Algorithmic tools for data-oriented law enforcement

Cocx, T.K.

Nederlandse Samenvatting

Op de achtergrond van de data-explosie van de late jaren 1990 is er een onderzoeksgebied geëvolueerd uit de statistiek en informatica. Het hoofddoel van deze vorm van computergestuurde data analyse, die bekend staat als *data mining* (het graven in gegevens) of *Knowledge Discovery in Databases* (KDD) ("kennis ontdekking" in databases), is om kennis te extraheren uit een, vaak grote, collectie van "ruwe" data, waarbij elementen uit de statistiek, database technologie, kunstmatige intelligentie, visualisatie en uit het machine-leren worden gecombineerd. Een belangrijk aspect van het vakgebied is het omgaan met gegevens die niet specifiek ontworpen werden om er computergestuurde analyses op uit te voeren. Bij het bedrijven van data mining draait het meestal om het vinden van onverwachte patronen en andere verrassende resultaten waar niet direct of concreet op gezocht werd.

Het toenemen van de mogelijkheden in de informatietechnologie van het laatste decennium heeft geleid tot een grote toename van de hoeveelheid opgeslagen gegevens, zowel als een zij-product van bedrijfs- en overheidsadministratie, als resultaat van wetenschappelijke analyses. Hoewel de meeste van deze gegevens een inherent nut met zich meebrengen, zoals klant management, het archiveren van belastingteruggaves of het uitvoeren van DNA analyses, heeft data mining software als doel om kennis te aggregeren uit deze gegevens, door het automatisch opsporen van (onderliggende) patronen, gedragsklassen of communicatienetwerken. Respectievelijk kan zo waardevolle kennis worden ingewonnen over klant gedrag, belastingontduiking of over stukjes DNA die garant staan voor, biologische afwijkingen. Meestal kunnen de algoritmes, opgesteld voor dit soort taken betrekkelijk eenvoudig overgebracht worden naar andere expertise-domeinen.

Een van deze potentiële domeinen is dat van de wetshandhaving. Als een onderdeel van nationale of regionale overheden dat gebonden is aan strenge regelgeving, is de overgang van een papieren administratie naar een digitale informatie infrastructuur langzaam gegaan, maar de laatste jaren, zeker na de aanslagen van 11 september 2001, hebben politieorganisaties meer geïnvesteerd in specifieke, bruikbare en uniforme informatiesystemen. Zoals in alle andere gebieden, heeft dit proces geleid tot zeer veel gegevens, die waarschijnlijk zeer bruikbaar zijn voor data mining doeleinden. Dit proefschrift beschrijft een aantal projecten in deze richting en rapporteert de resultaten die bereikt werden door toepassing van de ontwikkelde algoritmes op daadwerkelijke politie gegevens. Hiermee wordt een basis gelegd voor een toekomst waarbij speciaal ontwikkelde algoritmische assistentie een waardevolle rol kan spelen bij het bestrijden van de misdaad.

Het gerapporteerde onderzoek in dit proefschrift is onderverdeeld in twee delen, die direct gerelateerd zijn aan twee vormen van de operationele wetshandhaving zoals die in Nederland bestaan: de strategische en de tactische wetshandhaving. Bij het eerste moet gedacht worden aan het uitzetten van (korps-)beleid met betrekking tot locaties, personeelsbezetting, et cetera, op basis van bestaande gegevens over criminaliteit. De tactisch geörienteerde activiteiten zijn vaak direct gerelateerd aan opsporing en zaak-analyses.

In Hoofdstuk 2, als eerste hoofdstuk van het strategisch deel, wordt een eerste analyse uitgevoerd van een grote database van strafbladen (zie Appendix B), waarbij gepoogd wordt verschillende misdaden aan elkaar of aan demografische gegevens te relateren, op basis van het feit dat ze vaak samen "voorkomen" in een strafblad. Om dit te bereiken wordt het bekende algoritme A-PRIORI aangepast om in dit specifieke bestand te zoeken naar dergelijke verbanden, terwijl het rekening houdt met een variëteit aan problemen die inherent zijn aan criminaliteitsgegevens. Dit bestand, dat in geanonimiseerde versie beschikbaar werd gesteld, is eveneens gebruikt in andere analyses.

