A Geospatial Implementation of a Novel Delineation Clustering Algorithm Employing the K-means

Tonny J. Oyana

Department of Geography,
Southern Illinois University
1000 Faner Drive, MC 4514,
Carbondale, IL 62901-4514
Tel: +1 618-453-3022
Fax: +1 618-453-6465
Email: tjoyana@siu.edu

Abstract
The overarching objective of this study is to report the implementation and performance of a novel delineation clustering algorithm employing the k-means. This study explores a newly proposed algorithm designed to increase the overall performance of the k-means clustering technique—the Fast, Efficient, and Scalable k-means algorithm (FES-k-means*). The algorithm reduces the computational load and produce quality clusters. Resulting improvements reside in three major areas: 1) minimization of cluster number fluctuation; 2) efficient handling of large geospatial datasets; and 3) adequate analysis of large geospatial datasets, be it compact or scattered.

1. Introduction

Several algorithms are used to determine natural groupings or "clusterings" within a dataset. Of all the different forms of clustering, the improvements suggested in this study are for the unsupervised, partitioned learning algorithm of the k-means clustering method proposed by J. MacQueen (1967). MacQueen describes k-means as a process for partitioning an N-dimensional population into k sets on the basis of a sample. According to (Kanugo et al. 2002; Likas 2003) k-means is the most widely used and simplest form of clustering.

The FES-k-means* algorithm uses a hybrid approach that comprises the k-d tree data structure that enhances the nearest neighbor query, the original k-means algorithm, and an adaptation rate proposed by Mashor (1998). This algorithm was tested using two real datasets and one synthetic dataset. It was employed twice on all three datasets; once on data previously trained by the innovative MIL-SOM* (Mathematically Improved Learning-Self Organizing Map) method (Oyana 2006; Oyana et al. 2006), and secondly on the actual, untrained data. This two-step approach of data training prior to clustering provides a solid foundation for knowledge discovery and data mining, otherwise unclaimed by clustering methods alone.

Many attempts have been made to rid the conventional k-means algorithm of its limitations, namely 1) computational expensive nature for segregating large-scale datasets; 2) inaccurate cluster initialization; and 3) the local minima problem (Pelleg and Moore 2000). When used in conjunction with the MIL-SOM* training technique, FES-k-means* algorithm shows improvements in two of the aforementioned limitations. It reduces the computational load and produce quality clusters.

The benefits of this method are in four major areas: 1) it produces similar clusters as the original k-means method at a much faster rate than conventional methods as shown by runtime comparison data; 2) scalability is measured by its transferability to other platforms and its innateness to adequately handle small and large datasets; compact or scattered; 3) efficient analysis of large geospatial; and 4) the multiplicity in the algorithm makes it complex, yet innately stable and robust.

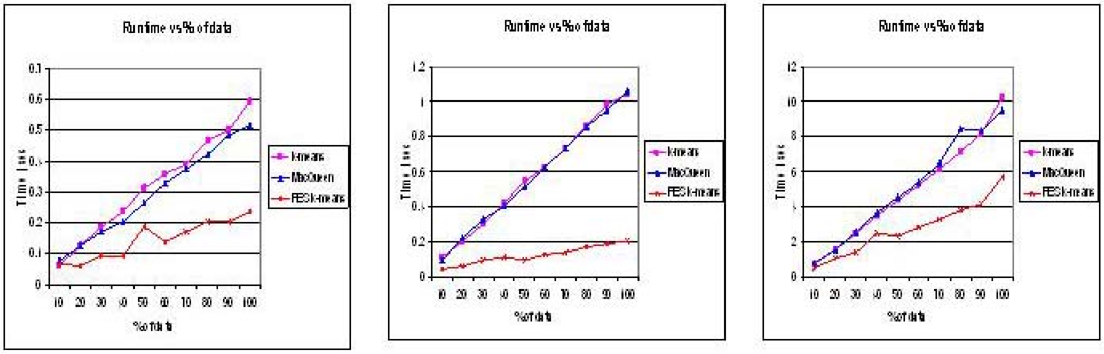
2. Data and Methods

The focus of this study is to evaluate the characteristics and assess the quality and efficiency of a new k-mean clustering technique, FES-k-means*. I invoked three distinct datasets to realize this goal; two published real datasets and one synthetic dataset. The real datasets were 1) georeferenced elevated Blood Lead Levels (BLL) in Chicago, Illinois; and 2) georeferenced physician-diagnosed adult asthma data for Buffalo, New York. There are two versions of each of the datasets: the raw data in its entirety, and the reduced MIL-SOM* trained version. The MIL-SOM* algorithm is essentially an improved version of the self-organizing map (SOM), an unsupervised neural network that is used to visualize high dimensional data by projecting it onto lower dimensions by selecting neurons or functional centroids to represent a group of valuable data.

