

1 Inleiding

Dit project heeft tot doel om efficiënte algoritmen te ontwikkelen voor relationele inductieve kennisbanken. Hierbij wordt speciale aandacht gegeven aan een aantal applicaties. In dit verslag wordt eerst de vooruitgang in een aantal toepassingsgebieden besproken, en daarna de ontwikkelde algoritmen en theorie.

2 Applicaties

2.1 Intensive zorgen

De vorige jaren werd reeds samengewerkt met de eenheid intensieve zorgen van het UZ-Leuven. Een aantal eerste resultaten werden o.a. gepubliceerd in [3]. Sinds midden 2006 is het online monitoring systeem van de afdeling voldoende operationeel en uitgetest om betrouwbare hoge-resolutie data te leveren. Deze data bevat o.a. gegevens over de waarden van meerdere parameters van de patienten die elke minuut worden gemeten. In eerste instantie is er gewerkt aan het voorspellen van de evolutie van deze parameters met Gaussiaanse processen [2]. Voor korte-termijnvoorspellingen blijkt deze aanpak succesvol, maar voor langere-termijnvoorspellingen neemt de nauwkeurigheid sterk af. Meer recent wordt daarom gewerkt aan twee ideeën die dit probleem mogelijk kunnen oplossen. De eerste is het beter parameteriseren van de dynamische eigenschappen van de data (in samenwerking met de Biores groep van de faculteit Bioingenieurswetenschappen). De tweede is een benadering met Dynamische Bayesiaanse netwerken waarbij we het probleem op een hoger niveau aanpakken.

2.2 Coherente lasercontrole

Het artikel [1] werd gepubliceerd en bevat resultaten van de samenwerking met de computational biology groep van Aberystwyth, UK en het departement scheikunde van Leeds tijdens mijn verblijf in Aberystwyth.

2.3 HIV-onderzoek

HIV is een snel muterend virus. Het ontwikkelt gemakkelijk mutaties die resistentie tegen bestaande therapiën opleveren. Om meer inzicht te krijgen in de manier waarop HIV resistent wordt is het nuttig te begrijpen wat de mutatiepaden (opvolging van mutaties die tot resistentie leiden) zijn. Nu beschikt men op dit ogenblik niet over grote hoeveelheden longitudinale data (meerdere sequenties van het virus in eenzelfde patient op verschillende tijdstippen). Om mutatiepaden te ontdekken moet daarom zoveel mogelijk gebruik gemaakt worden van transectionele data (een of twee sequenties per patient). Echter, directe toepassing van data mining methoden levert geen goed resultaat op omdat HIV

veel polymorphismen heeft (verschillende mogelijke sequenties die in het wild-type voorkomen).

We willen een algoritme ontwerpen dat gegeven een aantal sequenties van virussen van verschillende patienten, enerzijds een phylogenetische boom bouwt die zo goed mogelijk alle polymorphismen verklaart en anderzijds een functie leert die voor een gegeven sequentie zegt hoe waarschijnlijk elke mogelijke volgende mutatie is. Die functie geeft dan impliciet de mutatiepaden aan. [6] is een eerste aanzet in die richting.

3 Data mining algorithmen

3.1 Bayesiaanse logische programma's

Er werd gewerkt aan het ontwikkelen van methoden voor Bayesiaanse netwerken op relationele data. Vorig jaar werd reeds een vergelijkende studie gedaan van verschillende methoden voor het leren van relationele probabiliteitsbomen. Deze bomen laten toe om functies voor te stellen van relationele data naar probabiliteiten. Om dergelijke bomen optimaal te leren bleek dat men beter andere heuristieken gebruikt dan voor bv. regressiebomen.

Dit jaar werd deze bijdrage verder uitgewerkt en werd een tijdschriftartikel voorbereid. De relationele probabiliteitsbomen werden ondertussen in meerdere applicaties (o.a. HIV) toegepast.

[4] bouwt in zekere zin hierop verder en beschrijft een veralgemening van ordering search voor relationele data. Deze methode laat toe om de structuur van Logische Bayesiaanse Netwerken (LBN) en gelijkaardige modellen te leren in de vorm van ordeningsfuncties die aangeven in welke volgorde de knopen van het netwerk moeten beschouwd worden. Als lokale modellen gebruiken we relationele probabiliteitsbomen.

3.2 Incrementeel leren en reviseren van relationele beslissingsbomen

Niet alle kennis blijft onveranderd geldig. Omstandigheden kunnen veranderen en soms is het nodig geleerde patronen aan te passen. Recentelijk is er in het domein van reinforcement learning veel aandacht voor "transfer learning". Hiermee bedoelt men dat men eerst een taak leert oplossen en daarna probeert de opgedane kennis te hergebruiken voor het oplossen van een andere taak. Niet alle kennis is nog geldig, maar een aantal elementen blijven wel toepasselijk.

