



Constrained k -Means Clustering Validation Study

Southwestern Oklahoma State University

Nicholas McDaniel | Dr. Stephen Burgess | Dr. Jeremy Evert | Department of Computer Science and Engineering Technology

Abstract

Machine Learning (ML) is a growing topic within Computer Science with applications in many fields. One open problem in ML is data separation, or data clustering. Our project is a validation study of, "Constrained k -means Clustering with Background Knowledge" by Wagstaff et. al. Our data validates the finding by Wagstaff et. al., which shows that a modified k -means clustering approach can outperform more general unsupervised learning algorithms when some domain information about the problem is available. Our data suggests that k -means clustering augmented with domain information can be a time efficient means for segmenting data sets. Our validation study focused on six classic data sets used by Wagstaff et. al. and does not consider the GPS data of the original study. We have published our code on a public SWOSU Github repository to enable other researchers to use our code as a starting point. Validation studies such as this provide great learning opportunities for students interested in working with Machine Learning, Artificial Intelligence, and other related applications. This research was funded in part by the Dr. Snowden Memorial Scholarship with the NASA OKLAHOMA Space Grant Consortium. This material is based upon work supported by the National Aeronautics and Space Administration issued through the Oklahoma Space Grant Consortium.

Project Summary

Validation studies can be used to introduce students to a topic area while gaining valuable experience and insight into the topic at hand.

For example, this project serves as an introduction to machine learning and some of the research related to data clustering. The goal for the researchers is to increase familiarity with popular tools (eg. MatPy) and fundamental concepts. An emphasis was placed on projects related to current research trajectories for NASA.

Conclusion

This project serves as an introduction to machine learning and some of the research related to data clustering. Additionally, this project served as a way to introduce myself into the topic, which I feel was a more effective way of learning the material with a set boundary in which to learn the different tools.

Discussion

This project serves as an introduction to machine learning and some of the research around data clustering. The goal for the researchers is an increase familiarity with popular tools and fundamental concepts. An emphasis was placed on projects related to current research trajectories for NASA.

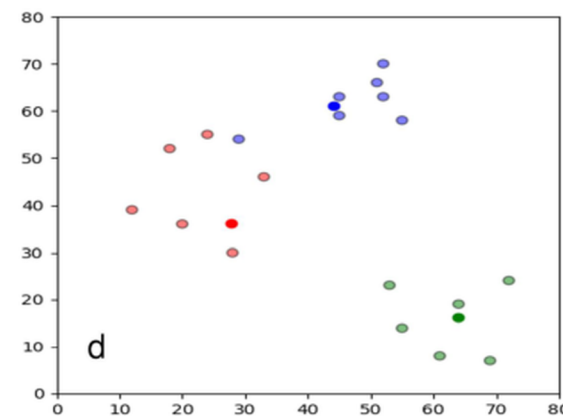
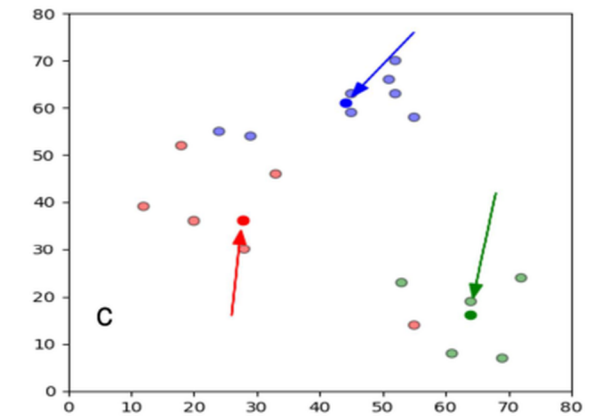
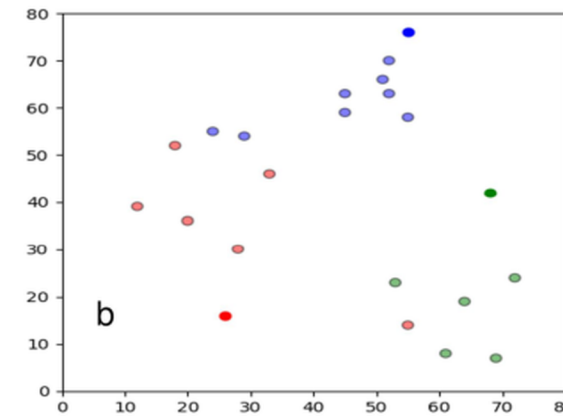
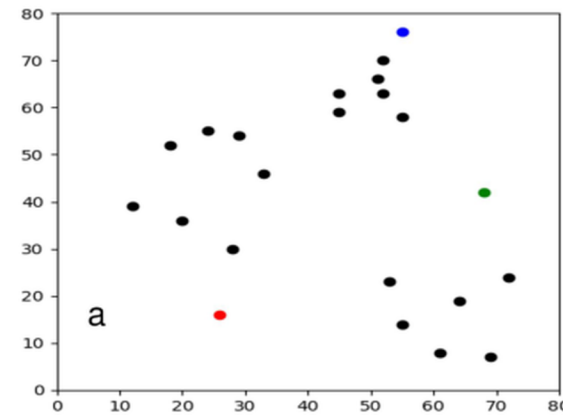
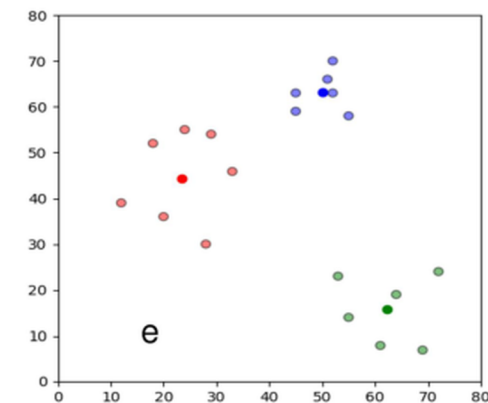


