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

25 February 2020

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 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.

All in all, 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

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]

PIKA – What's new?

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]. But what if you cannot move to Austin because of work, family, or even travel restrictions? With PIKA, we will expand upon the US News attributes and introduce new attributes to find similar cities through clustering instead of rankings.

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]. Because of this, PIKA will also focus on allowing the user to customize their experience as factors may not weigh equally for everyone. The novelty of PIKA is that it will allow users to specify the attributes important to them and the visualize which cities are similar to cities they know well or have an interest in moving to. With PIKA, it's not about what's best or better but rather what's similar though your eyes.

Affected Parties

Any user of PIKA will be able to, not only quickly see which cities are similar to theirs without having to do much research but also will be able to personalize what they mean by similarities. Users of our application will be able to make life changing decisions such as moving with peace of mind knowing what they are getting into. In addition, since users are able to adjust the variables, they will be able to personalize what attributes matter to them – this will be something that different users of PIKA will care to use, as an older couple with no children might not care about the school district as compared to their younger counterpart.

Application Impact

With PIKA, we aim to demonstrate how cities are not necessarily better or worse, but rather they are similar or dissimilar in the eyes of the beholder. If successful, users of PIKA will respond positively to prompts regarding social impact. That is to say, the key would be to survey people who have lived in two similar or dissimilar areas and have them discern the similarity index of the two locations with threshold determining if our application is proving accurate clustering. In addition, user response from people who have made the move using our application will be what can be used to determine if the application, overall, is successful.

Cost/Benefit Analysis

Given that the core of the application is unsupervised learning, even though PIKA allows for customization of variables, there lies a risk that the application will provide a generalization that may not fit each individual taste as the only thing the application knows is what the user provides. If a user determines, arbitrarily, that they do not care about the crime rate of an area, related factors may be omitted in calculations. Additionally, it can be difficult to discern the relationships within the clusters and therefore would not provide the user with exactly how each are similar or different.

The payoff of the application would be a shift in the moving mindset. With PIKA, moving is not going to a new and unfamiliar place, but rather just relocation of your possessions.

Cost

We expect each developer to devote 8 hours a week until the project due date. We will be volunteering our time through the class and will be able to use free Azure credits. On the open market the developer dollars would be $20,880 (~$87/hr * 5 developers * 8 hrs/week * 6 weeks) to create the final version. Hosting the application will cost ~$1,200/year. Assuming hosting the site for 1 year plus development costs we estimate $22,080.

Time Estimation

The project is estimated to be completed within six weeks with one week of presentation prep. Please see Appendix Figure 1 for the task breakdown.

Checkpoints for Success

At the Midterm stage, 3/27, we expect to be done with local application setup and have a working development environment: including the ability to pass data between our front and back ends. By the Final stage, 4/17, we expect for visualization to be accurate and the clustering algorithm to be able handle dynamic inputs. Each week we will have a checkup on our assigned epics each with a designated completion date.

Final Notes

All team members contributed a similar amount of effort for this proposal.

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

Appendix

Figure 1 – Time Breakdown

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