Een vergelijkbaar probleem doet zich voor bij online leren, waar men niet alle informatie meteen krijgt, maar de informatie slechts geleidelijk aan binnenkomt. In onze intensieve zorgen toepassing bijvoorbeeld heeft men bij opname van een patient weinig informatie, maar moet men toch al beslissingen nemen. Naarmate de patient langer op de afdeling blijft zal de kennis over de patient toenemen.

Mogelijks moet men daarom in de loop van het verblijf bepaalde hypothesen bijstellen. Aan de andere kant bestaat de hoop dat een nieuwe patient op een redelijk gelijkaardige manier zal reageren als de vorige, zodat men vroeger opgedane kennis (gedeeltelijk) kan hergebruiken.

In [5] beschrijven we een benadering om incrementeel beslissingsbomen te leren en, als fouten worden ontdekt, te reviseren. Het artikel beschrijft een aantal experimenten in het domein van reinforcement learning.

3.3 Het minen van graaf patronen

Er werd verder gewerkt rond het minen van frequente relationele patronen. Een probleem dat in de pattern mining literatuur veel aandacht krijgt is het efficiënt genereren van kandidaat-patronen. Zeker voor complexe patronen, zoals grafen, is het niet triviaal om alle patronen exact een keer te genereren (en bv. geen twee grafen te genereren die isomorf zijn). Als men elke nieuwe kandidaat met alle eerder gegenereerde kandidaten moet vergelijken bestaat het risico dat men een groot aantal isomorfie-testen moet doen, en er is geen algoritme bekend dat in het algemeen isomorfie tussen grafen kan beslissen in polynomiale tijd. Eenmaal er een methode is om elk patroon een keer te genereren kan men voor elk gegenereerd patroon de frequentie berekenen en op gepaste wijze in de zoekruimte knippen om de verzameling van alle frequente of interessante patronen te berekenen.

[7] beschrijft een algoritme dat voor een brede waaier van graaf klassen alle grafen die tot die klasse behoren en waarvoor een willekeurig anti-monotoon predicaat geldig is precies een keer genereert, waarbij het algoritme slechts een polynomiale hoeveelheid tijd besteedt per uitgevoerde grafe. Interessant is dat dit niet betekent dat we isomorfie tussen grafen kunnen beslissen in polynomiale tijd. Er kunnen immers een exponentieel aantal grafen zijn in functie van de maximale grootte van de grafen. Eigenlijk hergebruikt het algoritme berekeningen op zo'n manier dat de totale rekentijd aan de claim voldoet.

4 Andere activiteiten

Er werd tijd besteed aan het begeleiden van en samenwerken met een aantal doctoraatsstudenten (zie bijgevoegde lijst).

Ik ben betrokken bij de organisatie van de ILP2007 conferentie als verantwoordelijke voor de proceedings.

References

- [1] N. Form, R. Burbidge, J. Ramon, and J. Whitaker. Parameterisation of an acousto-optic programmable dispersive filter for closed-loop learning experiments. *Journal of Modern Optics*, 99999(1):1–13, January 2007.
- [2] F. Guiza, J. Ramon, and H. Blockeel. Gaussian processes for prediction in intensive care. In Neil D. Lawrence, Anton Schwaighofer, and Joaquin Quinero, editors, *Proceedings of the Gaussian Processes in Practice Workshop*, Bletchley Park, U.K., June 2006.
- [3] J. Ramon, D. Fierens, F. Guiza, G. Meyfroidt, H. Blockeel, M. Bruynooghe, and G. Van Den Berghe. Mining data from intensive care patients. *Advanced Engineering Informatics*, 21(3):243–256, 2007.
- [4] Jan Ramon, Tom Croonenborghs, Daan Fierens, Hendrik Blockeel, and Maurice Bruynooghe. Generalized ordering-search for learning directed probabilistic logical models. *Machine Learning*, 2007. to appear.
- [5] Jan Ramon, Kurt Driessens, and Tom Croonenborghs. Transfer learning in reinforcement learning problems through partial policy recycling. In *Proceedings of the 18th European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence. Springer-Verlag, 2007.
- [6] Jan Ramon, Snezhana Dubrovskaya, and Hendrik Blockeel. Learning resistance mutation pathways of HIV. In *Proceedings of The Sixteenth Annual Machine Learning Conference of Belgium and the Netherlands, Amsterdam, The Netherlands*, 2007. URL: <http://www.cs.kuleuven.ac.be/cgi-bin-dtai/publ.info.pl?id=42687>.
- [7] Jan Ramon and Siegfried Nijssen. General graph refinement with polynomial delay. In *Proceedings of the Workshop on Machine Learning and Graphs (MLG'07)*, 2007.