Patterns in the mind – Part 1

Introduction

Recently we have added k-means clustering to p:IGI+, to supplement the data exploration capabilities of the software. But what actually is clustering, how do we know when to use clustering and what techniques can we use to identify how many clusters we have?

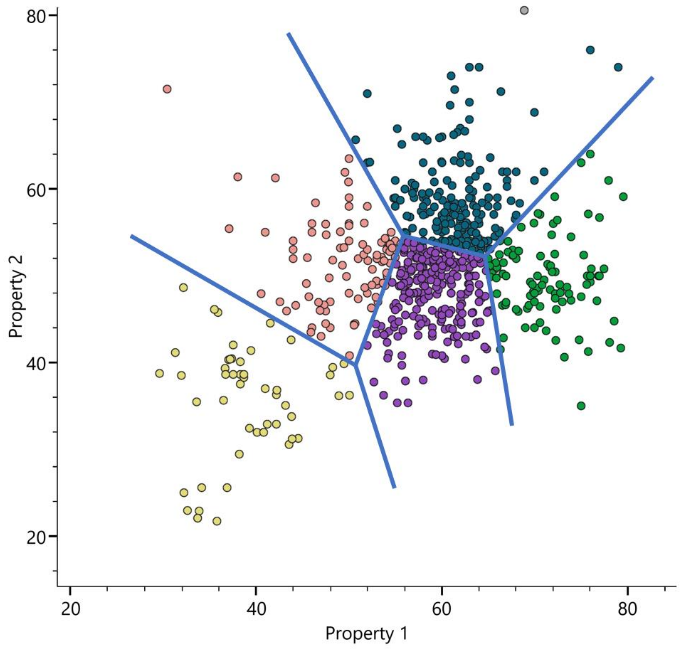
In this note, part 1 of a two part series, we consider two clustering methods, and ask the question when might clustering be an appropriate tool to use in data interpretation, using real and synthetic examples.

What is clustering?

Put simply, clustering is the act of finding similarly behaving groups of samples, based on a set of their characteristics (Everitt, 2011). The definition of similar is typically based on some form of metric, often a Euclidean (straight line) distance. A cluster model is just that, a model. All cluster models are wrong, some are useful. For this reason, there are many clustering methods used, and none is particularly better than any other overall (Estivill-Castro, 2002).

Here we’ll contrast k-means, with hierarchical clustering before we look at some more general issues of clustering

K-means clustering



K-means clustering is a method of finding k centroids that ‘best explain’ the data given you try and minimise the distance between the centroids and the data. The learning algorithm is very simple and converges quickly to a local optimum. The resulting clustering induces a Voronoi (polygonal) tessellation of the input space.

K-means clustering suffers from several drawbacks:

1. You need to specify the number of clusters, k, in advance. This will rarely be known!
2. Cluster boundaries are linear (the Voronoi polygons).

3. The clusters themselves are sensitive to initialisation – the algorithm only finds a locally best model, so each run may produce a different clustering.
4. The algorithm prefers clusters of equal size due to the learning mechanism.

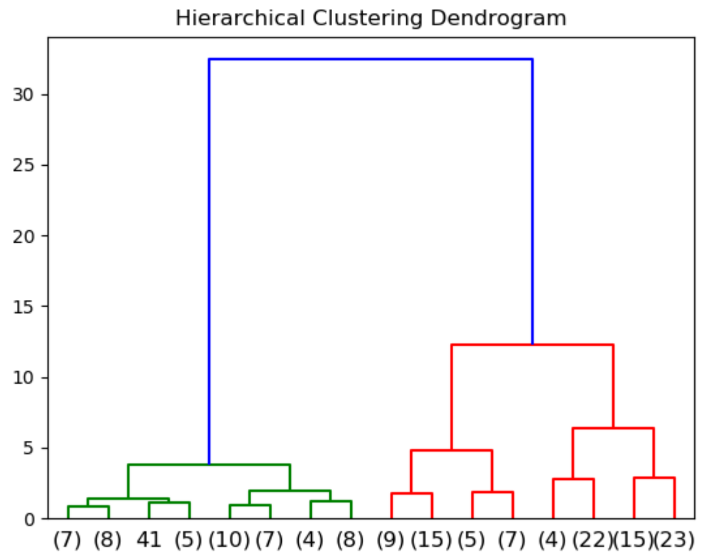
But k-means clustering has some benefits:

1. It is simple to understand and works reasonably well in higher dimensions.
2. It can be used after training to assign new points to clusters without needing to retrain the model.
3. It is fast to train and use.

