Statistical methods for mass spectrometry-based clinical proteomics

Kakourou, A.A.

Samenvatting

Dit proefschrift heeft betrekking op de ontwikkeling van nieuwe statistische methoden voor het construeren van diagnostische voorschriften op basis van proteomische gegevens verkregen met behulp van klinische massaspectrometrie. Onze interesse richt zich specifiek op de vergelijking van massa spectrometrische proteomische profielen gemeten in serum stalen van gezonde individuen en kankerpatienten in de context van case-control studies. Een hoofddoelstelling in zulke studies is de constructie van discriminerende regels voor het onderscheiden van individuen ten aanzien van de aanwezigheid of afwezigheid van de ziekte, evenals voor het voorspellen van de gezondheidsstatus van de toekomstige patienten.

In dit proefschrift onderzoeken we de analyse van twee verschillende types proteomische data in case-controle onderzoek, verkregen met twee verschillende massaspectrometrische technologieen. Het eerste databestand betreft een borstkanker case-controle onderzoek waarin het proteomische profiel van elk individu gegenereerd wordt met behulp van MALDI-TOF massaspectrometrie. De tweede dataset die we analyseren bestaat uit proteomische expressiemetingen van gezonde individuen en alvleesklierkanker patienten door middel van MALDI-FTICR-technologie. In het hierna volgende, bespreken we kort de nieuwe statische methoden die in deze thesis worden ontwikkeld en beschreven.

Hoofdstuk 1 geeft een algemene inleiding in klinisch proteomisch onderzoek. We beginnen met het introduceren van de basisprincipes van het meetproces waarmee de data worden gegenereerd, evenals de voorbewerkingsstappen die noodzakelijk zijn om de onbewerkte proteomische expressie te vertalen in een dataset die geschikt is voor verdere analyse. Vervolgens beschrijven we de speciale kenmerken van de MS-proteomische data alsook de statistische uitdagingen die zich kunnen voordoen bij het analyseren. We bespreken daarna de basiscomponenten van proteomische diagnose. Deze omvatten: a) de keuze van het model, b) de kalibratie van de diagnostische regel en c) de evaluatie van de diagnostische regel. We geven een gedetailleerde beschrijving van elk van deze stappen.

Hoofdstukken 2 en 3 beschrijven combinatie methoden om de prestatie van classificatie regels in klinisch diagnostisch onderzoek te verbeteren - ofwel door verschillende classificatie modellen te combineren - ofwel door combinatie van verschillende data sets per patient in de kalibratie van de diagnostische regel. Hoofdstuk 2 presenteert methodiek voor het combineren van een verzameling verschillende classificatiemodellen. Hoofdstuk 3 behandelt verschillende methoden, waaronder een aanpassing van de combinatieaan-

pak van hoofdstuk 2, voor het combineren van meerdere soorten of sets omics data per patient. De voorgestelde combinatiemethoden worden geillustreerd in de analyse van de MALDI-TOF case-control borstkanker data.

Hoofdstuk 2 bespreekt combinatie van verschillende classificatie regels voor verbeterde voorspellingen. In eerste instantie worden individuele classificatie modellen apart gekalibreerd, elk op basis van de gehele proteomische dataset. Een dubbelkruisvalideringsaanpak wordt gebruikt om de onafhankelijke classificatieregels te kalibreren en onbevooroordeelde schattingen van de klasse waarschijnlijkheden te verkrijgen, die later gebruikt worden als nieuwe predictie variabelen in de constructie van de gecombineerde classificatieregel. We presenteren twee verschillende methoden om de geschatte klasse waarschijnlijkheden te combineren en zo de uiteindelijke combinatie-gebaseerde voorspelregel op te stellen. De eerste aanpak is gebaseerd op gewogen sommen en is eenvoudig in toepassing daar het geen optimalisatie van gewichten vereist. Een onbevooroordeelde evaluatie van de prestaties van zulke gewogen-som combinatieregel is dan relatief gemakkelijk af te leiden, aangezien deze combinatie methode de kruisvalidatie van de individuele voorspellingen ongemoeid laat. In de tweede methode worden de afzonderlijke gekruisvalideerde voorspellingen op basis van elk afzonderlijk model zelf als "nieuwe" predictie-data gebruikt in de kalibratie van een combinerend "meta-model". Deze methode is daardoor meer flexibel. Echter, aangezien het combinerend model zelf ook kalibratie nodig heeft, is bijkomende kruis-validatie noodzakelijk om onbevooroordeelde evaluatie te verkrijgen van het uiteindelijke gecombineerde model. De combinatie ideeen gepresenteerd in dit hoofdstuk zijn deels gemotiveerd door consultatie, maar ook door suggesties en discussie gegenereerd in de proteomic massaspectrometrie competitie (Mertens, 2008; Hand, 2008). Een belangrijk resultaat van de competitie was dat de meerderheid van de deelnemers eenvoudige lineaire methoden koos om de proteomische gegevens te analyseren en daarbij ook vergelijkbare classificatie resultaten rapporteerden.