Omdat dit bestand een behoorlijke omvang "ruwe" data heeft, zijn standaard methoden om deze data te visualiseren vaak niet erg geschikt. In Hoofdstuk 3 wordt daarom een methode beschreven die het weergeven van verbanden tussen criminelen uit dit bestand optimaliseert door de domein kennis van de analist te betreken bij het "clusteren". Doordat deze expert direct de fysieke weergave van de gegevens kan manipuleren kan met een relatief lage "berekeningscomplexiteit" een zeer hoge kwaliteit visualisatie behaald worden.

Een belangrijk concept dat geëvalueerd kan worden met behulp van het strafbladenbestand is dat van de *criminele carrière*, dat beschouwd kan worden als een temporeel geordende serie misdaden die een individu tijdens zijn leven begaan heeft. Een ad-hoc methode wordt gesuggereerd in Hoofdstuk 4, waarbij vier belangrijke factoren van een carrière gebruikt worden om afstanden tussen verschillende carrières te berekenen. Deze afstanden kunnen vervolgens gevisualiseerd worden in een twee-dimensionale clustering. Hoofdstuk 5 stelt een aantal verbeteringen op deze methode voor die allemaal al functioneel zijn gebleken op een ander gebied. Eveneens worden daar methoden beschreven om uit deze gegevens nieuwe carrières te kunnen voorspellen, waarop verder ingegaan wordt in Deel II.

Nu een systeem ontworpen is voor het clusteren en classificeren van criminele carrìeres, is het mogelijk te zoeken naar vaak voorkomende subcarrières. Een nog belangrijkere onderneming is het zoeken naar subcarrières die vaak voorkomen in een specifieke klasse maar juist niet in alle anderen, zodat zij de rol op zich kunnen nemen van defini erende subcarrière, die gebruikt kan worden om bepaalde klasses te identificeren. Deze mogelijkheden worden behandeld in Hoofdstuk 6, waar een bestaande methode voor winkelmandjes-analyse wordt aangepast om tegemoet te komen aan de specifieke eisen voor de zoektocht naar subcarrières.

In het tweede deel wordt één van deze mogelijkheden beschreven in Hoofdstuk 7, via een methode om de kracht van een visualisatie aan te wenden om via simpele mathematische berekeningen een betrouwbare voorspelling te doen. Deze methode wordt uitgewerkt en speciaal toegespitst op criminaliteitsgegevens in Hoofdstuk 8, waar de verschillende variabelen van deze methodiek worden getoetst op daadwerkelijke data.

Via deze methode kunnen criminele carrières onder bepaalde omstandigheden met grote nauwkeurigheid voorspeld worden.

In Hoofdstuk 9 staat een onderzoek beschreven dat zich buigt over de vraag in hoeverre bestanden van in beslag genomen computers een indicatie kunnen zijn voor welke misdrijflocaties gerelateerd zijn aan dezelfde criminele organisaties. Voor dit doel werd een speciale afstandsmaat gedefinieerd die bepaald hoe groot de kans is dat twee computers bij dezelfde organisatie behoren. Hierbij werd gebruik gemaakt van tekstherkenningssoftware die tekst extraheerde van computers uit drugslaboratoria.

In Hoofdstuk 10 staat beschreven hoe *online predators*, "kinderlokkers op internet", automatisch herkend zouden kunnen worden op sociale netwerkomgevingen, zoals het Nederlandse Hyves. Er wordt beschreven hoe een genetisch algoritme automatisch groepen selecteert waartussen een significant verschil op te merken is tussen deze predators en andere gebruikers in het aantal minderjarige vrienden op hun profiel. Het blijkt dat deze variabele in sommige gevallen een zeer sterke indicator kan zijn voor risico classificatie van bepaalde gebruikersgroepen.

Dit proefschrift eindigt met Appendix A waarin een aantal overwegingen worden gegeven met betrekking tot statistiek, recht en privacy die een cruciale rol spelen voor iedereen die (een deel van) ons werk gebruikt of van plan is te gebruiken in de dagelijkse omgeving van het politiewerk. Het bespreekt de toepasbaarheid, statistische relevantie en inzichtelijkheid van en wat voorbehoudens met betrekking tot onze methodieken in het algemeen en voor politiegebruik in het bijzonder, waarbij vooral gefocust wordt op de gevoeligere toepassingen, besproken in Deel II. Om er zeker van te zijn dat onze methodes op de correcte manier bekeken en onze tools op een nauwkeurige en gepaste wijze gebruikt worden, is een behandeling van hun mogelijkheden en beperkingen voor maatschappelijk gebruik zowel belangrijk als vanzelfsprekend.