The benefits are why we added k-means clustering to p:IGI+ first. It is also widely considered to be a good benchmark for other algorithms.

Hierarchical clustering

Hierarchical Cluster Analysis (HCA), sometimes called connectivity-based clustering, aims to achieve the same clustering goal, and typically works agglomeratively (bringing together similar samples) since this is more computationally efficient. As with k-means the user needs to define the distance metric used, but also needs to define the linkage criterion (how to decide when to merge clusters).



The results of hierarchical clustering are typically shown as a dendrogram with the distance between each cluster (based on the linkage criterion) being shown on one axis, and the clusters shown arbitrarily on the other. Typically in HCA, the user selects the ‘cluster distance threshold’ at which to define the ‘clusters’. The drawbacks of hierarchical clustering include:

1. It does not scale well with data set size so cannot be applied to very large data sets.
2. Results can be sensitive to the choice of linkage function – for example Single Linkage (choosing the minimum distance between cluster members to decide whether to merge clusters) works well in some (topological) cases, but Ward linkage (minimising the increase in variance for the clusters being merged) works better for distinct (‘blob’) clusters.
3. The algorithm can be sensitive to outliers / noise, which can significantly change the clustering.
4. It is not easy to assign a new sample to a cluster without ‘re-training’ the model (which *could* produce a very different clustering).

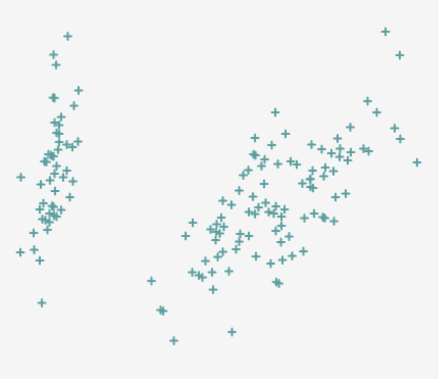
HCA also has some benefits:

1. While the results depend on the distance measure, and linkage criterion, there is no particular assumption on the ‘shape’ or balance of the clusters (there is no model for the cluster boundaries).
2. The user has a visual method for selecting the number of clusters based on the dendrogram.

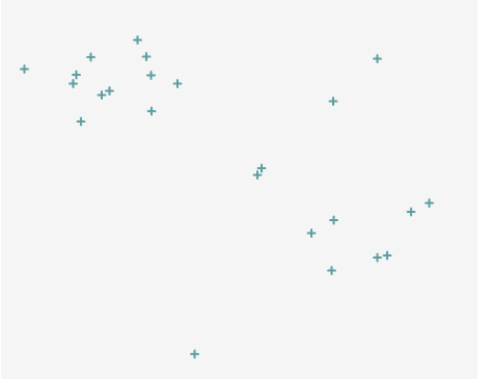
How to decide on the number of clusters?

The first question is what are you aiming to achieve with clustering? Do you have contextual information that might lead you to believe you have clusters, or even a certain number of clusters? Or are you asking a question of only your data?

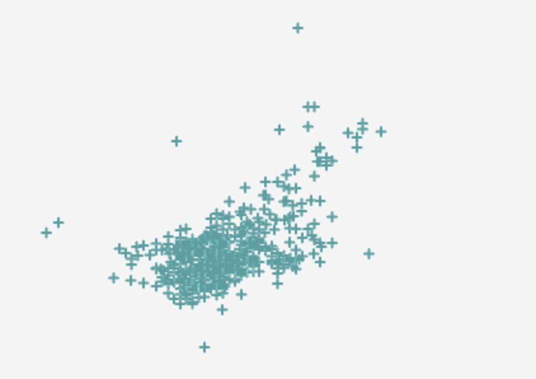
To help us think about this we show 5 data sets below. These are a mixture of real and synthetic data with between 24 and almost 1000 samples. They are deliberately unlabelled, and the axes have been hidden. The first question is can you tell which data sets are real, and which are synthetic?