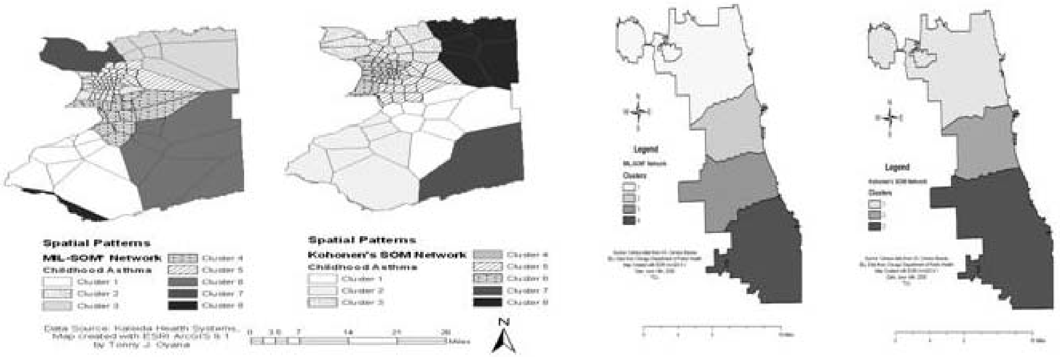
During experimentation, I assessed the performance of the FES-k-means* algorithm using three tasks: 1) evaluated speed efficiency using run-time; 2) evaluated mean square error for processed data; and 3) trained the data. I compared the FES-k-means* method with the standard k-means, and MacQueens k-means methods. MacQueen’s rendition is a convergent subsequence of the standard k-means. Using run-time, in seconds, speed efficiency was measured against percent of data processed for each of the three aforementioned clustering methods. The percent of data processed was randomly selected based on percentages that ranged from 10% to 100% and increased in ten percent increments (10%, 20%, 30%, etc.). To further explore clusters and outliers, field work and communal/ housing investigations in Chicago, Illinois was conducted. Photos have been provided to confirm findings for the Chicago BLL housing dataset, particular interest was paid to delineated clusters with outliers.

3. Results and Discussion

Results are presented in Figures 1 through 3. All test plots revealed the same general characteristics: an increasing linear relationship of the data processed and runtime, and similar positioning of the standard k-means and MacQueen methods for all three tests. From these plots, it is clear that FES-k-means*, which employs Mashor’s updating rate, has a faster convergence than the standard k-means and MacQueen’s methods. Each test plot yielded similar characteristics for the three methods: linearity of the data processed and mean square error, an increasing relationship between the percent of data processed and mean square error, linearity of the number of centers and mean square error, and parallel similarities between the three tested methods—the standard k-means, MacQueen, and FES-k-means*.



Figures 1a to 1c: Elevated BLL Linked with Age of Housing Data (left panel); Adult Asthma dataset (middle panel); synthetic dataset (right panel). Comparison of three k-means algorithms using runtime versus percent of data processed.



Figures 2a and 2b: left panels show childhood asthma dataset and right panels show elevated blood levels. Comparison of Kohonen's SOM and MIL-SOM* (classes delineated using FES-k-means*).

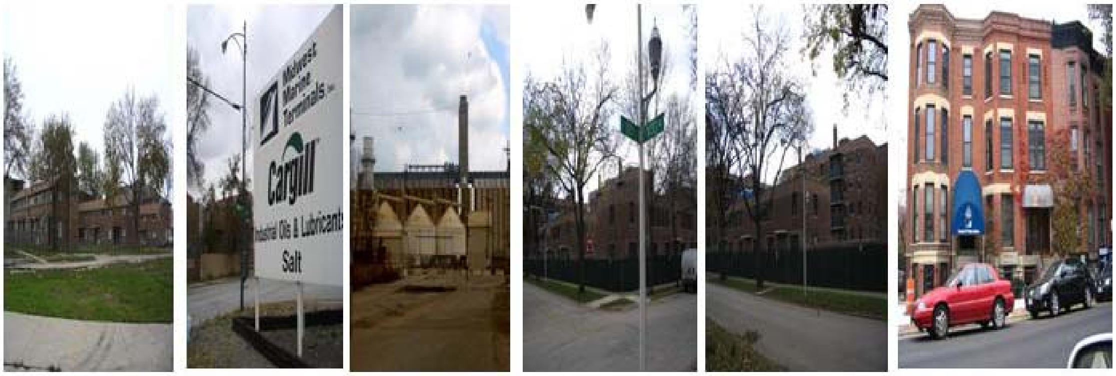


Figure 3 shows photos depicting housing conditions in the City of Chicago (taken in November 2006). From the left to right panels: the first five photos were taken in the west and southern part of city and the last photo was taken in the North part of the city. Outliers were detected in each class using the box plot.

