

Unsupervised Learning for Hydrogen Breath Tests

Michael NETZER^{a,1}, Friedrich HANSER^a, Maximilian LEDOCHOWSKI^b and Daniel BAUMGARTEN^a

^a*Institute of Electrical and Biomedical Engineering, UMIT – Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria*

^b*Akademie für Ernährungsmedizin, Innsbruck, Austria*

Abstract. Hydrogen breath tests are a well-established method to help diagnose functional intestinal disorders such as carbohydrate malabsorption or small intestinal bacterial overgrowth. In this work we apply unsupervised machine learning techniques to analyze hydrogen breath test datasets. We propose a method that uses 26 internal cluster validation measures to determine a suitable number of clusters. In an induced external validation step we use a predefined categorization proposed by a medical expert. The results indicate that the majority of the considered internal validation indexes was not able to produce a reasonable clustering. Considering a predefined categorization performed by a medical expert, a novel shape-based method obtained the highest external validation measure in terms of adjusted rand index. The predefined clusterings constitute the basis of a supervised machine learning step that is part of our ongoing research.

Keywords. Unsupervised Learning, Clustering, Time Series, Hydrogen Breath Tests, Carbohydrate Malabsorption

1. Introduction

The disaccharide lactose and the monosaccharide fructose play an important role in our diet. Lactose is an integral part of most dairy products whereas fructose can be directly linked to highly processed foods containing sugars in large quantities. Physiologically, lactose is broken down by the enzyme lactase into its monosaccharide components glucose and galactose which are then absorbed by the brush border of the epithelial cells lining the small intestine. About 65% of humans, however, downregulate the production of lactase after weaning resulting in lactose malabsorption [1]. In contrast, fructose is directly transported across intestinal epithelial cells by the Glucose Transporter 5 (GLUT5). This transport process is mediated via facilitated diffusion and its transport capacity is limited. On entering intestinal epithelial cells, fructose is converted to glucose to a high degree [2]. Excess fructose enters the liver via the portal vein system, possibly inducing nonalcoholic fatty liver disease. Not surprisingly, 34% of humans develop fructose malabsorption but the underlying detailed mechanisms are still elusive [3]. In general, carbohydrate malabsorptions lead to intestinal fluid retention, causing diarrhea

¹ Corresponding Author: Michael Netzer, Institute of Electrical and Biomedical Engineering, UMIT – Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria, E-Mail: michael.netzer@umit.at

and abdominal distention [4]. Carbohydrate malabsorption (CM) is often accompanied by Small Intestinal Bacterial Overgrowth (SIBO), which is defined as an increase in the number and/or alteration in the type of bacteria in the small intestine [5]. Hydrogen breath tests are a well-established diagnostic tool to identify pathophysiology of functional gastrointestinal disorders such as CM and SIBO [6].

In this work, we analyze hydrogen breath tests resulting from lactose and fructose tolerance tests using unsupervised machine learning approaches with the aim of identifying new characteristic patterns in H₂ time series data. In general, unsupervised methods search patterns without requiring a label. Clustering methods are one popular representative of this category. Objects within one cluster have, in general, a higher similarity compared to objects of another cluster. The number of clusters can be either specified by the scientist or determined by evaluating different cluster metrics. Recently, Rubio-Escudero et al. [7] applied data mining techniques to lactose-based hydrogen breath test data including clustering based on *k*-means using Euclidean distance. The authors used three cluster indexes (i.e., Silhoutte, Davis-Bouldin, and Dunn Index) to obtain a suitable number of clusters. In our work, we compare different clustering method using a total of 26 clustering indexes for internal validation. In a next step, we perform external validation using a predefined categorization scheme. To our best knowledge, the combined application of unsupervised learning methods using internal and external validation measures to results of lactose and fructose hydrogen breath tests has not been described previously.

2. Methods

2.1. Dataset

We included 4302 breath gas samples in our study, where 1879 samples result from a lactose tolerance test (LTT) and 2423 samples from a fructose tolerance test (FTT). Each trajectory (i.e., time series) includes five H₂ measurements taken in intervals of 30 minutes. Tab. 1 summarizes the experimental details. For a detailed description of the implementation of the hydrogen breath tests the authors refer to [8]. The persons included in this study gave written informed consent. The consecutive statistical analysis was performed using anonymized data.