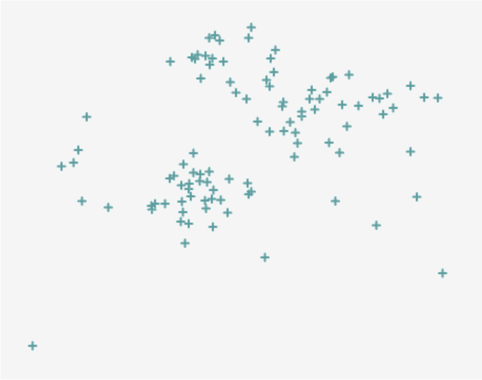
a.



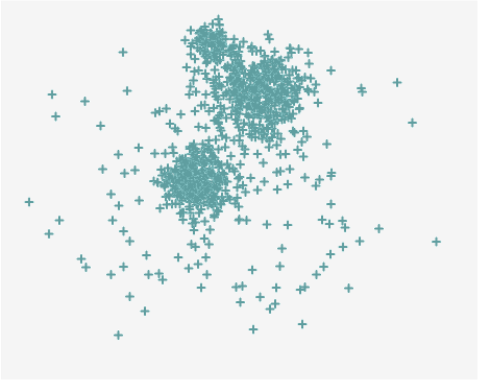
b.



c.



d.



e.

The second question is how many clusters do you think there are in the data sets? The point of this exercise is not to get the right answer, but rather to emphasise that there is no right answer!

Different people will make different interpretations here, just as different clustering algorithms will produce different results.

High dimensional data

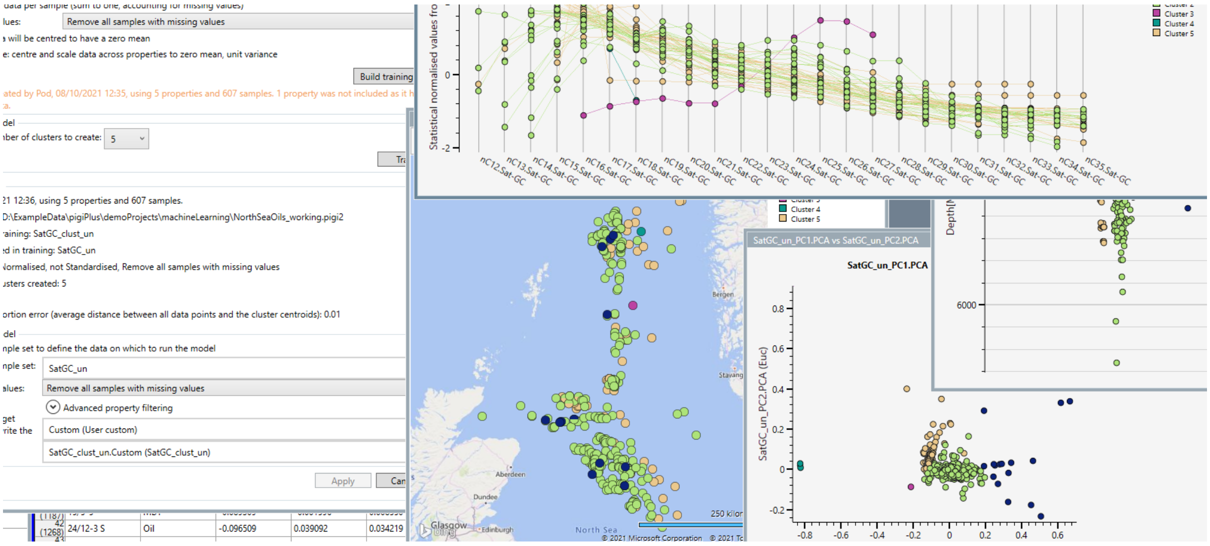
It is worth remembering this is the ‘easy’ problem. We are dealing with 2-dimensional data here and humans are good at visualising that. How might this look with 10 properties (in 10-dimensional space)? Even our intuition of distance can break down in high dimensional spaces – every point is likely to be on the edge of the convex hull bounding the points in high dimensions – that is the points are likely to be closer to the edge of the space than any other point in the data set! Of course, with correlation between the properties the ‘intrinsic dimension’ of the data might be a lot lower – if that is so then undertaking Principal Components Analysis prior to clustering might be a good idea. That’s a topic for another day.

Goodness of fit?

