Lys hjælper os med at forstå data – Fra kvantesamples til klarhed i datastrukturer

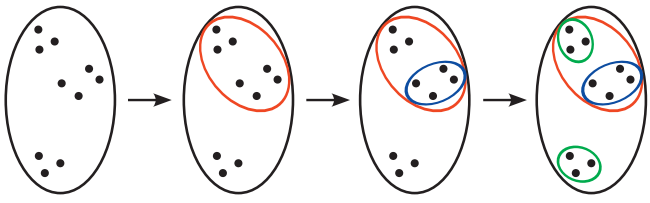
Josefine Bjørndal Robl, Institut for Fysik og Astronomi, Aarhus Universitet

Verden bombarderer os hele tiden med information. Den vælter ind gennem vores skærme, målinger og samtaler – vi drukner i mængden. Og alligevel er vi i stand til at finde mening i den. Hvorfor? Fordi vi ser mønstre.

At kunne finde struktur i en uoverskuelig kompleksitet er ikke blot en menneskelig evne – det er en nødvendighed for at forstå den verden, vi lever i. I dataanalyse kaldes denne opgave for *clustering*; kunsten at gruppere datapunkter, der ligner hinanden, uden på forhånd at vide, hvad vi leder efter. Men selv om idéen lyder enkel, så er implementeringen det langt fra, især når datasættene bliver store.

Hvordan vi finder strukturer ved at splitte

En intuitiv tilgang til clustering er den såkaldte *divisive hierarchical clustering* metode [1]. Her starter alle datapunkter i én samlet gruppe kaldet en *supercluster*. Metoden forsøger herefter at opdele denne gruppe i mindre *subclusters* ved at undersøge grupperinger af et faldende antal punkter. Hvis vi eksempelvis starter med 10 datapunkter som i figur 1, så leder metoden først efter mulige *subclusters* bestående af 9 punkter, dernæst 8, 7 og så videre. Når en gruppe opfylder et foruddefineret tæthedskriterium, så markeres den som en subcluster – på figur 1 er dette eksempelvis tilfældet for de 7 punkter omgivet af en rød ring. Herefter fortsætter metoden, men nu udelukkende inden for de etablerede grupper, i vores eksempel enten inden for den røde gruppe med de 7 datapunkter eller inden for de resterende 3 datapunkter.

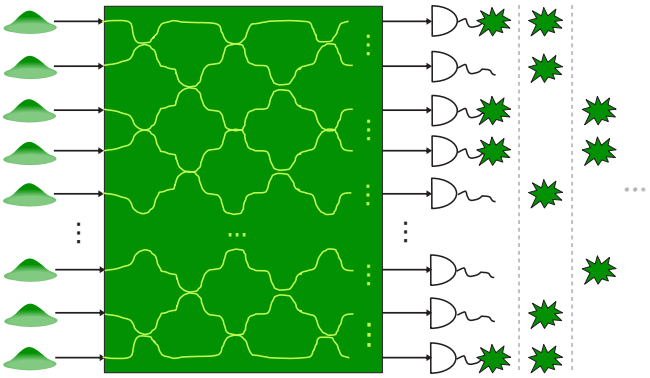


Figur 1. I divisive hierarchical clustering starter alle datapunkter i én stor cluster (sort), her af størrelse 10. Metoden tjekker nu iterativt om der er clusters af størrelse 9, dernæst 8, 7, osv.. Når et antal punkter, fx 7, ligger tættere end andre kombinationer af samme størrelse, så markeres de som en mindre subcluster (rød), og processen foregår inde i hver af disse; de 7 punkter i en subcluster og de resterende 3 punkter. Efter sidste tjek viser algoritmen et hierarki af clusters.

Denne metode er også kendt som en top-down-tilgang, da den gradvist afslører datastrukturer i flere lag – fra store overordnede grupper til små tætte clusters. Dette kommer dog med en pris: Antallet af mulige opdelinger vokser eksponentielt med antallet af datapunkter. Hvis man vil at finde de optimale grupperinger, så skal metoden gennemgå et enormt antal mulige kombinationer, hvilket hurtigt bliver beregningsmæssigt uoverskueligt.