Table 1. Description of the considered datasets including test substance, dosage, readings and class distribution of the subgroups.

Dataset	Test substance	Dosage in 250 ml water	Readings (min)
LTT	Lactose	50 g	0-30-60-90-120
FTT	Fructose	25 g	0-30-60-90-120

2.2. Unsupervised Learning Approach

We use partitional, hierarchical and a relatively new category of shape-based clustering methods to identify characteristic patterns in the H₂ time series data resulting from LTT and FTT.

2.2.1. Partitional Methods

Partitional clustering methods separate objects into a set of disjoint clusters [11]. The most popular method of this category is the k -means algorithm, which is employed in this work. Each object is assigned to its nearest cluster center according to the Euclidean distance between the two [12]. The parameter k represents the number of clusters and is a user defined parameter. A more robust version uses centric objects in the cluster as representatives.

2.2.2. Hierarchical Methods

Hierarchical methods identify clusters by merging (agglomerative) or splitting (divisive) objects into clusters [13]. Finally, a dendrogram indicates the resulting cluster hierarchies. The similarity between objects is quantified by using distance measures such as Euclidean distance. Methods to measure the distance between clusters (each consisting of multiple objects) include single link (minimum distance) or complete link (maximum distance between objects of each cluster). We use the more advanced Wards method that is based on a classical sum-of-squares criterion [14].

2.2.3. Clustering According to Shape

Transferring classical clustering approaches to longitudinal data, the majority of methods consider similarities at local time points. In contrast, Genolini et al. propose a method to group trajectories based on similarities in shapes [10]. In particular, the authors use a generalized distance of Fréchet and curve alignment to quantify similarity of trajectories. In our work we use the R implementation of this method (package *kmlShape*).

2.3. Cluster Evaluation Approach

2.3.1. Internal Validation Measures

Internal validation measures evaluate the quality of clusters based on measures such intracluster compactness and isolation, geometric or statistical properties, number of data objects and dissimilarity or similarity measurements [15]. Thus, these measures do not need external information. Popular representatives include the Silhouette [16] index that uses comparison of cluster tightness and separation. Davies-Bouldin [17] and Dunn index [18] use intercluster (i.e., distances between clusters) and intracluster (i.e., distances within a cluster) metrics to evaluate a clustering. Internal validation measures can be used to determine a suitable number of clusters. Charrad et al. propose a method that uses different indexes and determines the final number of clusters by majority voting [15]. In particular, the method includes the following steps:

1. Calculate $k \in K$ clusters where $k = 2, 3, \dots, k_{max}$ using a particular clustering method $m \in M$
2. Calculate $v \in V$ internal validation measures for each k
3. Generate a recommendation of relevant number of clusters $r \in R$ for each $v \in V$.
Depending on the methodology, select r with the smallest or highest index value.
4. Summarize all relevant number of clusters by using majority voting

As an example, let the relevant number of clusters suggested by Silhouette and Dunn index (both use a number of clusters where index reaches a maximum) be 3 and the relevant number of clusters suggested by Davis-Bouldin (relevant number of clusters is where index reaches a minimum) be 4. In this case, the final number of clusters obtained by majority voting would be 3. In our analysis, we use k -means and hierarchical clustering as clustering methods (M). The maximum number of clusters (k_{max}) was set to 8. A total of 26 clustering indexes were used (see [15] for further information).

2.3.2. External Validation Measures

External validation measures require external information. The Rand index (RI) and the adjusted Rand index (ARI) are two common measures to quantify how close clusters are to predefined reference partitions [18]. Our reference partition was defined by a medical expert. Formally, for a 2×2 contingency table (see Table 2), the adjusted rand index (ARI) is computed by Equation 1 [19].

Table 2. A 2×2 contingency table comparing partitions U and V.

