Food Group Classification Based On Micro- and Macro- Nutrient Content

|  |  |  |
| --- | --- | --- |
| Ahmad Hariri Computer and Information Sciences University of Delaware Newark, DE ahmadgjh@udel.edu | Ankush Saini Computer and Information Sciences University of Delaware Newark, DE anksaini@udel.edu | Michela Taufer Computer and Information Sciences University of Delaware Newark, DE taufer@udel.edu |

***Abstract***

***In this paper we reproduce and validate the results of Wyatt’s, Johnston, Papas, and Taufer 2016[1] work by classifying meaningful food groups in dietary datasets by National Health and Nutrition Examination Survey (NHANES) that are less subjective in nature by dening a new objective method of identifying food groups exclusively based on the food's micro- and macro-nutrient content. We first perform extensive preprocessing of the NHANES raw data to mitigate impacts of missing nutrient values, redundancies, and different food intake quantities and scales. We then utilize an unsupervised learning clustering algorithm to create food groups within the preprocessed NHANES data and identify food groups with similar nutrient content. Finally, we parallelize our method to benefit from the scalable MapReduce paradigm. Our results show that our method identifies food groups with a smaller diameter and larger cluster separation distances than the standard, expert-informed, method of grouping food items.***

***Keywords—Clustering methods, data processing, Apache Spark, dietary data, micro- and macro-nutrients***

# MOTIVATION

We observe that the diet statistics of an individual or population is huge and complex. The reason behind this is the availability of various foods, having different nutrient values and its processing after consumption. Modelling and analysis of such data can play an important role in medical interpretation by developing guidelines based on food consumption habits and its effect on diseases. On the other hand, design and development of such a robust model are equally difficult and complex.

This modelling is traditionally done through data-driven and criterion-driven approaches, which were computationally intense and subjective in nature. The first method involved how diet conforms to pre-established dietary guidelines. This method relied on the appropriateness of the score to portray the overall quality of an individual. The second method involved clustering, finding the correlation between different food items. Thus, new techniques were required to develop meaningful food group classifications.

We address this by defining a new method of identifying food groups based on the micro- and macro-nutrient values of the food items. We start with a well-known dietary dataset, the National Health and Nutrition Examination Survey (NHANES), process the raw data. Preprocessing includes the removal of duplicate and missing food items, thus introducing objectivity into the food grouping. This makes all the nutrient values to have equal weight. Thereafter, we use an unsupervised learning clustering algorithm in order to create food groups within the NHANES data. The algorithm is able to group the food items together based on nutrient content similarity. The ever-increasing amount of collected data must be processed in parallel in order to maintain the ability for time efficient analysis. Thus, the solution relies on the MapReduce paradigm for the parallel analysis of the food's nutrient values. The ultimate goal of our method is to identify high quality food groups which have a smaller diameter (more internally similar) and larger cluster separation (more externally different) than the traditional algorithms.

Our method is unique as it relies on all available nutritional information of food item only. We will be replicating the results of Wyatt’s et al. 2016 work and thus, try to validate their work.

# contribution

We build a system to compare the results between three different approaches namely 2-digit, 3-digit and DBSCAN. DBSCAN is selected because the algorithm must be efficient with large datasets and be insensitive to noisy data. We need an algorithm which is able to distinguish noise or outliers from actual clusters. And we do not want a clustering algorithm which preferentially creates globular clusters. Furthermore, two different metrics - Euclidean and Cosine similarity are used to compare the distances within each approach. Our Euclidean distance function can be defined as follows: (∑ni=1(xi−yi)2)1/2

