Universiteit
Leiden
The Netherlands

Robust rules for prediction and description

Manuel Proenca, H.

Citation
Manuel Proenca, H. (2021, October 26). *Robust rules for prediction and description*. *SIKS Dissertation Series*. Retrieved from https://hdl.handle.net/1887/3220882

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3220882

Note: To cite this publication please use the final published version (if applicable).

Samenvatting

Regels bieden een eenvoudige vorm voor het opslaan en delen van informatie over de wereld. Als mensen gebruiken we elke dag regels. Een arts die iemand met griep diagnosticeert, gebruikt bijvoorbeeld de regel: ”als een persoon koorts of keelpijn heeft, dan heeft hij of zij griep”. Hoewel een individuele regel alleen eenvoudige gebeurtenissen kan beschrijven, kunnen verschillende, geaggregeerde regels gezamenlijk complexere scenario’s beschrijven, zoals de volledige set diagnostische regels die (impliciet) door een arts worden gebruikt.

Gezien het veelvuldige gebruik van regels in verschillende domeinen, is het geen verrassing dat op regels gebaseerde modellen tot een van de eerstegebruikte technieken behoorden om computers uit te rusten met besluitvormingsmechanismen. Oorspronkelijk voerden mensen regels rechtstreeks in een computersysteem in; met de beschikbaarheid van grote hoeveelheden data verschoof de interesse naar het leren van regels uit data. De elektronische dossiers van een arts, waarin is vastgelegd of zijn patienten wel of geen griep hebben op basis van hun symptomen, kunnen bijvoorbeeld gebruikt worden om het besluitvormingsproces van de desbetreffende arts te achterhalen.

Het gebruik van regels omvat vele gebieden in de informatica; in dit proefschrift richten we ons op de op regels gebaseerde modellen voor machine learning en datamining. Machine learning is erop gericht om op basis van data het model te leren dat toekomstige (niet eerder waargenomen) gebeurtenissen het beste voorspelt. Datamining heeft als doel om interessante patronen te vinden in de beschikbare data. In het bijzonder richten we ons op de volgende onderzoeksvraag: “Hoe leren we robuuste en interpreteerbare op regels gebaseerde modellen van data voor machine learning en datamining, en definiëren we het optimale model?”

Om deze vraag te beantwoorden gebruiken we het Minimum Description Length (MDL) principe, waarmee we de optimaliteit van op regels gebaseerde modellen voor een bepaalde dataset kunnen bepalen. Informeel is het beste model voor een specifieke dataset het eenvoudigste model dat de data goed beschrijft. De specifieke modelklasse waarop we ons concentreren zijn zogenaamde regellijsten, d.w.z. geordende verzamelingen regels die opeenvolgend worden geïnterpreteerd. Helaas is het vinden van een optimaal model in de meeste gevallen computationeel onhaalbaar. Daarom stellen we heuristische algoritmen voor die goede modellen vinden en daarbij enkele garanties geven. We testen onze algoritmen empirisch om onze aanpak te valideren, en laten zien dat ze in de meeste gevallen beter of vergelijkbaar presteren in vergelijking met de state of the art.