Partition U	V	
	Pair in same group	Pair in different group
Pair in same group	a	b
Pair in different groups	c	d

$$ARI = \frac{\binom{n}{2(a+d)} - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{22} - [(a+b)(a+c) + (c+d)(b+d)]} \quad (1)$$

3. Results

3.1. Internal Cluster Validation

Fig. 1 depicts the frequencies among all 26 considered indices using lactose and fructose tolerance test results and using k -means and hierarchical as clustering methods. A number of 2 relevant clusters was proposed using the lactose dataset and the k -means clustering method (a majority of 7 indexes suggested this value). Using hierarchical clustering, the relevant number of clusters was 2 and three (each suggested by 7 indexes). The number of relevant clusters in the fructose dataset was 2, as well, suggested by a majority of 10 indexes using k -means and 8 using hierarchical clustering. Fig. 2 visualizes the corresponding mean trajectories for the datasets LTT and FTT. We observed no substantial difference between k -means and hierarchical clustering setting the number of clusters $k = 2$. The mean trajectory of the first LTT cluster describes a flat pattern increasing at 120 minutes. The mean trajectory of the second cluster shows a strong H_2 increase with a maximum at 120 minutes. The mean trajectory of the first FTT cluster shows a small increase with a maximum ppm value at 90 min. The mean trajectory of the second cluster shows again a strong increase with a maximum value at 90 minutes.

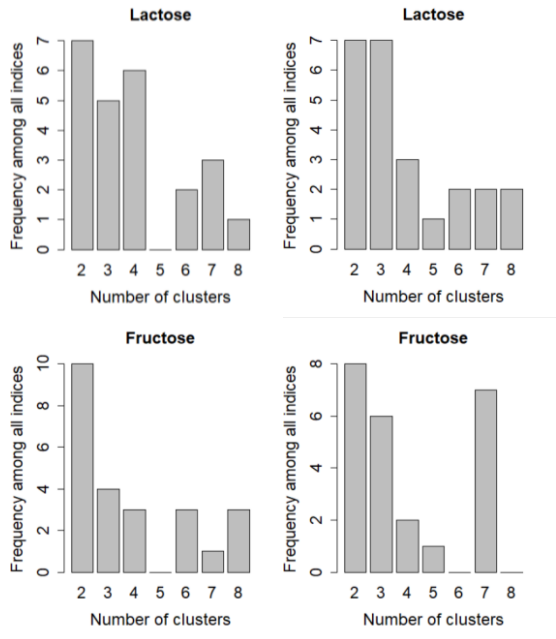


Figure 1. Frequencies among all indices using lactose tolerance test (top row) and fructose tolerance test (bottom row) using k -means (left column) and hierarchical (right column) as clustering methods.

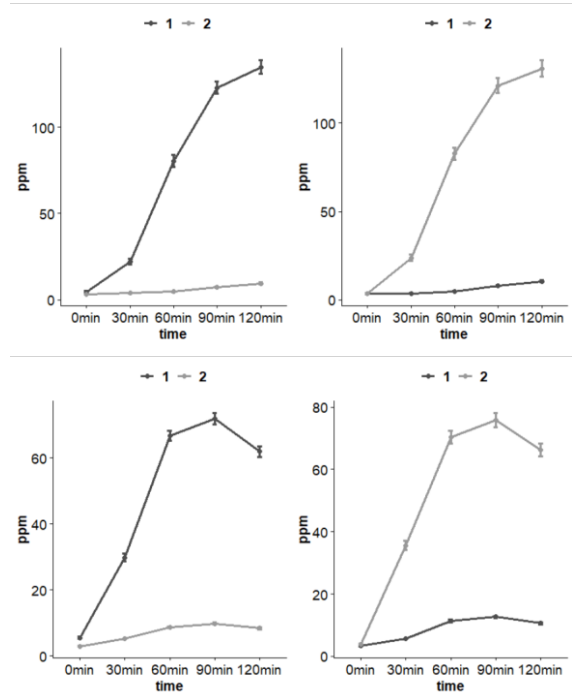


Figure 2. Mean trajectories for the resulting suggested number of clusters ($n=2$) using lactose tolerance test (top row) and fructose tolerance test (bottom row) using k -means (left column) and hierarchical (right column) as clustering methods.

Table 3. Proportions of groups NEG, POS and SIBOS for the two considered datasets (LTT, FTT).