Implementation of the suggested algorithm during the k-means clustering procedure has proven to be efficient in each of the problem areas mentioned. This improvement in the traditional k-means clustering method, allows for an even more efficient tool for visualizing and mining vast and extensive datasets. The improved k-means algorithm (FES-k-means) has provided a better result than the original k-means algorithm, which delineates cluster boundaries based on the best DBI validation.

In comparison with the MIL-SOM* trained data, I find that both, the trained and untrained datasets, returned comparable major clusters due to the robust nature of the FES-k-means* algorithm. The clusters for the MIL-SOM* trained data captures the clusters of the full dataset that it represents. Because, the untrained data is plagued with noise and outliers, initializing the clusters using the MIL-SOM* algorithm enables effective management of these outliers. The FES-k-means* clustering employed on MIL-SOM* trained and untrained data display similar clustering characteristics for each of the test situations. Data training prior to clustering affords the establishment of the k-value based on visual heuristics provided by MIL-SOM*.

Not only can FES-k-means* be utilized in the SOM toolbox, the efficiency and robustness of the algorithm enables it to be used on multiple platforms and program applications. For example, the original code was written in C, and then it was exported to work in Matlab and SOM neuron-computational environments. Also, due to its scalability we find that FES-k-means* yields equally successful clustering results on small or large datasets; unlike the fast global k-means algorithm in which the quality begins low and gradually increases as the number of clusters increase with larger cluster separation, and smaller dimensionality.

4. Conclusion

The engaging implementation and performance of the FES- k-means* clustering algorithm when applied to three different geospatial datasets undoubtedly reveals great potential. Implementation of the algorithm has so far been successful and the main benefits have been established. This study is, however, continuing to explore these potentials, along with other properties, while anticipating additional benefits. The FES-k-means* algorithm has provided a better result than the original k-means algorithm. Further research is also required in four key areas: 1) how to explore and handle outliers; 2) how to evaluate resulting clusters; 3) how to measure reliability of results; 4) and how to effectively display more than three dimensions. Presently, we are only able to classify multidimensional large geospatial datasets; however, the challenge is with the visual display of datasets containing more than three dimensions.

5. Acknowledgements

Supported in part by SIUC Faculty Start-up Fund and SIUC ORDA Faculty Seed Grant. Dharani Babu Shanmugam and Haricharan Padmanaban Computer Programmers were responsible for code implementation. Special thanks to Ms. Kara Scott, NSF Research Fellow and Dr. Jamson S. Lwebuga-Mukasa.

6. References

Kanungo, T., D.M. Mount, N. S. Netanyahu, et al. 2002. An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on pattern analysis and machine intelligence. 24(7): 881-92.

Likas A., N. Vlassis, and J.J. Verbeek. 2003. The global k-means clustering algorithm. Pattern Recognition. 36: 451-61.

MacQueen, J.B. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Statistical Laboratory, University of California. 282-97.

Mashor, M.Y. 1998. Improving the performance of k-means clustering algorithm to position the centers of RBF network. International Journal of the Computer, The Internet and Management. 6(2).

Oyana TJ. Introducing an improved fast and computationally-efficient SOM algorithm—MIL-SOM*: issues and benefits for large-scale geospatial data. In M. Raubal, H. Miller, A. Frank, and M. Goodchild, (eds.) Geographic Information Science. In Proceedings of the Fourth International Conference, GIScience 2006, Münster, Germany, September 20–23, 2006. pp.141–145.

Oyana TJ, Achenie LEK, Cuadros-Vargas E, Rivers PA, and Scott KE. 2006. A Mathematical Improvement of the Self-Organizing Map Algorithm. Chapter 8: ICT and Mathematical Modeling (pp 522–531). In Mwakali J.A. and Taban-Wani G. (eds.): Advances in Engineering and Technology, London: Elsevier Ltd, pp.847.

Pelleg, D., & A.W.Moore. 2000. X-means: Extending K-means with efficient estimation of the number of clusters. In Proceedings of the 17th International Conference on Machine Learning (ICML'00). 727–34.