Where x and y are two vectors

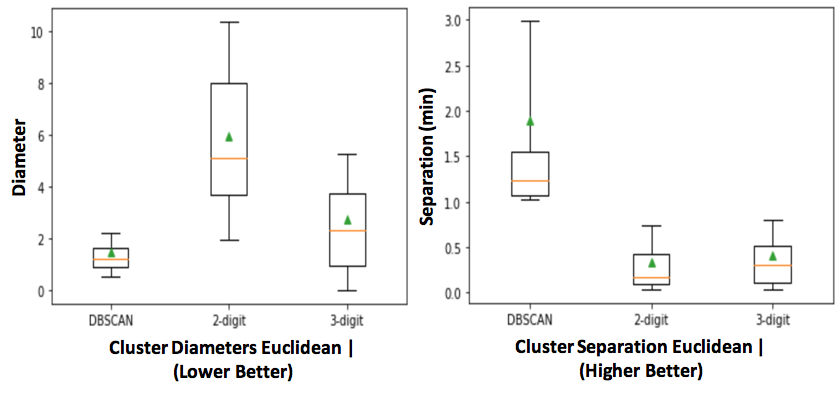
Whereas, our cosine similarity function can be defined as follows: x●y/(x●x)1/2\*(y●y)1/2

Where x and y are two vectors

The efficiency and accuracy of the system can be validated by measuring the inter- and intra-cluster distances. The greater the cluster separation implies well-separated clusters. On the other hand, the smaller the diameter, the more definitive clusters. Thus, these parameters play a crucial role in analysis and deduction of results.

# results

We use three different ways to classify food groups. First, we use DBSCAN to cluster the foods (based on nutrient values). Second, we cluster the foods into groups using first two digits of the USDA food code. Third, we cluster the foods into groups by the first three digits of the USDA food code. Using DBSCAN, 92 clusters are generated and cluster 60.05% of the unique foods into groups. Using the first 2 and 3 digits of the USDA food code and the same 60.05% of unique food items, 46 (2 digits) and 192 (3 digits) clusters are generated. In figure 2, the left Boxplot shows the diameters of the foods groups, the lower the diameter is better. Each box plot contains six different statistical information associated with the 92, 46, 192 groups generated by our method, 2-digit, 3-digit, respectively. The whiskers range from 10th (bottom) percentile to 90th (top) percentile. The top, bottom, line in the middle of the box are related to the 75th percentile, 25th percentile, 50th percentile, respectively. The average diameter is the green triangle. When using DBSCAN, we get average diameter 1.48 compared to 5.95 and 2.76 for the 2 digit and 3 digits, respectively. The 50th percentile is the median, we get 1.25 using DBSCAN compared to 5.14 and 2.35. In figure 2, the right box plot displays the separation of the foods groups. When using DBSCAN, we get average separation 1.89 compared to 0.33 and 0.41 for the 2 digit and 3 digits, respectively. The median for DBSCAN is 1.22 compared to 0.16 and 0.30 for the 2 digit and 3 digits, respectively. We got results similar to the work of Wyatt’s et al. 2016[1] results. Hence, the work of Wyatt’s et al. 2016[1] is reproducible.

Fig. 2. (*Box plots shows the clusters diameter(left) and separation(right))*

# conclusion

In this paper we provide a meaningful (objective) food groups classification based on nutritional data. We use parallel approach (using DBSCAN in a MapReduce framework) to cluster food groups based on the objective approach and compare it with USDA approach. We reproduce the work of Wyatt’s et al. 2016[1] to validate the results. We measure food groups quality based on the nutritional values similarity (diameter) of foods in the same a group and by the nutritional dissimilarity (separation) of foods from different groups. Our first test case (cluster similarity), we found an enhancement from a mean cluster diameter of 5.95 (2-digit USDA) or 2.76 (3-digit USDA) down to 1.48 units for (DBSCAN). Our second test case (cluster separation), we found an enhancement from an average of 0.33 (2-digit USDA) or 0.41 (3-digit USDA) up to 1.89 for (DBSCAN). Wyatt’s et al. 2016[1] approach classify food groups based on nutrient content (objective) which help in advising patients with dietary restrictions in which it will improve the health. Finally, our results match with Wyatt’s et al. 2016[1] results. Therefore, Wyatt’s et al. 2016 work is reproducible and trusted since we were able validate their work.

# **references**

1. 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.