There are error measures which can be used to assess the goodness of a clustering. Most of these are based around some measure of within cluster consistency, against between cluster difference e.g. Dunn (1974). If you are fortunate enough to have labels on your data these can also be used to assess the quality of a clustering (but really then you should be using a classifier!).

We don’t have scope to cover all the different clustering ‘goodness’ measures, but it is important to note there is no single measure or score we can use in all cases. There are some good heuristics such as used in the x-means algorithm (Pelleg & Moore, 2000), and some visual methods such as the ‘elbow method’ which shows the change in % variation explained as a function of the number of clusters (typically using the variance between clusters versus the total variance).

Primarily I would answer the question of the number of clusters, or equivalently how good is a clustering, with “Is it useful in some way?”. A good clustering might not look visually brilliant on a plot, but it might be very helpful during interpretation to focus on a smaller number of subsets of the data that are similar in certain ways. A key aspect of this might be to consider the spatial, depth or temporal structure of the clustered data, which is why the clustering tool in p:IGI+ interacts with the maps, statistics, depth plots, time series plots and more generally all other interpretative graphs.



data per sample (sum to one, accounting for missing values)

lues: Remove all samples with missing values

a will be centred to have a zero mean

e: centre and scale data across properties to zero mean, unit variance

ated by Pod, 08/10/2021 12:35, using 5 properties and 607 samples. 1 property was not included as it h
a.

del

nber of clusters to create: 5

21 12:36, using 5 properties and 607 samples.

D:\ExampleData\pigiPlus\demoProjects\machineLearning\NorthSeaOils_working.pigi2

raining: SatGC_clust_un

ed in training: SatGC_un

Normalised, not Standardised, Remove all samples with missing values

usters created: 5

ortion error (average distance between all data points and the cluster centroids): 0.01

del

nple set to define the data on which to run the model

nple set: SatGC_un

alues: Remove all samples with missing values

get
rite the Custom (User custom)

SatGC_clust_un.Custom (SatGC_clust_un)

42 (1268)	24/12-3 S	Oil	-0.096509	0.039092	0.034219

At the end of the day there is no magic. You must decide whether a particular clustering is useful!

Should you use clustering?

The first question should be is your data even suitable for clustering? If you have multiple oil families in your study region, maybe it is, but if these mix, then (hard) clustering might not be what you want to do. You should consider unmixing here, using something like partial least squares or independent components analysis. Again that is another topic.

The second question we might want to ask is do we have labels for the samples already. If so, we really want to undertake classification. Even if you have only a small proportion of your samples labelled, classification might produce a more useful, and interpretable result. You could even use semi-supervised methods, but that topic is also for another note.

Can you see clusters? The human brain remains significantly more capable than any machine-based mechanism in 2 or 3 dimensions. Beware confirmation bias – seeing what you expect to see. We are good at seeing patterns, indeed we want to see patterns, so we can imagine patterns in data that are simply random! Again, labels can really help here. But in 10 dimensions, what can you really say?

Summary

Clustering is an art. There are many methods and options within these, and no single 'right' answer. In general, the context should be used to inform decisions such as which method to use, how many clusters to select and what parameters to use. The key to effective use is integration with other views onto the data to support decision making. In essence a clustering is 'good' if it helps you understand your data better or improves your decision making.

In part 2 of this note we will discuss a wider range of clustering methods and illustrate, using synthetic data examples where they work well and where they do not.

References

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics. **4**, 95–104.

Estivill-Castro, V. (2002). Why so many clustering algorithms – A Position Paper. ACM SIGKDD Explorations Newsletter, **4** (1), 65–75.

Everitt, B. (2011). Cluster analysis. Wiley, Chichester, U.K.

Pelleg, D. and Moore, A.W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters (PDF). Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000).

Appendix: Answers to the 'clustering' data sets

The type, source and cluster numbers for the datasets shown above are described here.

a) *Real data, based on PCA applied to plant characteristics (PC1 vs PC2 shown). There are 3 clusters in the data based on knowing the plant families.*
b) *Synthetic data generated from 3 clusters, illustrating the challenges of smaller data sets.*
c) *Real data from North Sea oils (carbon isotope ratio, versus Pr/Ph ratio) – unknown number of clusters and mixing suspected. Arguably all one family with variation?*
d) *Synthetic data from four overlapping clusters.*
e) *Synthetic data generated from the same four overlapping clusters, but with many more samples taken.*