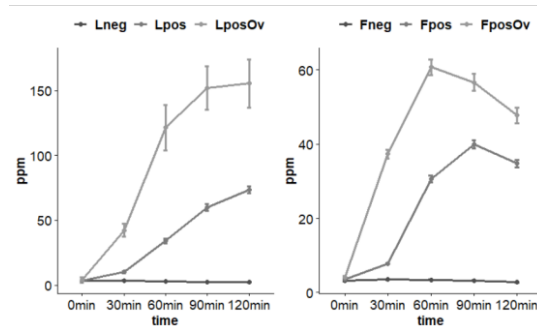
Dataset	NEG	POS	SIBOS
	Subgroup ratios (%)		
TT	75	24	1
FTT	59	27	14

3.2. External Cluster Validation

In contrast to obtaining the number of clusters by using internal validation measures, we used a categorization scheme proposed by a medical expert. The number of manually defined clusters was set to $k = 3$ representing the following groups

1. Negative lactate or fructose test result (LNEG, FNEG)
2. Positive lactate or fructose test result (LPOS, FPOS)
3. Small intestinal bacterial overgrowth syndrome (SIBOS)

The proportion of the groups as categorized by the medical expert is given in Tab. 3. The corresponding mean trajectories for dataset LTT and FTT are depicted in Fig. 3.

**Figure 3.** Mean trajectories for the manually determined number of clusters ($n=3$) using lactose tolerance test (left) and fructose tolerance test (right).

The resulting ARI values using k -means, hierarchical clustering and clustering according to shape are depicted in Tab. 4. The highest ARI values were obtained using clustering according to shape, which was developed to efficiently cluster longitudinal data.

Table 4. ARI values for the considered datasets using k -means, hierarchical clustering and clustering according to shape.

Dataset	ARI		
	k-means	Hierarchical	Shape based
LTT	0.54	0.17	0.68
FTT	0.53	0.21	0.67

4. Discussion

In this work we analyzed H_2 trajectories of breath gas samples resulting from lactose and fructose tolerance tests. In contrast to [7], we analyzed clusters in trajectories of lactose and fructose hydrogen breath tests using internal and external validation measures. First, we searched for number of clusters using 26 internal validation indexes. Based on the majority voting as proposed by [15], we determined a total of 2 dominant clusters. However, biologically we would expect three clusters in each dataset (LTT and FTT) representing groups NEG, POS and SIBOS. The SIBOS group was comparable small (see Tab. 4) which might explain that this group is not identified as a separate cluster by most indexes. However, the Ball index suggested a number of three clusters for all experiments (see Tab. 5). The Ball index [21] uses measurements based on sums of squares as measure of dispersion [22]. After setting the number of clusters to $k = 3$ we externally evaluated our clustering methods using ARI. In addition to k-means and hierarchical clustering, we applied a new method using shape information [10]. The shape-based method reached the highest similarity between the reference partition (ARI of 0.68 for LTT and 0.67 for FTT, respectively).

We conclude that only a small amount of internal validation indexes such as Ball index were able to obtain a reasonable number of clusters in our data. It is further important to externally validate clusters by using medical expert knowledge. Clustering methods considering shape information turned out to be the best method for clustering longitudinal data resulting from lactose and fructose tolerance tests. In our ongoing work, we will apply supervised machine learning techniques to automatically predict the result of the tolerance test (i.e., positive, negative or SIBOS).

Table 5. Listing of all clustering indexes that suggested a total of three clusters for LTT and FTT dataset and the considered clustering methods k -means and hierarchical clustering.

Dataset	Clustering method	Indexes suggesting $k=3$
LTT	k -Means	TrCovW, Duda, PseudoT2, Ratkowsky, Ball
	Hierarchical	Hartigan, Scott, TrCovW, TraceW, Ratkowsky, Ball, Frey
FTT	k -Means	Scott, TraceW, Friedman, Ball
	Hierarchical	Scott, TrCovW, TraceW, Ratkowsky, Ball, Frey

References

[1] Y. Itan, B.L. Jones, C.J. Ingram, D.M. Swallow and M.G. Thomas, A worldwide correlation of lactase persistence phenotype and genotypes, *BMC evolutionary biology* **10**(1) (2010), 1–11.