Het idee van onderzoek naar een alternatieve aanpak voor het gebruik van een enkele methode om de diagnostische regel te kalibreren - in bijzonder door een verzameling van verschillende classificatie procedures te gebruiken en een gecombineerde classifier te bouwen - werd in eerste instantie aanbevolen door Hand (2008) als een effectieve manier om voorspellingen te verbeteren. Inderdaad, resultaten van onze data-analyse en een simulatie-onderzoek gebaseerd op hergebruik van de proteomic massaspectra-data tonen aan dat in veel situaties winst te behalen is in classificatie prestatie via een combinatiegebaseerde benadering. We beperkten de samenstellende methodes die werden gebruikt bij de bouw van de gecombineerde classifier tot lineaire classificatiemethoden omdat a) deze gunstige resultaten opleverden in de proteomische competitie oefening, en b) deze methoden worden over het algemeen als betrouwbaar en stabiel beschouwd voor hoogdimensionale dataproblemen met relatief kleine steekproefgrootten, zoals in het geval van omics onderzoek.

Hoofdstuk 3 gaat in op het probleem van het integreren van verschillende omics data om een gecombineerde diagnostische regel af te leiden die beter presteert dan een voorspellingsmodel gebaseerd op een enkele omic data bron. Terwijl de combinatie methoden

gepresenteerd in hoofdstuk 2 zijn gebaseerd op het combineren van voorspelde schattingen verkregen uit verschillende methoden, maar op basis van een gemeenschappelijke dataset, zijn de combinatiemethoden van hoofdstuk 3 in de context van een combinatie van voorspellingen op basis van verschillende datasets.

We beschouwen twee verschillende soorten procedures voor de combinatie van verschillende single-omics data voor predictie, beide op basis van geregulariseerde regressie. Het eerste type is een parallelle benadering, waarbij een nieuwe vector van voorspellingen wordt verkregen als een gewogen som van single-source voorspellingen van de uitkomst. De gewichten kunnen van tevoren worden vastgesteld, wat leidt tot een gewogen combinatie met vaste gewichten (middeling van de single-source voorspellingen) of geschat, leidend tot een model-gebaseerde combinatie aanpak. Beide benaderingen werden geintroduceerd in hoofdstuk 2 in de context van het combineren van schattingen van verschillende classifiers gebaseerd op dezelfde (single-omics) proteomische dataset. In hoofdtuk 3 worden deze methoden toegepast op de aparte voorspellingen gebaseerd op verschillende (omic) datasets. Zoals in hoofdstuk 2, worden de individuele voorspellingen in de model-gebaseerde combinatie behandeld alsof het nieuwe covariaten zijn in het uiteindelijke combinerende model. Daarom is bijkomende kruis-validatie nodig om de combinatieregel te kalibreren en onbevooroordeelde gecombineerde voorspellingen te verkrijgen. Het tweede type combinatiemethoden dat we beschouwen, is een sequentiele benadering waarin we a priori een van de omic bronnen als primair kiezen. In deze tweede aanpak worden (gekruisvalideerde) predicties verkregen uit kalibratie op de primaire omics-bron ingevoegd als “offsets” in de kalibratie van een predictie regel op de tweede omics-bron, waardoor de combinatie sequentieel tot stand komt.

We evalueren de combinatiemethoden met twee omics toepassingen. In de eerste, beschouwen we borstkanker data gegenereerd door verschillende fractioneringen met proteomic massa spectrometrie. In de tweede beschouwen we transcriptomics en metabolomics als aparte omics-bronnen voor predictie van obesitas. Resultaten van beide toepassingen laten zien dat voorspellingen gebaseerd op combinaties beter zijn dan single-omics voorspellingen. Bovenal zijn combinatie-methoden beter en daardoor dus ook noodzakelijk, in vergelijking met naive kalibratie op een gezamenlijk databestand waarin de omicsbronnen zonder onderscheid aan elkaar geplakt worden.

Hoofdstukken 4 tot 6 behandelen methoden voor de andere statistische uitdagingen bij het analyseren van hoge-resolutie MALDI-FTICR mass spectrometry data. De analyses in deze hoofdstukken zijn gebaseerd op pancreas kanker case-control data.