Fotoniske kvanteberegninger

Men hvad gør man, når de klassiske metoder ikke længere rækker? Når datasættene bliver så store, at det bliver urealistisk at gennemgå alle mulige opdelinger? I stedet for at kæmpe imod kompleksiteten med mere klassisk regnekraft må man skifte spor og vende sig mod kvanteverdenen, hvor der findes beregningsmodeller, der udnytter de fundamentale fysiske principper på en helt anden måde. Én af disse er *Gaussian boson sampling*: En specialiseret kvanteberegningsmodel, som benytter fotoner til at generere output baseret på komplekse statistiske fordelinger [2].



Figur 2. En Gaussian Boson Sampler består af et lineært interferometer, hvori der indsendes squeezed vacuum, og hvor output er et fotonmønster ved hjælp af fotondetektorer, der registrerer antallet af fotoner i hver udgang. Når man sampler fra en Gaussian Boson Sampler, vil antallet af målte fotoner og hvilke detektorer, der registrerer dem, variere. Dette skyldes en underliggende sandsynlighedsfordeling, der er for kompleks til at kunne simuleres klassisk.

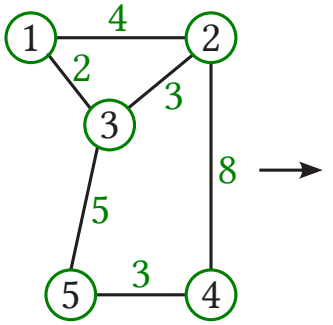
I en Gaussian Boson Sampler (figur 2) sendes *squeezed vakuumtilstande* – ikke-klassiske kvantetilstande af lys med reduceret kvantestøj i enten position eller impuls og forøget støj i den anden – ind i et lineært optisk interferometer. Her spredes og interfererer lyset gennem en række beam-splitters og phase-shifters, hvorefter det måles af fotondetektorer, som registrerer antallet af fotoner i hver udgang. En sådan måling kaldes en *sample* – et specifikt fotonmønster, der angiver, hvilke detektorer har målt hvor mange fotoner.

Hvis vi sampler fra en sådan Gaussian Boson Sampler, altså gentager eksperimentet mange gange, vil vores samples ændre sige både i antal og i selve mønsteret. Sampler vi længe nok, så vil vi se, at disse samples nærmer sig en kompleks sandsynlighedsfordeling. Og det er netop strukturen i denne sandsynlighedsfordeling, som gør Gaussian boson sampling så interessant.

Sandsynligheden for at måle de forskellige samples afhænger nemlig af en underliggende matrixfunktion kaldet *Hafnianen* og måden denne beregnes på gør den meget svær at beregne klassisk.

Kan fotoner gøre clustering nemmere?

Men hvordan hænger denne sandsynlighedsfordeling sammen med clustering? For at besvare dette spørgsmål skal vi først tale om grafer – matematiske objekter bestående af et antal knuder og kanter mellem dem. Som eksempel kan vi tage grafen på figur 3, hvor vi kan se, at alle punkterne er forbundet med kanter, som hver især er vægtet baseret på eksempelvis punkternes afstand til hinanden. En sådan graf kan beskrives ved en matrix kaldet en *adjacency matrix*, som på elegant vis repræsenterer en grafs struktur uden at tegne den.



	1	2	3	4	5
1	0	4	2	0	0
2	4	0	3	8	0
3	2	3	0	0	5
4	0	8	0	0	3
5	0	0	5	3	0

Figur 3. En graf består af en række knuder forbunder med kanter. Disse kanter kan være vægtede med baseret på eksempelvis afstanden mellem knuderne. En sådan graf kan beskrives ved dens adjacency matrix, som har 0'er, hvis der ikke er en kant mellem to knuder, og vægtningen af kanten, hvis der er en kant.

I dette grafteoretiske billede kan hvert fotonmønster, altså hver sample, tolkes som en unik *subgraf* – et udsnit af knuder og deres indbyrdes kanter – af den oprindelige graf. Det vil sige, at hver gang vi måler en sample, så svarer det til at udvælge en bestemt gruppe af datapunkter og deres indbyrdes værdier.

