Machine Learning to Investigate the Relationship Between Nutrition and Gut Biome

by Lauren "Zoe" Baker

The human gut microbiome is a population of microorganisms that inhabit the human gastrointestinal tract. The host bacteria that comprise the human gut microbiome are interconnected with nutrition and human health outcomes, such as obesity, type 2 diabetes, inflammatory bowel disease, and cardiovascular disease [1]. One of the challenges in investigating the nature of this relationship between the microbiome, human health, and nutrition is that there are hundreds of species of bacteria which reside in the gut microbiome. An additional challenge lies in the fact that this relationship may be largely undefined or high-dimensional – that is, consisting of several interrelated features. Without a clear hypothesis or direction, clinical studies would likely be infeasible; empirically isolating and evaluating the significant impact of thousands of nutrients on thousands of different species of bacteria would be incredibly time and resource intensive. To better uncover this hidden relationship, one might then turn to the use of classical statistical methods. However, these methods often lack in their ability to perform exploratory, unguided analysis on large datasets [2]. Thus, the scope of this problem suggests the use of machine learning.

Machine learning uses automated statistical models and algorithms to investigate patterns and form inferences from data. As a part of the Machine Learning, Informatics, and Data Science team at Mines (MInDS@Mines), one of our core focuses is the development and application of machine learning algorithms to human health conditions whose mechanisms are not well understood. Thus, our interest in the human microbiome is centrally motivated by the microbiome's connection to health conditions. However, nutrients and the bacteria which reside in the microbiome have a complicated,interconnected relationship. The nutrients we consume impact the bacteria that exist in our gut, while also independently impacting our health. Therefore, to truly understand the connection between the microbiome and our health, our team first set out to characterize the relationship between nutrients and bacteria. To accomplish this task, we examined a dataset of 2000 patients which included bacteria abundance levels found from their stool samples, and a dataset on that same group of 2000 patients with a background questionnaire, and nutrient information from a log of their food intake.

Machine learning algorithms build mathematical models from a set of data to perform tasks. Given a set of input data, these models produce outputs, like assigning samples from the input data into categories, or predicting the value of a certain variable. Machine learning models are built through either supervised or unsupervised learning. In supervised learning, the model is given data that includes both the inputs and the correct, or previously known, output for each input. The model is then trained to make predictions as close as possible to these correct outputs. We were not able to provide the "correct/ previously-known" relationships to our model. Instead, our model had to learn these relationships by itself. Thus, we looked at unsupervised learning. Specifically, we examined a type of unsupervised learning called clustering. Clustering is the process of grouping together objects into clusters in order to maximize how similar objects are to other objects within its cluster and maximize how different objects are to objects outside its cluster. Clustering is unsupervised

because instead of requiring outputs, it merely examines how inputs relate to each other.

Clustering is useful because understanding how data falls into groups can provide insight into possible patterns and structures in the data, an insight that we could then apply to further research questions. For instance, say we cluster bacteria into several groups. By examining patients' levels of bacterial abundance in relation to a certain group of bacteria species, we might learn that having a higher abundance of bacteria within this group is linked to a certain health condition. To perform this clustering, we first must note that our datasets have two key dimensions. One dimension, the feature dimension, are the attributes of patients such as their bmi, levels of a certain nutrient, or their stool abundance of a certain bacteria. The other dimension is the patients themselves. Traditional clustering methods examine how to group together data points from a single dataset along either the feature dimension, grouping attributes of the data, or the subject dimension, grouping individual patients. We wanted to examine if we could improve the clustering results through two methods: (1) co-clustering, which is simultaneously clustering along both the feature dimension and subject dimension; and (2) transfer learning, which is incorporating information learned from clustering one dataset into clustering another dataset.

