A Deeper Look into Scalable Methods for Creating Food Groups Using NHANES Dataset

Samuel Jones, Michela Taufer, Michael Wyatt, and Nigel Tan
Tickle College of Engineering, University of Tennessee, Knoxville, Tennessee

Abstract

Every year there are surveys conducted by the National Center for Health Statistics to analyze the health and nutritional intake of children and adults in the United States. These surveys are known as the National Health and Nutrition Examination Survey (NHANES) and gathers its data by getting participants to record their food intake and all the information about the food they are eating. The dataset uses USDA food codes to classify each food categorically dependent on if it's a grain, dairy, etc. This project sets out to find a better way of grouping these foods by their nutritional value rather than what kind of food groups they're a part of. This classification is especially important for the NHANES dataset s the purpose of the survey is to conduct the health of the country. The NHANES dataset is also perfect for grouping the foods as it has all the nutritional values for each food. Having a way to group foods by their nutritional value is also helpful to consumers when they want to eat more healthfully or have dietary restrictions. This project is an extension of Michael Wyatt's project where I expand on his ideas of finding methods to distinguish how to group the food items.

Methodology

- Preprocess the data in order to parallelize computation
- Use DBSCAN clustering algorithms to group the foods together based on their nutritional value
- Use different correlation techniques to determine which items should cluster together and finding an appropriate epsilon area for each correlation using elbow method

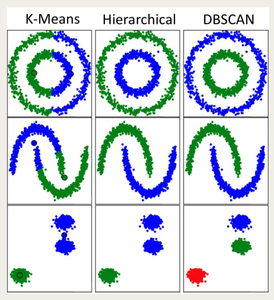

Figure 1 Examples of how common clustering algorithms perform under certain conditions

Feature	K-Means	Hierarchical	DBSCAN
Resource Efficient	Yes	No	Yes
Noise Insensitive	No	No	Yes
Outlier Detection	No	No	Yes
Spheroid Clusters	Yes	Yes	Yes
Non-Spheroid Clusters	No	Yes	Yes
Undefined Cluster Count	No	Yes	Yes

Figure 2 Showing the benefits of DBSCAN compared to other common clustering algorithms

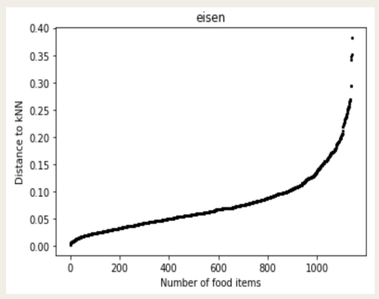

Figure 3 Example of how to use the elbow method to find a good epsilon value, for this example epsilon = 0.1

Preprocessing

- The data is set up into nutritional categories where each row represents a certain food item
- Need to normalize each food item and get rid of empty and redundant data
- Put into form to work with Pyspark efficiently

DR1IFDCD	DR1MC	DR1IGRMS	DR1IKCAL	DR1IPROT	DR1ICARB	DR1ISUGR	DR1IFIBE	DR1ITFAT
64132010	0	160	114	0.98	27.92	26.32	1.6	0.05
92101500	0	710.4	4	0.78	0	0	0	0.07
91200040	0	4	14	0.04	3.56	3.41	0	0
12120100	0	10.08	13	0.3	0.43	0.02	0	1.16
52215200	0	63.62	198	5.27	32.67	1.22	2	4.93
22600100	0	10	54	3.7	0.14	0	0	4.18
71000100	0	20.13	21	0.33	3.89	0.17	0.4	0.49
72201200	0	30	15	0.69	2.09	0.4	1	0.65

Figure 4 Table showing part of what the raw data from the NHANES dataset looks like

