking measures of Davies-Bouldin Index, Dunn Index, & Silhouette Coefficient to compare the quality of the clusters.

The backend will be a Flask application with the following endpoints: get_attributes, process_attributes, contact_us, and get_documents. The frontend will be a reactJS application that will call on the backend. Both of these components will be hosted on the cloud in a Heroku server - allowing users to test our application.

With PIKA, we are setting to answer three main questions. What is the data that is relevant to our model in that it's an attribute that a user can interact with and might make a difference in clustering cities. How will users be affected by being able to have a say in what the data model shows them. And lastly, we want to know if we can visualize these clusters in a meaningful way where the user scope of why these clusters are generated.

In order to analyze how the user will be affected by being able to have a say in their model, we mainly have to perform a survey where we show several users US New's 2019 best places to live and have them utilize our tool. Afterwards, we can compare the users' satisfaction regarding the results. In addition, we can compare the number of times our clusters pick up the top locations presented by US News. We are trying to measure 1. user satisfaction and 2. the number of times our clusters pick up top cities.

Because of the scale of our experiment, the only controls we can have in place are the compare to list (US News) and a constant number of attributes within our tools. While reproducing, the experimenters should collect data regarding attributes and weights for each of their users. This will also be a good indicator that each city weighs differently for each user. We hope to see that top cities in the US News ranking do not show up in our most desired clusters a majority of the time thus backing the hypothesis that each user's attributes will garner different results.

We are currently working on building out our visualization of the data and will attempt to display the data. <UnderConstruction />

Innovation

With PIKA, we are trying to group together the desirability to live in certain areas. US News Ranking, in 2019, used a methodology that included analyzing the Job Market Index, Value Index, Quality of Life Index, Desirability Index, and Net Migration to determine that Austin, TX was the most desirable place to live in the United States [1]. Our innovation is that we want to allow the user to be able to interact with the model they are seeing and be able to directly affect its outcome providing a more customizable experience. One current approach ranks happiness of cities using several different factors such as income equality, employment, commuting, housing, density, age, and climate, but an article cites contradictory results of each factor [4]. What this means for us is that different factors mean different things for each person and allowing a user to have a say in the tool is novelty we are trying to provide. In addition to allowing users to choose what attributes they want accounted for, we are going to allow users to weigh how important that attribute is giving another level of granularity. With PIKA, it's not about what's best or better but rather what's best for YOU.

Plan

Task	Person Assigned	Estimated Completion Time	Updated Completion Time	Status
Generate robust K-Means clustering capability	Philip Edwards	1.5 weeks	1.5 weeks	Completed
Generate robust Spectral clustering capability	Philip Edwards	1.5 weeks	1.5 weeks	Pending
Generate robust Agglomerative clustering capability	Philip Edwards	1.5 weeks	1.5 weeks	In Progress
~~Clean and normalize data~~ Clean, normalize, and collect data	Philip Edwards	~~1.5 weeks~~	2 weeks	In Progress
~~Create Django Backend with API capabilities~~ Create Flask Backend with API capabilities	Yongquan Tan and Bowen Yang	~~2.5 weeks~~	1.5 weeks	Completed
~~Create connection between Django Backend and clustering capabilities~~ Create connection between Flask Backend and clustering capabilities	Yongquan Tan	2.5 weeks	2.5 weeks	In Progress
Backend Unit Testing	Yongquan Tan	1 week	1 week	Pending
Setup Front-End Web Application	Hongji Wang	1 week	1 week	Completed
Create API Connections with the Backend and Data sharing	York Delloyd and Hongji Wang	1 week	1 week	Pending
Front End Unit Testing	York Delloyd and Hongji Wang	1 week	1 week	Pending
Create User Interface for input data and adjusting variables	York Delloyd and Hongji Wang	4 weeks	4 weeks	In Progress
Create Return Data Display	York Delloyd and Hongji Wang	~~4 weeks~~	3 weeks	In Progress
Document progress and submitting said documentation throughout the project	York Delloyd	2 weeks	2 weeks	In Progress
Onboard the Backend Application to ~~AWS~~ Heroku	Bowen Yang	1 week	1 week	Completed
Onboard the Front-End Application to ~~AWS~~ Heroku	Bowen Yang	1 week	1 week	Pending
Quality Analysis and User Acceptance Testing	Bowen Yang	1 week	1 week	Pending
Application Documentation and Replication Steps	Bowen Yang	1 week	1 week	Pending
~~Building the dataset~~	~~Team~~	~~1 week~~		Moved
~~Creation of Posters and Media~~	~~Team~~	~~2 weeks~~		Canceled

Conclusion and Discussion

<Under Construction />

Final Notes

All team members contributed a similar amount of effort.

Works Cited

[1] How We Rank the Best Places to Live & Retire. (n.d.). Retrieved from https://realestate.usnews.com/places/methodology

[2] Goetzmann, W. N., & Wachter, S. M. (1995). Clustering Methods for Real Estate Portfolios. Real Estate Economics, 23(3), 271–310. doi: 10.1111/1540-6229.00666

[3] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. DMKD, 3(8), 34-39.

[4] Florida, R., Mellander, C., & Rentfrow, P. J. (2013). The Happiness of Cities. *Regional Studies*, *47*(4), 613–627. doi: 10.1080/00343404.2011.589830

[5] Clustering Minnesota Cities. (2005, September 22). Retrieved from https://www.lcc.leg.mn/lga/Background/clustermethodology.pdf

[6] Geltner, D., MacGregor, B. D., & Schwann, G. M. (2003). Appraisal smoothing and price discovery in real estate markets. Urban Studies, 40(5-6), 1047-1064.

[7] Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. Organizational behavior and human decision processes, 39(1), 84-97.

[8] Francisco J. Goerlich Gisbert, Isidro Cantarino Martí & Eric Gielen (2017) Clustering cities through urban metrics analysis, Journal of Urban Design, 22:5, 689-708, DOI: 10.1080/13574809.2017.1305882

[9] Ng, A. Y., Jordan, M. I., & Weiss, Y. (n.d.). On Spectral Clustering: Analysis and an algorithm. Retrieved from http://papers.nips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf

[10] R. Xu and D. C. Wunsch, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, Institute of Electrical and Electronics Engineers (IEEE), May 2005. https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=1763&context=ele_comeng_facwork

[11] Chandrasekhar, T. , Thangavel, K. , Elayaraja, E. Effective clustering algorithms for gene expression data. Int J Comput Appl. 2011; 32(4): 25–9.

[12] Beata Calka, “Estimating Residential Property Values on the Basis of Clustering and Geostatistics ”, Received: 7 February 2019; Accepted: 21 March 2019; Published: 24 March 2019

[13] Aristidis Likas, et al. “Pattern Recognition.” *The Global K-Means Clustering Algorithm*, vol. 36, no. 2, 2003, pp. 451–61.

[14] L. R. Klein , and H. Rubin. *A Constant-Utility Index of the Cost of Living*. Oxford University Press, 1947–1948, p. 4.

[15] Findlay, Allan, et al. “Where to Live in Britain in 1988.” *Copyright © 1988 Published by Elsevier Ltd*.

[16] Goerlich Gisbert, F. j., Martí, I. C., & Gielen, E. (2017, April 10). Clustering cities through urban metrics analysis. Retrieved from https://doi.org/10.1080/13574809.2017.1305882