[2] C. Jang, S. Hui, W. Lu, A.J. Cowan, R.J. Morscher, G. Lee, W. Liu, G. Tesz, M. Birnbaum and J.D. Rabinowitz, The Small Intestine Converts Dietary Fructose into Glucose and Organic Acids, *Cell Metabolism* **27**(2) (2018), 351–361.

[3] R.P. Ferraris, J.-y. Choe and C.R. Patel, Intestinal absorption of fructose, *Annual review of nutrition* **38** (2018), 41–67.

[4] A. Amann, G. Poupart, S. Telser, M. Ledochowski, A. Schmid and S. Mechtcheriakov, Applications of breath gas analysis in medicine, *International Journal of Mass Spectrometry* **239**(2–3) (2004), 227–233.

[5] J. Bures, J. Cyrany, D. Kohoutova, M. Förstl, S. Rejchrt, J. Kvetina, V. Vorisek and M. Kopacova, Small intestinal bacterial overgrowth syndrome, *World journal of gastroenterology: WJG* **16**(24) (2010), 2978.

[6] S.V. Rana and A. Malik, Hydrogen breath tests in gastrointestinal diseases, *Indian Journal of Clinical Biochemistry* **29**(4) (2014), 398–405.

- [7] C. Rubio-Escudero, J. Valverde-Fernández, I. Nepomuceno-Chamorro, B. Pontes-Balanza, Y. Hernández-Mendoza and A. Rodríguez-Herrera, Data Mining Techniques Applied to Hydrogen Lactose Breath Test., *PloS one* **12** (2017).
- [8] A. Eisenmann, A. Amann, M. Said, B. Datta and M. Ledochowski, Implementation and interpretation of hydrogen breath tests, *Journal of breath research* **2** (2008).
- [9] M.G. Omran, A.P. Engelbrecht and A. Salman, An overview of clustering methods, *Intelligent Data Analysis* **11**(6) (2007), 583–605.
- [10] C. Genolini, R. Ecohard, M. Benghezal, T. Driss, S. Andrieu and F. Subtil, kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes, *Plos One* **11**(6) (2016), 0150738.
- [11] X. Jin and J. Han, *Partitional Clustering*, in: Encyclopedia of Machine Learning, C. Sammut and G.I. Webb, eds, Springer US, Boston, MA, 2010, pp. 766–766.
- [12] L. Rokach and O. Maimon, *Clustering methods*, in: Data mining and knowledge discovery handbook, Springer, 2005, pp. 321–352.
- [13] B. Clarke, E. Fokue, and H.H. Zhang, *Principles and theory for data mining and machine learning*, Springer Science & Business Media, 2009.
- [14] F. Murtagh and P. Legendre, Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion?, *Journal of classification* **31**(3) (2014), 274–295.
- [15] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, *J Stat Softw* **61** (2014).
- [16] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* **20** (1987), 53–65.
- [17] D.L. Davies and D.W. Bouldin, A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence* (1979), 224–227.
- [18] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, *Journal of cybernetics* **4**(1) (1974), 95–104.
- [19] A. D’Ambrosio, S. Amodio, C. Iorio, G. Pandolfo and R. Siciliano, Adjusted Concordance Index: an Extension of the Adjusted Rand Index to Fuzzy Partitions, *Journal of Classification* (2020), 1–17.
- [20] J.M. Santos and M. Embrechts, *On the use of the adjusted rand index as a metric for evaluating supervised classification*, in: International conference on artificial neural networks, Springer, 2009, pp. 175–184.
- [21] G. Ball and D. Hall, *A novel method of data analysis and pattern classification*, SRI International: Menlo Park, CA, USA (1965).
- [22] F.J. Abarca-Alvarez, F.S. Campos-Sánchez and R. Mora-Esteban, Survey Assessment for Decision Support Using Self-Organizing Maps Profile Characterization with an Odds and Cluster Heat Map: Application to Children’s Perception of Urban School Environments, *Entropy* **21**(9) (2019), 916.