

The Data Stream Phenomenon

State-of-the-
Art in Data
Stream
Mining
(Part I)

João Gama
and Mohamed
Gaber

Conclusions
and Open
Issues

- Highly detailed, automatic, rapid data feeds.
 - Radar: meteorological observations.
 - Satellite: geodetics, radiation,.
 - Astronomical surveys: optical, radio,.
 - Internet: traffic logs, user queries, email, financial,
 - Sensor networks: many more *observation points* ...
- Most of these data will never be seen by a human!
- Need for near-real time analysis of data feeds.
- Monitoring, intrusion, anomalous activity Classification, Prediction, Complex correlations, Detect outliers, extreme events, fraud,

The Past of Machine Learning

In the last two decades, machine learning research and practice focus in batch learning using small datasets.

- The whole training data is available to the algorithm, that outputs a decision model after processing the data multiple times.
- This practice assumes that examples were generated at random accordingly to some stationary probability distribution.
- Most learners use a greedy, hill-climbing search in the space of models.
- Learning from small datasets: Emphasis in variance reduction.

What distinguishes current data sets from earlier ones is *automatic data feeds*. We do not just have people entering information into a computer. We have computers entering data into each other.

The Future of Machine Learning

Learning from small datasets: emphasis in variance reduction.
Whats about large datasets?

- Increasing data = Variance reduction. Stable statistics estimators
- Learning from large datasets may be more effective using algorithms that places greater emphasis on bias management
- Solutions to these problems require
 - New Sampling and Randomize Techniques,
 - New Approximate, Incremental Algorithms,
 - Management the cost of Model's update and the Gains in Performance.
 - Incorporation of Change Detection Algorithms inside the Learning Process.

Thanks for your attention!

More information:

- Sensors** J. Gama, R. Pederson; *Predictive Learning from Sensory Data*, Learning from Data Streams – Processing Techniques in Sensor Networks, Springer Verlag, 2007.
- Streams** Learning from Data Streams – Processing Techniques in Sensor Networks, Editores J. Gama and M. Gaber, Springer Verlag, 2007.
- Streams** S. Muthukrishnan, *Data Streams: Algorithms and Applications*, Now Publishers, 2003.
- VFDT** P. Domingos, G. Hulten; *Learning from Infinite Data in Finite Time*, Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002
- VFDT** J. Gama, R. Fernandes, R. Rocha, *Decision Trees for Mining Data Streams* Intelligent Data Analysis, Vol. 10, Number 1, IOS Press, 2006.
- ODAC** P. P. Rodrigues, J. Gama and J. P. Pedroso. *ODAC: Hierarchical Clustering of Time Series Data Streams*. In *Proceedings of the Sixth SIAM International Conference on Data Mining*, 2006.

Thanks for your attention!

State-of-the-Art in Data Stream Mining (Part I)

João Gama and Mohamed Gaber

Conclusions and Open Issues

