

Mining High-Speed Data Streams

Authors:

Pedro Domingos
Geoff Hulten
University of Washington

Presented By:

Vinod Kulkarni
Frederic Bihan
Lars Eriksson

Problem

- Huge Volume of data growing very fast
- Higher data rate arrivals than that can be mined
- Data bases should be scanned several times

Overview

- Introduction
- **Prior Work**
- Key Contributions
- Methodology
- Results
- Our Opinion of the paper

Prior Work

- Efficient but no guarantee that the model will be similar to the "Batch Mode" model.
- Highly sensitive to Example Ordering.
- Some produce similar models but are sometimes slower than the batch algorithm.

Key Contributions

- Constant time per example
- Nearly Identical result to the Batch Algorithm results
- Need to see examples only Once
- Available ready to use model after some computation.

Hoeffding Bounds

- Statistical result :

$$\epsilon = R \cdot \sqrt{(\ln(1/\delta)) / 2n}$$

ϵ : Hoeffding Bound

R : Range

δ : one minus the desired probability of choosing the correct attribute at any given node

n : number of examples

Methodology

- No need to scan the entire database to determine the best attribute to test at a given node
- But how many examples do we need ?
 - Hoeffding Bounds

- As the number of examples increases ϵ decrease since the numerator is constant.
- This ensures that with high probability the attribute chosen using n examples is the same that would be chosen using infinite examples.
- This leads to the creation of a new decision tree learner algorithm using this Hoeffding Bound property, the Hoeffding tree algorithm.

Hoeffding Tree Algorithm

- Algorithm on appendix A.
- New Features:
 - X_0 : NULL attribute that represents the fact that not splitting the node.
 - $G(X_a) - G(X_b) > \epsilon$: leads to a split

Results cont..

- Running the VFDT, VFDT boot and C4.5 on data sets, it appears that C4.5 is more accurate than both VFDT based systems under 100K examples.
- But we can also see that VFDT begins to be interesting when C4.5 stops, taking advantage of the large number of examples to reach a very high accuracy (like 88.8%).

Results

- These results have been obtained using VFDT which is the system based on the Hoeffding tree algorithm.
- It has also been compared to C4.5 release 8.

Results cont..

- As we look at the accuracy as a function of the noise, it appears the difference between C4.5 and the VFDT based systems increases as the noise level increase.
- This result tends to show that Hoeffding bound is an effective pruning method.

Opinion

- We have graded the paper in two aspects, the paper as a whole and the algorithm both on a scale 1 - 10.
- Algorithm : 8
- Paper : 7