Metrics and Correlation Factors

Manhattan Distance

- Similar to Euclidean distance
- Not good for multidimensional data analysis
- Clusters have very large diameter

$$d_{man}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)|$$

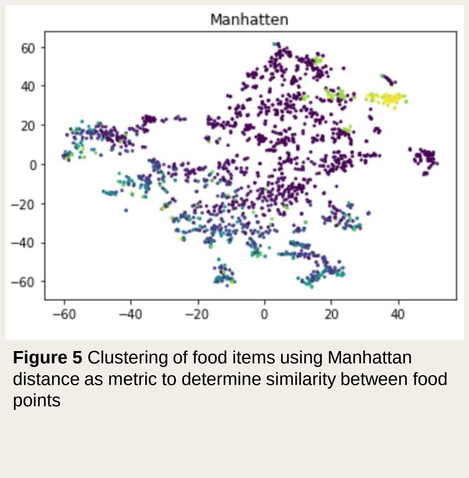

Figure 5 Clustering of food items using Manhattan distance as metric to determine similarity between food points

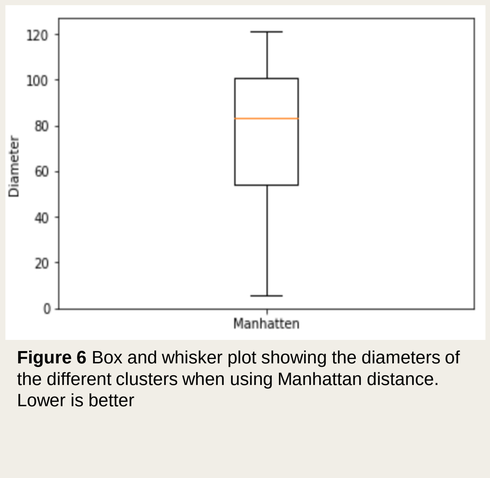

Figure 6 Box and whisker plot showing the diameters of the different clusters when using Manhattan distance. Lower is better

Acknowledgements:

The poster and research was made possible because of the past work of Michael Wyatt which is what this project is based on and the results it is trying to replicate

Pearson Correlation

- Shows how data relates to each other linearly making it better for multidimensional data
- Clusters have smaller diameter which means better clustering

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

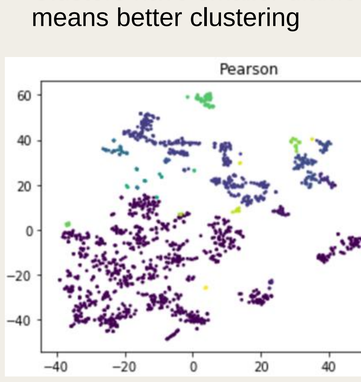

Figure 7 Clustering of food items using Pearson correlation as metric to determine similarity between food points

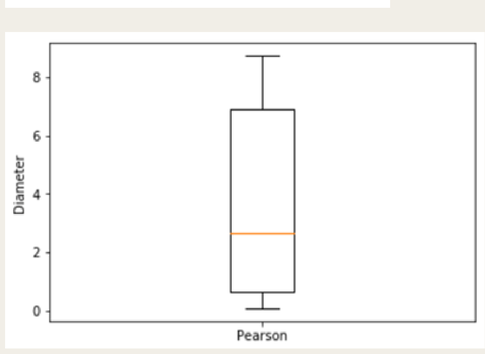

Figure 8 Box and whisker plot showing the diameters of the different clusters when using Pearson correlation. Lower is better

Eisen Cosine Similarity

- Same as Pearson Correlation but with the mean set to 0
- Approach Michael Wyatt used in his paper
- Clusters have smaller diameter

$$d_{eisen}(x,y) = 1 - \frac{\left|\sum_{i=1}^{n} x_i y_i\right|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

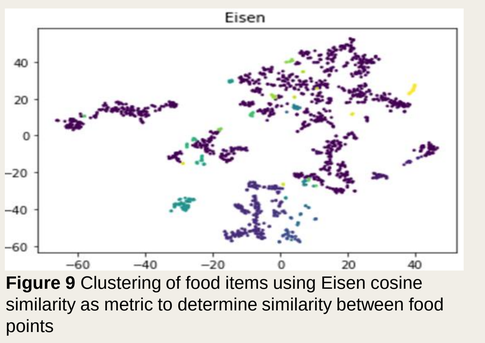

Figure 9 Clustering of food items using Eisen cosine similarity as metric to determine similarity between food points

