

RECENT ADVANCES IN DATA STREAM MINING  
JESUS S. AGUILAR-RUIZ  
SCHOOL OF ENGINEERING  
PABLO DE OLAVIDE UNIVERSITY

In recent years, designing scaling-up and scalable algorithms has consolidated as an important challenge within data mining research community. Memory and time limitations compel make such system to give an approximate answer from few scans (ideally only one) assuring that both result and performance are not adversely affected by the order of the examples. In addition, when the distribution is not stationary (records are collected over months), algorithms based on data partitioning techniques (instance/feature sampling) are oversensitive to both underfitting and overfitting. Many scalable learning algorithms are based on decision trees, modelling the whole search space hierarchically as disjointed hypercubes. The large and complex trees given by these systems cast doubts on its capabilities as suitable knowledge representation due to the user need to explore paths of several dozen of levels to know interesting patterns. In addition, mining time-changing data streams may involve to check an *out-of-date* sub-tree, increasing the computational cost. A common approach in these systems consists in repeatedly applying the learner to a sliding window of  $w$  examples. The main goal in these systems is then to find the value for  $w$  that optimizes the performance as a function of the input data. So interactive and user-controlled data mining systems are becoming increasingly developed. Such approaches trade accuracy for simplicity providing more meaningful and understandable models.

This work will give an introduction to general concepts and an overview of the recent systems and challenges on data streams.