Tidsskriftet Tidsskrift for Den norske legeforening

Clusters

MEDICINE AND NUMBERS

JO RØISLIEN

jo@joroislien.no
Jo Røislien, professor of medical statistics at the University of Stavanger and science communicator.
The author has completed the ICMJE form and declares no conflicts of interest.

METTE LANGAAS

Mette Langaas, professor of statistics and deputy head of studies at the Department of Mathematical Sciences, Norwegian University of Science and Technology.
The author has completed the ICMJE form and declares no conflicts of interest.

Not all data sets have explanatory variables and outcomes. The data may nevertheless contain associations that are worth revealing.

In the 2010s, the Intervention Centre at Oslo University Hospital Rikshospitalet worked to develop a computer algorithm that could automatically detect tumours in a radiological image. The result of the computer algorithm was a two-dimensional geometric shape: the outline of a tumour. To check whether the algorithm worked, the outline generated by the automatic method was compared to outlines produced manually by four experienced radiologists. A geometric shape is mathematics, but it is not a *number*, and comparing the outlines of tumours required a different quantitative approach from the one used in traditional statistical methods.

Overlapping

To quantify how similar the different outlines were, the Dice similarity coefficient (1) was used. This is a measure of the degree of overlap between two geometrical figures, with values ranging from 0 to 1 – from none to complete overlap. The four radiologists and the data algorithm created the outline of a tumour in eight radiological images. For all pairs of observations, between the radiologists and the automatic method, the Dice similarity coefficient varied from 0.72 to 0.95, traditionally considered very good overlap. The researchers nevertheless felt that something was not quite right.

Distance

To visualise which geometrical shapes were most similar, agglomerative hierarchical cluster analysis was applied (2). Cluster analysis is a collection of mathematical techniques used to split a data set into groups – so-called clusters – so that the observations within each cluster are more similar to each other than are observations from different clusters.

To produce such clusters, a measure of the similarity between two observations is needed. This is done by measuring distance. It can be Euclidian distance – a straight line that can be measured with a ruler – but other measures of whether things are 'close' to one another or 'similar' can also be used, such as correlation, which indicates the degree of association, or the Dice similarity coefficient.

Dendrogram

The result of a cluster analysis can be visualised in a dendrogram, a tree-like figure where elements that are close to each other (similar) are linked at the bottom of the figure, while elements that are far from each other (dissimilar) are linked further up. In testing the algorithm, the cluster analysis showed that the outlines produced by the automatic method were generally less similar to the radiologists' outlines than the radiologists' outlines were to each other (Figure 1). In other words, the radiologists constituted one cluster, the automatic method constituted another. The automatic method 'saw' another outline of the tumours than the radiologists did (1). The cluster analysis revealed a structure in the data that otherwise would have gone undetected.

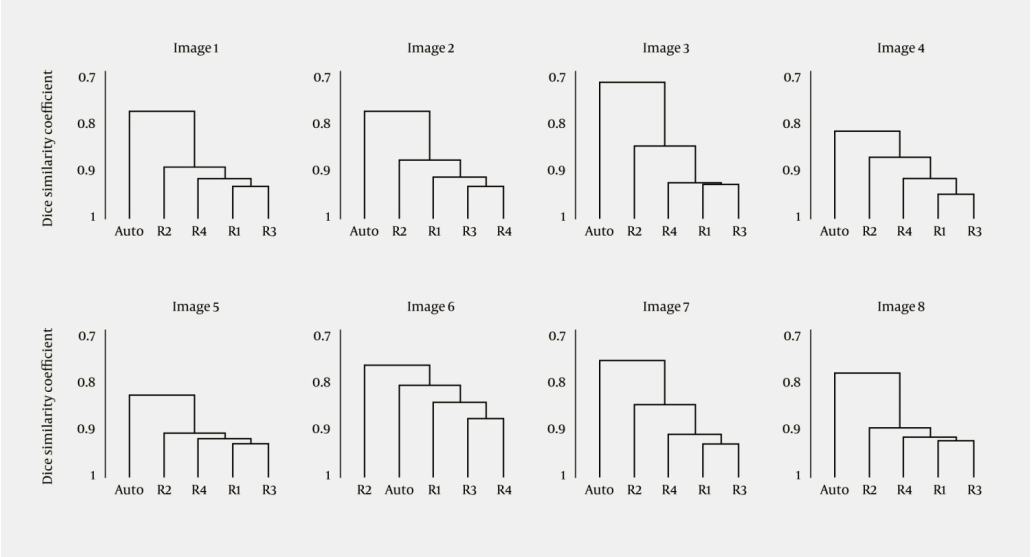

Figure 1 Dendrograms from an agglomerative hierarchical cluster analysis based on data from Røislien and Samset (1). Four radiologists (R1–R4) and an automatic method (Auto) were shown eight radiological images of frozen liver tissue and asked to produce an outline of a tumour in each image. The outlines were subsequently compared using the Dice similarity coefficient as a measure of distance.

Breast cancer

Similar issues are found in many other disciplines, including in genetic research, in which the activity of many genes are often measured in relatively few individuals in order to reveal associations and structures.

In a study published in 2000, the activity of 1 753 genes was analysed in 65 breast cancer tumours (3). Using hierarchical cluster analysis it was discovered that the tumours could be divided into a small number of clusters with different molecular characteristics – so-called *molecular portraits*. It turns out that such molecular portraits can be used to suggest personalised treatment of breast cancer. The PAM50 method, which recommends a treatment based on a patient's gene expression for 50 genes and is used in hospitals worldwide, stems from hierarchical cluster analysis (4).

Learning

Not everything that can be quantified can easily be reduced to a single figure on a number line. For *high-dimensional* observations – such as the outline of a tumour or the simultaneous activity of multiple genes – analytical methods that can learn from data are essential; methods where we feed an algorithm into the computer and let it trawl through the data searching for structures without human interference. The result from such non-guided learning – such as hierarchical cluster analysis – can give valuable insight into the issue that we are studying.

And learning from our quantitative data is exactly what we want to do.

REFERENCES

1. Røislien J, Samset E. A non-parametric permutation method for assessing agreement for distance matrix observations. Stat Med 2014; 33: 319–29. [PubMed][CrossRef]

2. James G, Witten D, Hastie T et al. An introduction to statistical learning. 2. Utg. New York, NY: Springer, 2021.

3. Perou CM, Sørlie T, Eisen MB et al. Molecular portraits of human breast tumours. Nature 2000; 406: 747–52. [PubMed][CrossRef]

4. Pu M, Messer K, Davies SR et al. Research-based PAM50 signature and long-term breast cancer survival. Breast Cancer Res Treat 2020; 179: 197–206. [PubMed][CrossRef]

Publisert: 9 December 2022. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.22.0703