Figure Description

This set of graphs shows an example of how k -means are found. A shows that there are three different centroids chosen in the color, and then the centroid will have the nearest colors assigned to them, as shown in B. After this, in C it shows the centroids moving to the average of the colors assigned to them. Then it will repeat until it won't move at all, ending up with E.



Related Literature

Wagstaff et. al. found that in a clustering algorithm, the more constraints that are added, k -Means the more accurate and efficient the algorithm becomes. The following studies have expanded this work.

In "Integrating Constraints and Metric Learning in Semi-Supervised Clustering" by Bilenko et. al., the constraints are increased by a completely separate algorithm to better increase the accuracy of clustering grouping.

In "Constrained k -Means Clustering" (Bradley et. al.), we see a validation of the Wagstaff et. al. assertion that the more constraints added to the k -Means clustering algorithm, the more accurate it will become. This study was completely separate from Wagstaff et. al., but resulted in nearly identical findings.

Finally, in a "Survey of Clustering Algorithms" (Xu & Wunsch), there is an extremely in-depth look into the use of constraints with different clustering algorithms in various situations, ranging from the traveling salesman problem, bioinformatics, and different datasets.

Future Work

This project will be evolving over the next few years. It will continue to follow the papers as outlined by Wagstaff et. al. The researchers will create a python script that will be able to take in any sort of spreadsheet or dataset and be able to find the k -means of that dataset.

Author Biography

Nicholas McDaniel is a SWOSU Computer Science Senior. Nick has been working as a NASA student researcher for three years. Nick has experience as a Cluster Computer systems administrator. Nick demonstrated a love for working on projects. He has been integral in setting up several different networking channels for the university.

Works Cited

Raso, Sugath, Mikhail Bilenko, and Raymond J. Mooney. "A probabilistic framework for semi-supervised clustering." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 59-68. ACM, 2004.

Bradley, P. S., K. P. Bennett, and Aylan Demirci. "Constrained k -means clustering." *Microsoft Research, Redmond* (2000): 1-8.

Wagstaff, Kim, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k -means clustering with background knowledge." In *ICML*, vol. 1, pp. 577-584. 2001.

Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." *IEEE Transactions on neural networks* 16, no. 3 (2005): 645-678.