Improved strategies for distance based clustering of objects on subsets of attributes in high-dimensional data

Kampert, M.M.D.

Summary in Dutch (Samenvatting)

Dit proefschrift richt zich op het clusteren van objecten in hoogdimensionale data, waarbij de aanname geldt dat de objecten niet op alle attributen clusteren, zelfs niet op een enkele subset van attributen, maar vaak op verschillende subsets van attributen. Met het doel om een dergelijke clusterstructuur te kunnen vinden, stelden Friedman en Meulman (2004) een raamwerk voor met een specifiek algoritme COSA genaamd. In plaats van rechtstreeks clusters te produceren, geeft COSA een representatieve afstandsmatrix die vervolgens kan worden geanalyseerd met behulp van een groot aantal methoden voor afstandsanalyse, zoals hiërarchische clustering of multidimensional scaling analyse. Het doel van dit proefschrift is om het gedrag van COSA te bestuderen zodat we zowel het nut als de zwakke plekken kunnen aantonen en verbeteringen kunnen voorstellen die de zwakke punten aanpakken.

Hoofdstuk 1

Belangrijke begrippen voor COSA zijn cluster analyse, gereguleerde weging van de attributen in hoogdimensionale data, en afstanden. Een cluster analyse kan worden uitgevoerd om inzicht te krijgen in data en om hypothesen te kunnen genereren, of om anomalien en opvallende attributen te detecteren. Cluster analyse is het zoeken naar natuurlijke groeperingen in een dataset van objecten die worden gemeten op attributen, terwijl de groepenlabels van de objecten ontbreken. Binnen deze natuurlijke groepen zijn de objecten homogeen, en zijn de groepen bij voorkeur erg verschillend van elkaar wat betreft de waarden op de attributen.

Omdat we kunnen aannemen dat er veel irrelevante attributen zijn die niet bijdragen aan de clustering van een groep in hoogdimensionale data, dient de cluster analyse een strategie te hanteren waarmee de attributen gewogen kunnen worden. Met name zou elk cluster een eigen subset van relevante attributen kunnen hebben. De belangrijkste uitdaging voor dit soort strategieën voor het wegen van attributen is ervoor te zorgen dat de cluster analyse stabiele resultaten produceert. Door de oplossingsruimte voor de waarden van de gewichten voor ieder attribuut te beperken, kan een dergelijke stabiliteit worden vergroot.

Voor elk type cluster analyse is er altijd een notie van (dis)similariteit tussen de ob-

jecten. In COSA zijn dit de representatieve afstanden. De COSA afstanden bieden de mogelijkheid om clusters te onthullen die zijn gevormd in bepaalde subsets van de attributen in hoogdimensionale data. Multidimensional scaling analyse (MDS), evenals het dendrogram dat resulteert uit hiërarchische clustering, zijn bruikbare technieken waarmee deze specifieke clusterstructuren kunnen worden onthuld in een visualisatie van de COSA-afstanden.

Hoofdstuk 2

Het COSA-algoritme dat centraal staat in dit proefschrift afgeleid van het algoritme uit Friedman en Meulman (2004) dat gebruik maakt van een K naaste buren strategie, en hier COSA-KNN genoemd wordt. Het belangrijkste doel van COSA-KNN is om een afstandsmatrix te produceren die de afstanden tussen N objecten bevat. Een afstand tussen object i en j, aangeduid met D_{ij}, en is opgebouwd uit een gewogen som van P afstanden, één afstand voor elk attribuut: de attribuutafstand. Deze P attribuutafstanden, aangeduid met $\{d_{ijk}\}_{k=1}^{P}$, worden op hun beurt weer opgebouwd uit metingen verzameld in een $N \times P$ dataset. Er wordt aangenomen dat de dataset een onderliggende clusterstructuur heeft waarin de objecten zijn geclusterd op clusterspecifieke subsets van de attributen. Een voordeel hierbij is dat het niet noodzakelijk is om in COSA-KNN het verwachte aantal clusters op te geven; dit wordt namelijk slim omzeild door gebruik te maken van de K naaste buren strategie.

De attribuutafstanden worden gewogen op basis van een λ-gereguleerde beperking op de *negatieve entropie* van de attribuutgewichten. Deze specifieke beperking zorgt ervoor dat COSA-KNN bruikbare oplossingen kan bieden voor de attribuutgewichten van elk cluster in de hoogdimensionale data waarbij P veel groter is dan N. De definitie van de afstandsmatrix zelf zorgt ervoor dat de afstanden voor de objectparen binnen een cluster altijd kleiner zijn dan de afstanden voor de objectparen in verschillende clusters. We zullen deze eigenschap van de afstanden aanduiden als 'majorizing '.

