Tidsskriftet
DEN NORSKE LEGEFORENING

Klynger

MEDISIN OG TALL

JO RØISLIEN

jo@joroislien.no
Jo Røislien er professor i medisinsk statistikk ved Universitetet i Stavanger, og vitenskapsformidler.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

METTE LANGAAS

Mette Langaas er professor i statistikk og nestleder for utdanning ved Institutt for matematiske fag ved NTNU.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Ikke alle datasett har forklaringsvariabler og utfall. Allikevel kan det finnes sammenhenger i dataene som er nyttige å avdekke.

På 2010-tallet arbeidet Intervensjonssenteret på Rikshospitalet med å utvikle en dataalgoritme som automatisk kunne finne tumorer i et radiologisk bilde. Resultatet av dataalgoritmen var en todimensjonal geometrisk form: omrisset av en tumor. Om algoritmen fungerte eller ikke, ble fastslått ved å sammenligne omrisset fra den automatiske metoden med omriss laget manuelt av fire erfarne radiologer. En geometrisk form er matematikk, men den er ikke et *tall*, og å sammenligne omriss av tumorer krevde en annen kvantitativ tilnærming enn tradisjonelle statistiske metoder.

Overlapping

For å tallfeste hvor like de ulike omrissene var, benyttet man Dice-koeffisienten (eng. *Dice similarity coefficient*) (1), et mål på grad av overlapping mellom to geometriske figurer som tar verdier fra 0 til 1 – fra ingen til fullstendig overlapping. I åtte radiologiske bilder laget de fire radiologene og dataalgoritmen omrisset av en tumor. For alle parene av observasjoner – mellom radiologene og den automatiske metoden – varierte koeffisienten fra 0,72 til 0,95, tradisjonelt ansett som utmerket samsvar. Allikevel følte forskerne at noe skurret.

Avstand

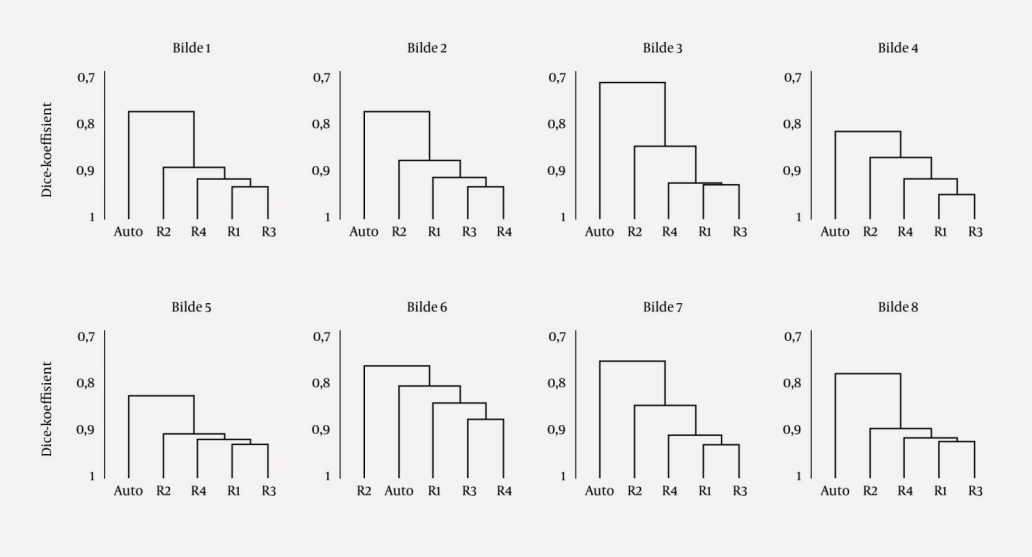
For å visualisere hvilke geometriske omriss som var likest, benyttet man agglomerativ hierarkisk klyngeanalyse (2). Klyngeanalyse er en samling matematiske teknikker for å dele et datasett inn i grupper – såkalte klynger – slik at observasjoner innenfor hver klynge er likere hverandre enn observasjoner fra ulike klynger.

For å lage slike klynger trengs et mål på hvor like to observasjoner er. Det gjøres ved å måle avstand. Dette kan være euklidsk avstand – en rett linje man kan måle med linjal – men også andre mål på om ting ligger «nær» hverandre eller er «like», kan benyttes, slik som korrelasjon, som sier noe om grad av samvariasjon – eller Dice-koeffisienten.

Dendrogram

Resultatet av en klyngeanalyse kan visualiseres i et dendrogram: en trelignende figur der elementer som er nær hverandre (ligner), er koblet sammen nederst i figuren, mens elementer som er lenger fra hverandre (ikke ligner), er koblet sammen høyere oppe.

I uttestingen av algoritmen viste klyngeanalysen at omrissene til den automatiske metoden generelt lignet mindre på omrissene til radiologene enn radiologenes omriss lignet på hverandre (figur 1). Radiologene utgjorde én klynge, den automatiske metoden en annen. Den automatiske metoden «så» altså et annet omriss av tumorene enn radiologene (1). Klyngeanalysen avdekket en struktur i dataene man ellers ikke hadde oppdaget.



Figur 1 Dendrogrammer fra agglomerativ hierarkisk klyngeanalyse basert på data fra Røislien og Samset (1). Fire radiologer (R1–R4) og en automatisk metode (Auto) ble vist åtte radiologiske bilder av frossent levervev og bedt om å lage omriss av en kreftsvulst i hvert bilde. Omrissene ble deretter sammenlignet ved hjelp av Dice-koeffisienten som avstandsmål.

Brystkreft

Tilsvarende problemstillinger finnes i flere fagfelt, blant annet i genforskning, der man gjerne måler aktiviteten til mange gener i relativt få individer for å avdekke sammenhenger og strukturer.

I en studie fra 2000 analyserte man aktiviteten til 1 753 gener fra 65 brystkreftsvulster (3). Ved hjelp av hierarkisk klyngeanalyse fant man at svulstene kunne deles inn i noen få klynger med ulike molekylære karakteristika – såkalte *molekylære portretter*. Det viser seg at slike molekylære portretter kan brukes til å foreslå persontilpasset behandling av brystkreft. Metoden PAM50, som anbefaler behandling basert på en pasients genuttrykk for 50 gener, og som brukes på sykehus verden over, stammer fra hierarkisk klyngeanalyse (4).

Læring

Ikke alt som kan kvantifiseres, kan uten videre reduseres til ett enkelt tall på en tallinje. For *høydimensjonale* observasjoner – som omrisset av en tumor eller aktiviteten til mange gener samtidig – er analysemetoder som kan lære av data, viktige. Metoder der vi fôrer datamaskinen med en algoritme og lar den tråle gjennom dataene på jakt etter strukturer uten innblanding fra mennesker. Resultatet fra slik ikke-veiledet læring – som hierarkisk klyngeanalyse – vil kunne gi verdifull innsikt i problemstillingen vi studerer.

Og det å lære av tallene våre er vel akkurat det vi vil.

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. desember 2022. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.22.0703