A Cluster Based Assessment of Meal Similarity for Meal Recommendation

Pooja M. Desai, Tara V. Anand, Lena Mamykina

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

Department of Biomedical Informatics, Columbia University Medical Center

INTRODUCTION

- Diabetes is a chronic condition that requires rigorous nutrition and lifestyle management. Individuals often struggle to identify meals that keep their blood sugar in a healthy range and match their preferences.
- Patients seek nutrition guidance tailored to their food preferences. Specifically, they seek suggestions for meal modifications and an evaluation of how meals meet goals.
- We explore the use of free-text user-entered meal descriptions to assess similarity between meals and alignment between meals and selected nutrition goals. We use:
  1. Clustering approaches to identify groups of similar meals
  2. Binary Classification to characterize within-cluster meals as meeting or not meeting a specific nutrition goal

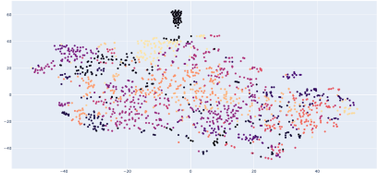
METHODS

- Data was collected from T2Coach, a mobile application for diabetes where individuals can select nutrition goals to work on and log free-text descriptions of meals.
- **Text Pre-Processing:** Steps included lowercase conversion, spelling correction (comparing against a dictionary of known food terms), and punctuation removal. Stop word removal included standard and measurement words (e.g. cup, tbsp).
- **Word Embeddings:** Food descriptions were encoded using multi-hot, TF-IDF, Average Word2Vec and Doc2Vec embeddings and evaluated.
- **Clustering**: Several methods were explored, specifically: K-Means, Hierarchical, and Mean Shift clustering. Clusters were evaluated through intrinsic and qualitative methods.
- **Binary Classification**: Several methods explored, including: Logistic Regression, Random Forest Classification, Gaussian Naive Bayes, and Gradient Boost.

FINDINGS

TASK 1: Clustering to Identify Similar Meals

- Word2Vec embeddings with K-means clustering performed best on internal metrics (silhouette coefficient = 0.19; Calinsky-Harabasz Index= 92.42) and in qualitative review.



Cluster 1 (rice & beans)
- yellow rice beans beef meatballs
- danish rice beans salad beef
- beans potato salad rice
- whites rice tofu eggplant string beans carrots

Cluster 2 (salads)
- chicken roast sweet potato salad
- ham swiss grapes potato salad
- salad sauteed shrimp
- panera caesar salad apple

Cluster 3 (chicken)
- chicken soups celery noodles
- pepperoni pizza chicken nuggets
- baked chicken vegetables
- chicken pasta alfredo
- garlic chicken spinach

TASK 2: Binary Classification of Meal Goal Alignment

- Trained classifiers for the two most common goals: "eat a variety of fruits and vegetables" and "choose lean proteins". Used .67/.33 test/train split and 5x cross-validation.

Goal	**"Eat a variety of fruits and vegetables"**				**"Choose lean proteins" goal**			
F1 scores	*Logistic Regression*	*Random Forest*	*Naive Bayes*	*Gradient Boosting*	*Logistic Regression*	*Random Forest*	*Naive Bayes*	*Gradient Boosting*
Multi-hot	**0.82**	0.72	0.55	**0.80**	**0.67**	0.62	0.49	0.63
TF	0.74	0.77	0.57	**0.81**	**0.66**	0.61	0.51	**0.69**
Word2Vec	0.57	0.63	0.59	0.66	0.57	0.56	0.58	0.60
Doc2Vec	0.58	0.71	0.70	0.72	0.55	0.58	0.55	0.60

CONCLUSIONS

- This work shows moderate success using meal descriptions to identify groups of similar meals and predict goal achievement.
- Future work can examine the incorporation of additional pre-processing, the inclusion of portion size information, leveraging properties of word embeddings and exploring addition of image data to improve clustering and classification.