Advanced Data Science with Spark Streaming

Albert Bifet
Huawei Noah’s Ark Lab: want to build

Intelligent Mobile Devices

Intelligent Telecommunication Networks

Data Mining & Artificial Intelligence

Intelligent Enterprise
Stream Data Mining

Data Streams

- Sequence is potentially infinite
- Large amount of data: sublinear space
- Arrival at high speed: sublinear time per example
- Once an element from a data stream is processed it is either discarded or archived
- Data is evolving, it is not static

Approximation algorithms

- Small error rate with high probability
Open Source Stream Data Mining Projects

- Non distributed
  - VFML (C)
  - Vowpal Wabbit (C++)
  - Sofia-ML (C++)
  - MOA (Java)

- Distributed
  - Streams (Java)
  - Mahout (Java)
  - Jubatus (C++)
  - Apache SAMOA (Java)
streamDM: Data Mining for Spark Streaming
IBM Announces Major Commitment to Advance Apache®Spark™, Calling it Potentially the Most Significant Open Source Project of the Next Decade

IBM Joins Spark Community, Plans to Educate More Than 1 Million Data Scientists
streamDM is designed specifically to be used inside Spark Streaming.

Diagram:
- Input data stream
- Spark Streaming
- Batches of input data
- Spark Engine
- Batches of processed data

Spark logo
Overview

- streamDM is near real-time (latency 1 sec).
- For real-time computation (latency < 1 msec),
  - use Storm or Flink with Apache SAMOA
Advantages

- **Advanced machine learning methods included** such as streaming decision trees, streaming Random Forests, CluStream and StreamKM++.

- **Ease of use.** Experiments can be performed from the command-line, as in WEKA or MOA.

- **High extensibility**

- **No dependency on third-party libraries**, v.s. MLlib uses Breeze, which depends on netlib-java, and jblas which depends on native Fortran routines.
streamDM Getting Started

- Download streamDM
  
git clone https://github.com/huawei-noah/streamDM.git

- Build streamDM
  
sbt package

- streamDM execute tasks
  
  ./spark.sh "EvaluatePrequential
  -l (SGDLearner -l 0.01 -o LogisticLoss
  -r ZeroRegularizer)"
streamDM methods

Pre-release 30/06/15
● SGD Learner and Perceptron
● Naive Bayes
● CluStream
● Hoeffding Decision Trees
● Bagging
● Stream KM++

Release 31/12/15
● Random Forests
● Frequent Itemset Miner: IncMine
Hoeffding Tree

- With high probability, constructs an identical model that a traditional (greedy) method would learn
- With theoretical guarantees on the error rate
CluStream / StreamKM++

• Clustream is based on the concept of *microclusters*.  
  • Microclusters are summary data structures

• CluStream has two phases:
  • In the *online* phase, a set of microclusters are updated and maintained incrementally
  • In the *offline* phase, it applies a weighted k-means algorithm on the microclusters, to obtain the final clusters from the stream.

• *StreamKM++* computes a small weighted sample of the data stream, called the *coreset* of the data stream.
STREAMDM C++

Stream Machine Learning in C++
streamDM C++ compared with VFML

● It is much faster and uses less memory.
  • Faster than VFML in C and MOA in Java.

● Number of powerful methods
  • The adaptive decision tree is more accurate and does not need an expert user to choose the optimal parameters.
  • It contains powerful ensemble methods.

● Easy to use
  • Evaluation and learning are separated, not linked together.
  • It contains several methods for learning numeric attributes.

● It is easy to extend and add new methods.
Open Source Software Projects 2015
Huawei Noah’s Ark Lab

Founders of:
• streamDM for Spark Streaming
• streamDM C++

Supporting:
• Apache SAMOA
• MOA
Contributors streamDM

Silviu Maniu (HUAWEI Noah's Ark Lab)
Albert Bifet (HUAWEI Noah's Ark Lab)
Jianfeng Qian (HUAWEI Noah's Ark Lab)
Guangjian Tian (HUAWEI Noah's Ark Lab)
Cheng He (HUAWEI Noah's Ark Lab)
Wei Fan (HUAWEI Noah's Ark Lab)
streamDM: Data Mining for Spark Streaming
Thanks!

http://huawei-noah/github.io/streamDM