



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).

## Resumo

Regras são uma forma simples de armazenar e partilhar informação sobre o mundo. Como seres humanos, usamos regras todos os dias, tal como o médico que diagnostica alguém com gripe, representada da forma “se uma pessoa tem febre ou dor de garganta (entre outras aflições), então ela tem gripe.”. Embora uma regra só possa descrever eventos simples, o agregamento de várias regras pode descrever cenários mais complexos, como o conjunto de todas as regras de diagnóstico usadas por um médico.

Dado seu uso abundante, não é surpresa que os modelos baseados em regras tenham sido umas das primeiras técnicas usadas para equipar computadores com poder de decisão. Inicialmente, as pessoas inseriam as regras diretamente nos sistemas dos computadores; no entanto, com a disponibilidade de grandes quantidades de dados, o interesse voltou-se para aprender regras a partir dos dados. Por exemplo, os relatórios de diagnósticos de um médico sobre se os seus pacientes têm ou não gripe com base nos seus sintomas podem ser usados para aprender o processo de tomada de decisão desse mesmo médico. O uso de regras abrange muitos campos da ciência da computação e, nesta dissertação, focamo-nos em modelos baseados em regras para as áreas de aprendizagem de máquina e prospecção de dados. A aprendizagem de máquina dedica-se a aprender o modelo que melhor prevê eventos futuros (não vistos até então) a partir dos dados. A prospecção de dados visa encontrar padrões interessantes nos dados disponíveis. Mais concretamente, estamos preocupados com a seguinte questão científica: “ Como aprender a partir dos dados modelos baseados em regras que sejam robustos e interpretáveis nas áreas de aprendizagem de máquina e prospecção de dados, e definir a sua proximidade ao ótimo? ”

Para responder a essa questão, empregamos o princípio da Descrição de Comprimento

Mínimo (MDL em inglês), que nos permite definir quão ótimos são os modelos baseados em regras para um conjunto de dados específico. Informalmente, o melhor modelo para um conjunto de dados específico é o modelo mais simples que melhor descreve os dados. A classe de modelo específica em que nos focamos é chamado lista de regras, ou seja, um conjunto ordenado de regras que têm de ser interpretadas sequencialmente. No entanto, encontrar um modelo ótimo é computacionalmente impossível na maioria dos casos. Por isso, propomos heurísticas que descobrem bons modelos mantendo algumas das garantias originais. Os nossos algoritmos foram empiricamente testados, validando desta forma a nossa abordagem e mostrando, na maioria dos casos, que eles atingem um desempenho melhor ou semelhante ao estado da arte em aprendizagem de máquina e prospeção de dados.