De 'majorizing'-afstanden worden verkregen uit een iteratief algoritme dat op heuristische wijze het COSA-KNN-criterium minimaliseert. We beginnen met een eerste set attribuutgewichten waaruit een eerste afstandsmatrix kan worden afgeleid. Vervolgens worden er op basis van de K naaste buren methode nieuwe attribuutgewichten berekend, waaruit weer een nieuwe afstandsmatrix kan worden afgeleid, enzovoort. Deze iteratieve procedure wordt voortgezet totdat een bepaald convergentiecriterium is bereikt. Binnen dit iteratieve proces gebruikt COSA-KNN een strategie dat gebruikt maakt van de homotopie tussen het criterium van COSA-KNN en een 'criterium bij benadering', zodat het algoritme op elegante wijze ongewenste lokale minima kan vermijden.

Hoofdstuk 3

De keuze van de waarden voor twee instellingsparameters in COSA is cruciaal om een subtiele clusterstructuur in de data te kunnen detecteren. Deze twee paramteters zijn λ en K. De definitie van λ is het bereik van de waarden die een groep van objecten op een attribuut kenmerken, en het heeft rechtstreeks invloed op de waarden van de

attribuutgewichten. De definitie van K is de grootte van het aantal naaste buren (de buurt) voor elk object in een cluster.

Het aantal succesvolle waarden van de parameters kan worden vergroot wanneer we een robuuste versie van COSA-KNN toepassen op een dataset. In de robuuste versie van COSA-KNN gebruiken we op de mediaan gebaseerde schattingen voor de attribuutgewichten, waar in de oorspronkelijke versie van COSA-KNN attribuutgewichten gebaseerd zijn op het gemiddelde. De robuuste versie van COSA-KNN is sneller en beter in het ontdekken van clusterstructuren in de data.

Om succesvolle combinaties van de waarden van de parameters automatisch te vinden, wordt de zogenaamde Gap-statistiek procedure gebruikt (gebaseerd op Tibshirani et al., 2001). De Gap-statistiek procedure is een permutatiemethode waarin we een specifieke combinatie van parameters selecteren die de grootste kloof (= Gap) biedt tussen enerzijds de waarde van het robuuste COSA-KNN criterium dat behoort bij de dataset, en anderzijd de benadering van de verwachte waarde die hoort bij een vergelijkbare dataset zonder clusterstructuur.

Hoewel het niet-robuuste COSA-KNN criterium een concaaf vlak oppervlak lijkt te geven voor een selectie van waarden voor zowel λ als K, toont de robuuste versie van het criterium een zigzagpatroon over de oneven en even waarden voor K. Dit specifieke zigzagpatroon is ook zichtbaar in de Gap-statistiek waarden, wat leidt tot de voorkeur voor even waarden voor K boven oneven waarden voor K. Omdat de schattingen op basis van even waarden voor K over het algemeen stabieler zijn, raden we aan om alleen met even waarden voor K te werken.

Hoofdstuk 4

We verhogen de kracht van COSA-KNN door deze vier verbeteringen in te voeren:

i. met een verandering in de notatie in de definitie van het cluster probleem voor COSA uit hoofdstuk 2, kunnen we het probleem ook generalizeren naar situaties waarbij de aanname van elkaar uitsluitende clusters niet nodig is;

ii. in plaats van λ de negatieve entropie van de attribuutgewichten voor elke buurt te laten reguleren, kunnen we λ de Kullback-Leibler divergentie tussen de attribuutgewichten en een vooraf gespecificeerde set attribuutgewichten laten reguleren. Op deze manier kunnen we vooraf gespecificeerde attribuutgewichten op laten nemen in de analyse die het belang van elk attribuut in de clustering aangeven;

iii. we herformuleren het criterium om ervoor te zorgen dat er ook attribuutgewichten een nulwaarde zullen krijgen, aangedreven door λ;

iv. we passen de COSA-afstand aan zodat deze beter is in het scheiden van objectenparen in verschillende clusters.

De originele versie van COSA-KNN is niet goed in staat om clusters te detecteren die hoofdzakelijk verschillen in hun gemiddelde op attributen, en binnen-clustervarianties hebben die gelijk zijn aan die van de ruisobjecten. Dit soort clusterstructuren kunnen nu ook door COSA-KNN gevonden worden wanneer deze verbeteringen

zijn doorgevoerd in COSA-KNN. Oftewel, deze verbeteringen maken COSA-KNN flexibeler en krachtiger in meer situaties.

In COSA-KNN krijgt elk object hetzelfde aantal K naaste buren toegewezen. In plaats van de grootte van elke buurt in te stellen op een vaste waarde voor K, kunnen we de grootte van elke buurt ook laten bepalen door de parameter λ. Op deze manier maken we K dus overbodig. Deze benadering en het bijbehorende algoritme noemen we COSA-λNN.

De prestaties van COSA-λNN zijn even goed als de verbeterde versie van COSA-KNN, zo niet beter. Het voordeel van COSA-λNN ten opzichte van COSA-KNN is niet alleen dat elke buurt een andere grootte kan hebben, het gaat ook gepaard met minder rekentijd, aangezien we niet meer een succesvolle waarde voor K hoeven te zoeken. Terwijl we naar een succesvolle waardecombinaties zoeken van λ en K in COSA-KNN, hoeft dit in COSA-λNN alleen nog maar voor λ te gebeuren.