Det interessante er netop, *hvilke* subgrafer der optræder hyppigst. Det er nemlig vist, at sandsynligheden for at måle en given subgraf afhænger direkte af dens tæthed [3], da denne egenskab er relateret til Hafnianen. Jo tættere en subgraf er, desto større er sandsynligheden for, at det tilsvarende fotonmønster bliver samplet.

Og lige præcis dette interne bias er nøglen til at løse vores clusteringproblem. For det er netop sådanne tætte grupperinger, vi forsøger at finde i clustering. Denne observation er grundlaget for min forskning, hvor jeg har kombineret samples fra en Gaussian Boson Sampler med den hierarkiske clusteringmetode for at udvikle en *sampling-baseret clustering-algoritme*. Min forskning har ikke blot vist, at denne metode faktisk virker og kan identificere clusters i en graf, men antyder også, at dette muligvis kan gøres uden brugen af eksponentielt mange samples [4]. Denne indikation er vital, da dette vil betyde, at vi ikke blot udnytter, at tætte subgrafer er mere sandsynlige at sample, mens ikke-tætte subgrafer er mere usandsynlige, men i stedet at de fleste ikke-tætte subgrafer er *så* usandsynlige, at de kan negligeres og vi altså kun behøver at kigge på et lille udsnit af subgrafer for at finde clusters. Dette er derfor stadig et åbent spørgsmål, jeg aktivt undersøger i min forskning.

Et kig fremad

Et naturligt næste skridt er at undersøge, om denne metode bevarer sin styrke i mere realistiske scenarier. Indtil videre er metoden kun testet med i støjfrie simuleringer, men skal den gøre sig gældende som et reelt værktøj, må den være robust og kunne fungere på virkelig kvantehardware, hvor støj og tekniske begrænsninger er uundgåelige.

Vi befinder os lige nu i *NISQ*-æraen; en tid hvor kvantecomputere stadig er små og udsatte for fejl, men alligevel kraftige nok til at løse problemer uden for klassiske computeres rækkevidde. Fotoniske platforme, hvilke Gaussian boson sampling baserer sig på, er blandt de mest modne eksperimentelle platforme, hvilket gør det oplagt at undersøge, hvordan man kan udnytte deres naturlige bias i praktiske algoritmer. Her handler det ikke nødvendigvis om at opnå universelle kvanteberegninger, men lige så meget om at finde nichemodeller, der gør nytte af de begrænsede ressourcer, vi har til rådighed. Netop derfor er Gaussian boson sampling interessant: Det er ikke et forsøg på efterligne en klassisk beregning, men i stedet en model som udnytter sit eget bias til at løse et klassisk hårdt problem.

Selvom metoden stadig befinder sig i et udviklingsstadie, hvor yderligere forskning er nødvendig, demonstrerer den, hvordan kvanteberegninger kan tilføre nye perspektiver til komplekse problemer. Det understreger vigtigheden af at fortsætte udforskningen af disse metoder, da de kan bane vejen for innovative løsninger – selv inden for de nuværende teknologiske begrænsninger.

Litteratur

[1] L. Kaufman og P. J. Rousseeuw (1990) "Finding Groups in Data: An Introduction to Cluster Analysis", *John Wiley & Sons*.

[2] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn og I. Jex (2017) "Gaussian Boson Sampling", *Phys. Rev. Lett.*, bind **119**, nr. 17, side 170501.

[3] J. M. Arrazola og T. R. Bromley (2018) "Using Gaussian Boson Sampling to Find Dense Subgraphs", *Phys. Rev. Lett.*, bind **121**, nr. 3, side 030503.

[4] J.B. Robl, F.K. Marqversen, A.B. Michelsen (under forberedelse) "Exponential Speedup of Divisive Hierarchical Clustering through Gaussian Boson Sampling".

Josefine Bjørndal Robl er ph.d-studerende ved Institut for Fysik og Astronomi på Aarhus Universitet samt ved Kvantify. Hun arbejder med Gaussian boson sampling og dets anvendelser inden for clustering.