Scalable, High-Performance Data Mining with Parallel Processing

Alex Alves Freitas

CEFET-PR, Dep. de Informatica (DAINF)
Av. Sete de Setembro, 3165
Curitiba - PR 80230-901
BRAZIL
alex@dainf.cefetpr.br
http://www.dainf.cefetpr.br/ alex

Abstract

Parallel processing seems to be the great hope to speed up and scale up data mining algorithms, in order to cope with the huge size of real-world databases and data warehouses. However, most projects on parallel data mining have focused on the parallelization of a single kind of algorithm or knowledge discovery paradigm. This tutorial will present a considerably broader view of the area of parallel data mining. In particular, it will discuss the parallelization of algorithms of four different knowledge discovery paradigms, namely rule induction, instance-based learning (or nearest neighbours), genetic algorithms and neural networks. In addition, this tutorial will address both the use of "general- purpose" parallel machines and the use of commercially-available parallel database servers. Different parallelization strategies will be discussed and compared, for each of the four above- mentioned knowledge discovery paradigms.