Hoofdstuk 4 beschouwt het probleem van het samenvatten en analyseren van MALDI-FTICR massa-spectrometrie data. Ten behoeve hiervan wordt een snelle en simpele procedure ontwikkeld voor het voorbewerken van de individuele spectra, voorafgaand aan de uiteindelijke analyse. Het doel is het identificeren van de isotopische clusters in de individuele spectra en het kwantificeren van hun isotopen pieken door het gebruiken van bestaande kennis over de isotopen clusterinformatie. De procedure is geheel gebaseerd op de bekende statistische eigenschappen van MALDI-FTICR MS data. Daarom is het in het bijzonder relevant in de context van grote studies met spectra verkregen over grote cohor-

ten patienten, aangezien het snel toegepast kan worden. Door middel van deze procedure wordt de volledige expressie in de individuele spectra gereduceerd tot samenvattingen voor de clusters van isotopen. Voor elke cluster kunnen dan samenvattende statistische maten worden gebruikt voor de expressie van die clusters. Vervolgens onderzoeken we verscheidene manieren voor het samenvatten van de expressie binnen de afgeleide clusters met als doel het evalueren van de verschillende keuzes met betrekking tot hun voorspellend vermogen. Om de samenvattende maten af te leiden, wordt informatie over - ofwel - de globale intensiteit - ofwel - de verdeling van de intensiteit binnen elk isotopen cluster gebruikt. Onze resultaten laten zien dat zowel intensiteit als verdeling informatie dragen over de klasse uitkomst en kunnen worden gebruikt voor voorspellingsdoeleinden.

Hoofdstuk 5 gaat dieper in op de eerder gerapporteerde resultaten om de voorspellende kracht van isotoopclusters in de pancreaskanker data aan te tonen. Het doel van dit hoofdstuk is om: a) een verzameling isotopen-clusters te identificeren die geassocieerd zijn met de uitkomst van de ziekte en b) om informatie over zowel intensiteit als verdeling optimaal te integreren. We stellen een benadering voor waarbij gebruik wordt gemaakt van voorkennis over de relatieve voorspellende kracht van elke bron, evenals de gepaarde structuur van de proteomische gegevens - zoals samengevat op basis van onze statistische maten binnen de isotopen clusters. Om het probleem van het selecteren van isotopen te onderzoeken en tegelijkertijd de toegevoegde waarde van de verdeling te beoordelen, gebruiken we een Bayesiaanse modelformulering waarmee we meerdere selectielagen kunnen introduceren. We presenteren een post-hoc analyse die onderzoekers in staat stelt zich te concentreren op een beperkte set potentieel interessante isotopenclusters voor verder onderzoek en die inzicht geeft in de relatieve voorspellende capaciteit van verdeling gecombineerd met intensiteit. Afsluitend presenteren we de resultaten van een simulatieonderzoek om te demonstreren hoe de methode zich gedraagt in een gecontroleerde situatie. De resultaten laten zien dat de Bayesiaanse benadering de belangrijkste clusters met succes kan identificeren en de effecten van zowel intensiteit als verdeling nauwkeurig kan schatten.

Hoofdstuk 6 beschrijft nieuwe methoden om diagnostische regels op te stellen voor hoge resolutie massaspectrometrie proteomics data die onderhevig zijn aan de detectielimiet (LOD). We passen gecensureerde regressiemethodologie toe om de gemiddelde individuele expressie in isotopen clusters te schatten, ervan uitgaande dat eventuele ontbrekende intensiteiten in onze dataset te wijten zijn aan links-censurerende mechanismen bij eiwitten die slechts in lage hoeveelheden aanwezig zijn. Ons doel is om informatie te verkrijgen over de gemiddelde proteomische expressie die het mogelijk maakt om voorspelregels te construeren die net zo goed zijn als wanneer we de volledige informatie tot onze beschikking hadden gehad. De voorgestelde methoden worden geevalueerd met betrekking tot hun voorspellende prestaties, door de onvolledige metingen te vervangen door de schattingen van individuele expressie, rekening houdend met de detectielimiet, en deze als nieuwe invoervariabelen te gebruiken in een voorspellingsmodel. We combineren gecensureerde regressie met het lenen van informatie door de toevoeging van individueel-specifieke random effecten om rekening te houden met mogelijk gebrek aan

informatie en meetonzekerheid. We demonstreren verschillende varianten van het gebruik van de random effecten-formulering voor voorspellingsdoeleinden. Resultaten van zowel interne als externe validatie laten zien dat het gebruik van deze schattingen als nieuwe voorspellers resulteert in een vergelijkbare voorspellende nauwkeurigheid als die werd bereikt met behulp van de volledige intensiteitsinformatie.