Hoofdstuk 5

Alhoewel de belangrijkste output van COSA een "cluster-happy" afstandsmatrix is, is het niet vanzelfsprekend hoe L-groepen uit deze matrix te extraheren. COSA op zichzelf kan niet automatisch worden vergeleken met cluster algoritmes die ten doel hebben om L groepen te vinden. Om L clusters uit de COSA-afstandsmatrix te extraheren, stellen we een nieuw algoritme voor, genaamd MVPIN (Minimum Variance strategy to Partition In Neighborhoods).

MVPIN is een beperktere vorm van Ward's methode. De beperking is gebaseerd op de populaire Jarvis en Patrick (1973) similariteitsmaat, het aantal overeenkomstige naaste buren. Bij het uitvoeren van Ward's minimale variantie methoden voegen we alleen twee cluster samen wanneer de similariteitsmaat voor deze twee clusters het hoogste is en ook hoger is dan een bepaalde vastgestelde drempel.

Vergeleken met concurrerende cluster algoritmes zoals Partitioning Around Medoids (PAM; Kaufman & Rousseeuw, 1987) of hiërarchische clustering gebaseerd op 'average linkage' of Ward's methode, laat een eerste onderzoek zien dat MVPIN beter presteert in het detecteren van de clusterstructuur voor prototypische datasets voor COSA. De prestaties van MVPIN zijn met name goed wat betreft het detecteren van kleine homogene clusters die erg vergelijkbaar zijn met elkaar omdat ze een groot overlappend clusteringspatroon hebben op een bepaalde subset van attributen. In combinatie met de Gap-statistiekprocedure laat MVPIN over het algemeen betere resultaten zien dan die van PAM en Ward's methode op de (geoptimaliseerde) afstanden van COSA-λNN toegepast op 12 benchmark datasets.

Hoofdstuk 6

COSA is grotendeels gemotiveerd om toegepast te worden op omics-data, net als een aantal van de huidige geavanceerde L-groepen cluster algoritmes die een vorm van gereguleerde weging van de attributen bevatten. Dit soort L-groepen cluster algoritmes die ofwel geïnspireerd door COSA, dan wel vergeleken met de oorspronkelijke COSA, dit zijn, o.a., Entropy Weighted K-means clustering (EWKM; Jing, Ng, &

Huang, 2007), Sparse clustering (SPARCL; Witten & Tibshirani, 2010), en Simple Approach to Sparse clustering (SAS; Arias-Castro & Pu, 2017). Wanneer we de resultaten van deze algoritmes repliceren en vervolgens vergelijken met de resultaten van MVPIN dat is toegepast op de oorspronkelijke COSA- KNN afstanden (Hoofdstuk 3) en de COSA-λNN afstanden (Hoofdstuk 4), dan zien we dat COSA-λNN en SAS over het algemeen het beste presteren. De vergelijking is gebaseerd op 11 omics benchmark-datasets.

In plaats van COSA-λNN en de andere algoritmes als concurrenten van elkaar te framen, kunnen we ze ook gebruiken om elkaars resultaten te valideren. Gegeven een bekend veronderstelde clusterstructuur, kunnen we een optimale COSA-afstand en optimale cluster attribuutgewichten berekenen op basis van een zelf-geleerde cluster-specifieke parameter λ_l. Vervolgens kunnen de zelf-geleerde cluster-attribuutgewichten samen met attribuut-belangrijkheidsmaten worden gebruikt om te filteren op de attributen waarop een cluster homogeen is (het attribuut waarop het cluster een lage variantie heeft). De validatie van de clusters op basis van het COSA-raamwerk kan als een permutatietest gebruikt worden.

Hoofdstuk 7

Dit proefschrift wordt afgesloten met een discussiehoofdstuk waarin een aantal onderwerpen worden besproken. Hieronder vallen mogelijk tekortkomingen met betrekking tot de keuze van de simulatievoorbeelden, de studie van de rekenkosten voor COSA-KNN en COSA-λNN, de studie naar convergentie-eigenschappen, en hoe ontbrekende data binnen het COSA-raamwerk kunnen worden behandeld. Wat er verder nog wordt besproken zijn mogelijke toekomstige verbeteringen en uitbreidingen. Onder de mogelijke verbeteringen wordt er gekeken naar een andere regularisatiestrategie voor de attribuutgewichten, de COSA-afstanden en de afstanden op het niveau van de attributen. We bespreken ook hoe het COSA-raamwerk kan worden uitgebreid tot een raamwerk voor het schatten van ontbrekende data, Self-Organizing Maps (Kohonen, 1980), en Point of View Analysis (Tucker & Messick, 1963; Meulman & Verboon, 1993). Deze tekortkomingen, verbeteringen en uitbreidingen tezamen zijn een pleidooi voor verdere studie van COSA.