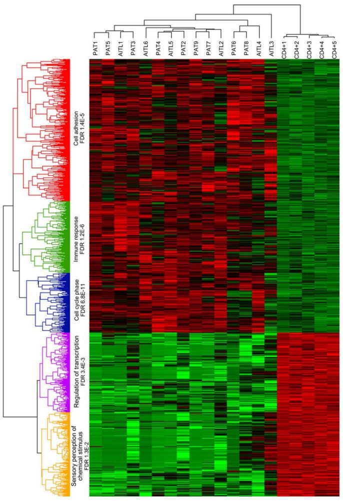
Clustering mixed data types with focus on a genetics context

by Paul Arndt

Clustering is an unsupervised machine learning approach to find patterns in large amounts of data. To do that, most clustering algorithms assume numerical data. In practice, though, most datasets contain a mixture of numerical, categorical, ordinal and/or binary data.
The basis for all clustering algorithms is the distance or similarity measure that determines how clusters are formed. Mostly these similarities are thought of as a n x n matrix, where the elements of the matrix represent the closeness of the observations to each other.



Heat Maps like this are often used to visualize clustering results.

The biggest problem for handling mixed data is to find a good similarity measure, meaning a function for comparing variables of different data types to each other.
There are basically two approaches for this. First, all categorical values could be converted into numerical or binary ones. An example for this is the Gower similarity, which compares the variables using a set of rules depending on the datatype of the variables. Second one could cluster the different data types separately and then merge the resulting clusters.

Another approach that does not compare variables but transforms entire tables is a modified multiple factor analysis. It uses Principal Component Analysis for numerical data and Multiple Correspondence Analysis for categorical data to create normalized distance tables that can then be analysed using normal clustering algorithms with euclidean distance or another Principal Component Analysis.

In the context of gene analysis, most clustering algorithms have been used on quantitative data like the output from the cuffdiff tool. My goal is to incorporate qualitative data in the form of variants into the clustering process in order to gain more meaningful results. In the future, even more diverse data could be added as input to further strengthen the results.

In our hands-on session, we will briefly show you what information you can gain from the variants today by using freely available tools and invite you to discuss with us the ethical consequences of having more and more precision in predicting future illnesses, as well as who should have access to this information.