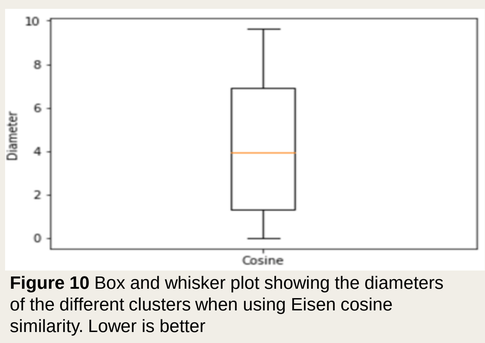

Figure 10 Box and whisker plot showing the diameters of the different clusters when using Eisen cosine similarity. Lower is better

Final Results

- USDA is clustered by how many digits at the beginning of the code match, not good for showing nutritional value
- Eisen and Pearson both perform very well when looking at clusters that are more compact and have farther separation when compared to USDA's clustering
- Manhattan performs poorly most likely due to the high dimensions of dataset

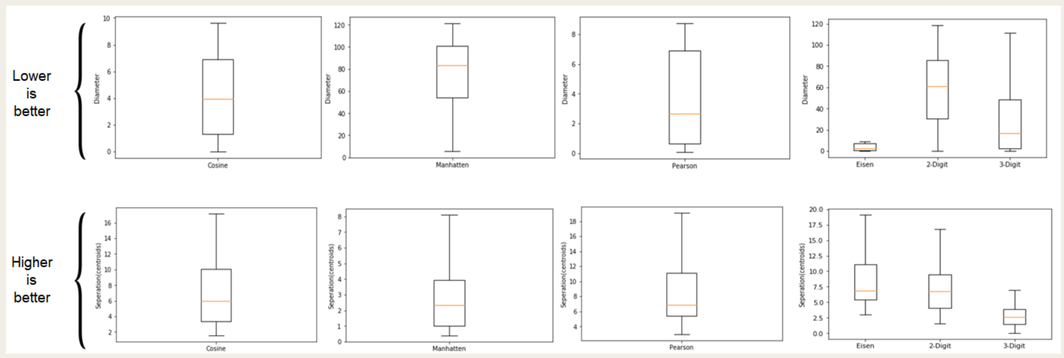

Figure 11 Showing off results from presented experiments and comparing results to USDA 2-digit and 3-digit clustering results

Conclusion

- High dimension correlation works better
- Infinitely many ways to define correlation
- Our clusters could allow people to make more informed choices when wanting to eat healthily
- USDA's groupings are misleading and are only convenient for categorizing foods

Method	# of Clusters	% of foods clustered	Avg. Diameter	Avg. Separation
Manhatten	47	90.33%	78.63	3.41
Eisen	34	69.54%	4.31	7.96
Pearson	23	73.93%	3.63	8.41
USDA 2-digit	49	100%	60.52	7.82
USDA 3-digit	203	100%	23.41	2.54

Figure 12 Table comparing results of each metric. Pearson performed the best

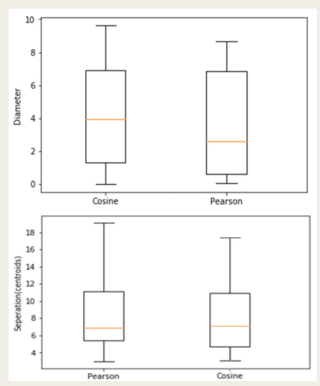

Figure 13 Comparing Diameter of clusters and the separation of clusters between Eisen cosine similarity and Pearson correlation. Higher separation is better, lower diameter is better

THE UNIVERSITY OF TENNESSEE KNOXVILLE

References:

M. Wyatt, T. Johnston, M. Papas, and M. Taufer. Development of a Scalable Method for Creating Food Groups Using the NHANES Dataset and MapReduce. In Proceedings of the ACM Bioinformatics and Computational Biology Conference (BCB), pp. 1 - 10. Seattle, WA, USA. October 2 - 4, 2016.