Context-aware clustering of temporal event sequences

INSIGNEO
Institute for *in silico* Medicine

Suraj Tailor, Jessica G Magallanes Castaneda

Background

The project aim work was to seek a solution to clustering data consisting of mixed data types. This includes categorical, binary, numeric and sequence data. It was required the sequence distance metric that is described in the paper *Context Aware Trace Clustering* [1] was implemented in the given solution. This referenced paper's method describes a context aware distance metric that modifies the Levenshtein distance metric to make substitutions and deletions according to probabilities calculated from all of the sequence data. This mixed data clustering solution is then to be applied to patient data which not only includes their personal attribute information, but sequence information which describes all steps and duration of a patient's treatment.

Figure 1 shows an example of clustering applied to numeric data as a basic example of clustering.

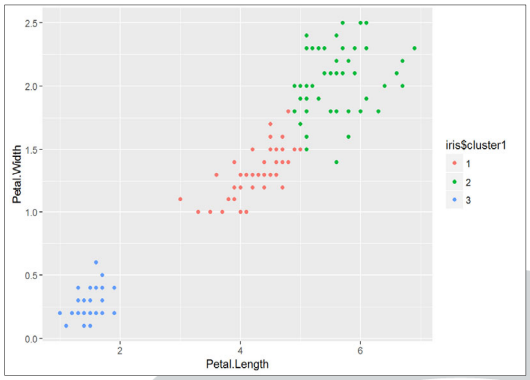

Figure 1: K-means clustering performed on two variables, petal length and petal width, of the popular Iris dataset. The three clusters closely relate to the original species labelling of the data.

Methods

The chosen method was an adaptation of the Gower's distance metric [2] which incorporates the context aware distance metric. This metric is then applied to individual patient data which gives a distance matrix for how dissimilar patients are to one another. Hierarchical clustering is then applied to the distance matrix to find all the combination of clusters for the data. Finally, a silhouette plot is used to obtain the optimum number of clusters for the data.

Results

The following was achieved during the project.

- A literature review was conducted on techniques in mixed data clustering.
- A formula to incorporate a sequence distance metric into the Gower's distance was devised:

$$Distance(i,j) = \frac{\sum_k^p \delta_{ijk} d_{ijk} w_k \; + \; \delta_{ij(p+1)} w_{p+1}}{\sum_k^{p+1} \delta_{ijk} w_k}$$

- An initial implementation of the adapted Gower's distance metric with the context aware distance metric was coded in R.
- Initial clustering experiments with the adapted Gower's distance metric were conducted, which has helped redefine the requirements for a clustering algorithm that takes attributes at unique and individual sequence level.

Conclusions

It is difficult to draw any conclusions on the effectiveness of the adapted Gower's distance metric in finding clusters until further tests are carried out using data where there is good understanding of the domain. This would be a data set where there is an idea of how clusters may be likely form. Once this is achieved, further work will be to implement an optimisation algorithm for the weightings within the Gower's distance equation

References

[1] J. C. B. R.P. and W. Aalst, "Context aware trace clustering: Towards improving processmining results," Apr. 2009.doi:10.1137/1.9781611972795.35.

[2] J. C. Gower, "A general coefficient of similarity and some of its properties,"Biometrics,vol. 27, no. 4, p. 857, 1971.doi:10.2307/2528823.

The Insigneo Institute for *in silico* Medicine is a collaborative initiative between the University of Sheffield and Sheffield Teaching Hospitals NHS Foundation Trust.

The University Of Sheffield.

Sheffield Teaching Hospitals NHS
NHS Foundation Trust