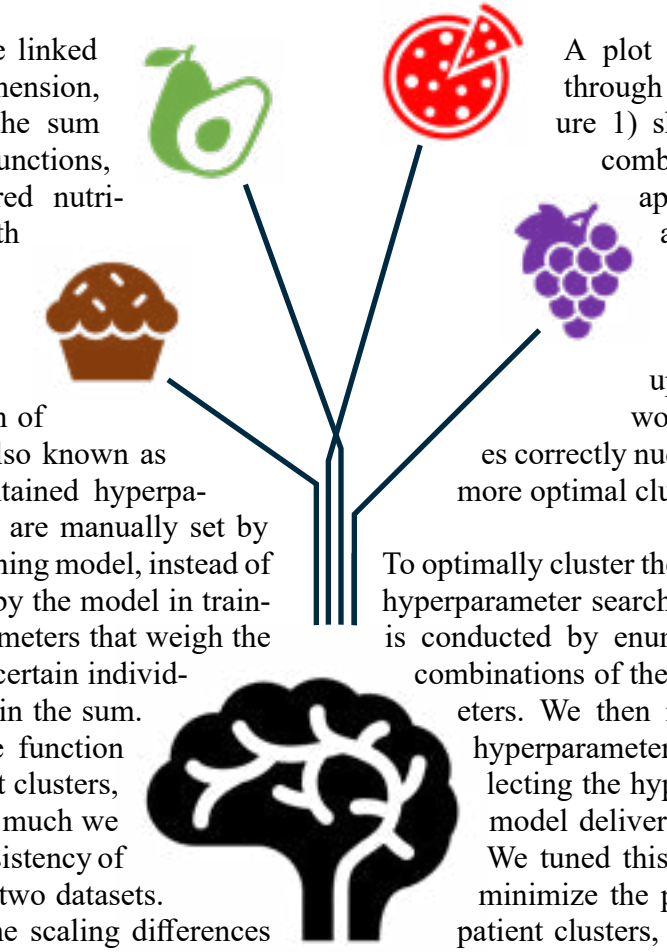
We first investigated co-clustering because feature and subject dimensions rarely have a strictly one-way relationship – that is, one dimension's similarity is rarely only unidirectionally dependent on the other dimension [3]. There are a multitude of algorithms that can be used to cluster data. For this project, we performed a Nonnegative Matrix Tri-Factorization (NMTF) on the dataset. Factorization is the process of writing a matrix of data as the product of several matrix factors. Intuitively, it makes sense that factorization is related to clustering: The goal of clustering is to split up data into several smaller components, and factorization splits up a data matrix into several matrices. NMTF, in particular, finds factors which are inherently clusters of the columns of the input matrix of data. Note that this factorization method finds approximate matrix factors. In other words, when these matrix factors are multiplied back together, they closely resemble the original matrix but do not perfectly equal the original matrix. An optimal NMTF minimizes the error in this approximation, which is equivalent to minimizing a distance function between the product of those factors and the original input matrix. An iterative algorithm is used to discover this optimal factorization. After finding the optimal factorization, we can derive from those factors the locations of the centers of the clusters, and which clusters each data point belongs to along both the feature and subject dimensions.

We also considered transfer learning, wherein if two separate machine learning problems share a relationship, solving one problem could provide additional information and data to aid in solving the related problem. For example, information gained by a machine learning model that categorizes cats could transfer that information to a machine learning model that categorizes dogs. Transfer learning for a clustering task would mean using the clustering structure and results of one dataset to aid in clustering the other. In the context of this project, if bacteria species and nutrients in the gut were related, clustering the bacteria genera dataset might aid in the clustering of the nutrient dataset. Given that both co-clustering approaches share information about patients, transferring this knowledge should be a natural fit. Transfer learning could yield a clustering analysis that better accounts for the underlying relationships between nutrients and bacteria.

To transfer knowledge from the co-clustering of separate datasets, H. Wang et. al. demonstrated that we could form an objective by summing the subproblems of co-clustering each dataset so that we can simultaneously solve both subproblems [4]. In other words, we could use NMTF to find co-clusters for the datasets at the same time. To do this, we combined the minimization of the distance function which leads to the best approximate factors for the bacteria, and the minimization of the distance function which leads to the best approximate factors for the nutrients, into a single sum to be minimized. In addition, we added a new distance function between the patient clusters of the nutrients and the patient clusters of the bacteria to our objective, to also be minimized to encourage the model to make clusters which are consistent across both datasets since they are on the same set of patients. Effectively,

nutrients and bacteria were linked together by their shared dimension, patients. By minimizing the sum of these three distance functions, we concurrently coclustered nutrients and bacteria along both dimensions while ensuring consistent patientwise clustering.

The formulation of the sum of these distance functions, also known as an objective function, contained hyperparameters. Hyperparameters are manually set by the user of the machine learning model, instead of being automatically tuned by the model in training. We had two hyperparameters that weigh the precedence of minimizing certain individual distance functions within the sum. One weighted the distance function between both sets of patient clusters, allowing us to control how much we wished to prioritize the consistency of the clustering between the two datasets. The other controlled for the scaling differences between the two datasets, as simply having data at a larger scale should not make it any more or less important in the training of a machine learning model. Using standard constrained optimization with Lagrangian multipliers, we derived a series of update processes for each variable to minimize the combined sum of distance functions to create clusters in the data. These update processes start each variable at a random point, then nudge that variable in a direction that minimizes the combined sum. With each iteration, we descend upon a more optimal solution. We use this iterative method since there is no closed-form solution to minimizing the sum.



A plot of the objective function through the training process (Figure 1) shows that the sum of the combined distance functions approaches zero through iterations of the derived updating method, supporting the hypothesis that the update process converges upon a minimum. In other words, our updating processes correctly nudge our variables towards a more optimal clustering.

To optimally cluster the data, we then performed a hyperparameter search. A hyperparameter search is conducted by enumerating over all possible combinations of the manually set hyperparameters. We then run the model with those hyperparameters and compare results, selecting the hyperparameters in which the model delivers the most optimal results. We tuned this hyperparameter search to minimize the percent difference between patient clusters, which aligns with our primary goal of discovering potential helpful relationships between the two datasets.

After finding the clusters using our co-clustering with transfer learning algorithm, we evaluated the quality of our clustering using a silhouette score. A silhouette score is an aggregate measure of how similar each object in a cluster is to other objects in its cluster, and how different each object is from other objects outside of its cluster. However, as we optimized for similarity between our patient clusters, the silhouette score consistently became worse. Prioritizing the relationship between the two separate datasets seemed to decrease the quality of

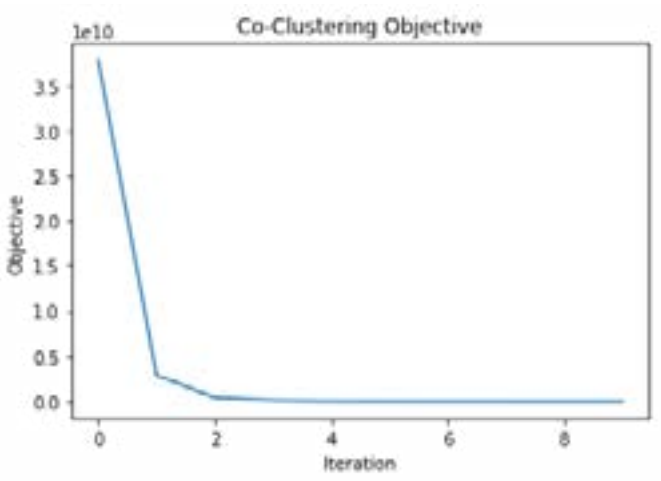


Figure 1: Plot of Objective Function

our clusters. This result indicates that sets of patients might be better clustered by nutrients or bacteria individually, and not together. These results may also simply indicate that the machine learning model we developed is ill-suited to the problem.

However, while the machine learning model derived in this paper may not have fit this particular dataset, and did not provide the insight we were looking for, this research has led to the creation of a co-clustering with transfer learning algorithm. This algorithm can simultaneously cluster multiple related datasets while ensuring that these datasets have consistent clusters along a shared dimension. In addition, this algorithm has theoretic guarantees. In other words, we have an update process that can be proven with each iteration to decrease the objective function of the algorithm and increase the quality of our solution. Thus, this algorithm can be applied to other datasets. It is also important to note that while this algorithm did not find a relationship between nutrients and bacteria, it does not mean that a relationship does not exist.

To continue our research and search for more definitive results, the team and I are investigating further modifications to our experiment along with other machine learning methods. One of the promising modifications is the incorporation of non-linearity into our relationship model through the use of the "kernel" trick. This method maps data into a higher dimension. Data that is inseparable with linear methods in a lower dimension could be separable in a higher dimension. In our case, perhaps our data cannot be usefully clustered in the dimension that it is in right now, but might exhibit more definitive patterns and clusters in a higher dimension. In addition, we are investigating an Active-Learning algorithm which uses sparse matrices to select out the most important or representative samples in a dataset. Perhaps clustering, or performing other machine learning algorithms, on these more representative samples will provide more insightful and meaningful results.

References

[1] R. Singh, H.-W. Chang, D. Yan, K. Lee, D. Ucmak, K. Wong, M. Abrouk, B. Farahnik, M. Nakamura,T. Zhu,T. Bhutani, and W. Liao, "Influence of diet on the gut microbiome and implications for human health," Journal of Translational Medicine, vol. 15, p. 73, Mar. 2017. doı: 10.1186/s12967-017-1175-y.

[2] R. Iniesta, D. Stahl, and P. McGuffin, "Machine learning, statistical learning and the future of biological research in psychiatry," Psychological Medicine, vol. 46, no. 12, pp. 2455–2465, 2016. doı: 10.1017/S0033291716001367.

[3] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. Mooney, "Model-based overlapping clustering," Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jan. 2005, pp. 532–537. doı:10.1145/1081870.1081932.

[4] H. Wang, F. Nie, and H. Huang, "Large-scale cross-language web page classification via dual knowledge transfer using fast nonnegative matrix trifactorization," ACM Trans. Knowl. Discov. Data, vol. 10, no. 1, Jul. 2015, ıssn: 1556-4681. doı: 10.1145/ 2710021.

Lauren "Zoe" Baker is a sophomore at Colorado School of Mines pursuing a B.S. in Computer Science and a B.S. in Computational and Applied Mathematics. Zoe currently conducts research with the Machine Learning, Informatics, and Data Science (MInDS @ Mines) laboratory under the guidance of mentors Dr. Hua Wang, Saad Elbeleidy, and Lodewijk Brand. Her research is centered around the development and application of machine learning algorithms to biological data.