HUAWEI Advanced Data Science with Spark Streaming

Albert Bifet (@abifet)
Huawei Noah’s Ark Lab Focus

Intelligent Mobile Devices

Data Mining & Artificial Intelligence

Intelligent Telecommunication Networks

Intelligent Enterprise
Big Data Stream Team Challenges

Goal:
- Focus on Big Data Stream Mining, Graph Mining systems and algorithms related research areas;
- Combine domain knowledge from MBB/FBB_IDC networks, address key challenges and construct competitive solutions based on novel techniques.

Challenge:

The explosive growth of traffic makes store all datasets and then mining knowledge from them become very expensive and impossible;

Special challenges for telecomm big data analytics

Core Challenges
- Mining knowledge from big telecom data fast, efficiently with limited software and hardware resources (CPU, Memory, Storage, Network I/O)

The value of the datasets will be decreased as the time of processing increases;

1.5x
2.5x

1.5x
2.5x

- The explosive growth of traffic makes store all datasets and then mining knowledge from them become very expensive and impossible;
- The value of the datasets will be decreased as the time of processing increases;

2.5x
1.5x
End-to-end Big Data Stream Analysis

- **StreamSMART**
  - Algorithm oriented System optimization
  - Distributed Stream Processing Platform
  - Stream mining & Online Learning
  - Algorithm oriented System optimization

**SOC**
- Fault Localization
- User-level Network quality management
- Sketch based anomaly detection algorithm

**NOC**
- Alarm correlation
- WFM
- Adaptive Dynamic pattern mining algorithm

**SDN**
- real time DDoS detection
- Graph divided based load balance

- **StreamMBT**
  - Distributed streaming query meta language
  - Distributed fault tolerance

- **First online big data stream analytical solution for FBB/MBB networks**
- online Adaptive Dynamic pattern mining algorithm
- Multi-condition anomaly detection Algorithm
- Fault tolerance Solution is in the world leading level
- The fastest Stream process System in the world

- ✓ real time DDoS detection
- ✓ Alarm correlation
- ✓ WFM
- ✓ Fault Localization
- ✓ User-level Network quality management
- ✓ Fault tolerance Solution is in the world leading level

#StrataHadoop
Stream / Online Big Telecom Data Analysis Solution

- Distributed stream processing system for high-speed large telecom-data online analysis
- Load-balanced & Fault-Tolerance
- Stream SQL for efficient high-level programming
- Support iterative Stream mining & online learning algorithms;
- Provide algorithm-level fault-tolerance and approximation;
- Algorithm oriented System optimization
- System-level enhancements
- Online Model Parallelization & Iterative online ML support
- Stateful, Robustness & Efficiency
- From reactive to predictive
- Large scale but sparse value density data analysis
- Automatic continuous data pattern mining
- Extremely imbalanced streams with concept drift;
- Eco-system

APP

- Data Driven Intelligent Network Management;
- Data Driven Intelligent Network Control;

Stream mining & Online Learning

- Framework for general stream mining and online learning algorithms.
streamDM

- streamDM is designed specifically to be used inside Spark Streaming
streamDM

- Software for data stream mining on Spark Streaming

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

- Defined by V’s
  - Volume
  - Velocity
  - Variability
  - Variety
  - Value
  - Veracity
5W?
Why?
Huawei Embraces Open-Source Apache Spark

June 9, 2015 | by Bing Xiao

This is a guest blog from one of our partners: Huawei.

Join us at the Spark Summit to hear from Intel and other companies deploying Spark in production. Use the code DataBricks20 to receive a 20% discount!

It’s not unusual that one or more terabytes data flows in a telco network every second — this translates to roughly exabytes every month. In fact, the challenges go beyond the speed and volume of network flow data. For example, the location data is in original wireless coding format with complex nested structure, and leaves little room for compression; the signaling data, derived from multi-interfaces device of multi vendors in real-time and batch mode, requires complex association rules to make it meaningful and easily interpretable. Finally, the dynamic relationships across those data layers and...
Reduce the distance between Research discovery and implementation in Production
streamDM

- **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.
What?
Data Stream Setting

- 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, not static

- Approximation algorithms
  - Small error rate with high probability
streamDM

- Release 31/12/15
  - SGD Learner and Perceptron
  - Naive Bayes
  - CluStream
  - Hoeffding Decision Trees
  - Bagging
  - Stream KM++
- Release 30/06/16
  - Random Forests
  - Frequent Itemset Miner: IncMine
streamDM

- Methods for Regression
  - Adaptive Rules
  - Hoeffding Regression Trees
  - Bagging
  - Random Forests
- Methods for Clustering
  - DenStream
  - ClusTree
streamDM

- Methods for Frequent Pattern Mining
  - Itemsets
  - Sequences
  - Trees
  - Graphs
- Methods for Outlier Detection
  - AnyOut, Storm
  - Abstract-C, COD, MCOD
- Methods for Recommendation Systems
  - Collaborative Filtering: BRISMF
streamDM

- Methods for Novel Challenging Settings
  - Active Learning
  - Semi-supervised Learning
  - Multi-Label
  - Multi-Target
  - Unbalanced Data
How?
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 (SGD Learner -l 0.01 --o LogisticLoss
    -r Zero Regularizer)"`
Hoeffding Tree

- With high probability, constructs an identical model that a traditional (greedy) method would learn.

- With theoretical guarantees on the error rate.
Bagging builds a set of M base models, with a bootstrap sample created by drawing random samples with replacement.

Dataset of 4 Instances: A, B, C, D

- Classifier 1: B, A, C, B
- Classifier 2: D, B, A, D
- Classifier 3: B, A, C, B
- Classifier 4: B, C, B, B
- Classifier 5: D, C, A, C
Bagging builds a set of M base models, with a bootstrap sample created by drawing random samples with replacement.

Dataset of 4 Instances: A, B, C, D

- Classifier 1: A, B, B, C: A(1) B(2) C(1) D(0)
- Classifier 2: A, B, D, D: A(1) B(1) C(0) D(2)
- Classifier 3: A, B, B, C: A(1) B(2) C(1) D(0)
- Classifier 4: B, B, B, C: A(0) B(3) C(1) D(0)
- Classifier 5: A, C, C, D: A(1) B(0) C(2) D(1)
Random Forests

- Random Forests is one of the most popular methods in machine learning.
- **Bagging** builds a set of M base models, with a bootstrap sample created by drawing random samples with replacement.
- **Random Trees**: trees that in each node only uses a random subset of the attributes
CluStream

- 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++

- StreamKM++ computes a small weighted sample of the data stream, called the coreset of the data stream.

- Coreset of a set $P$ with respect to some problem
  - Small subset that approximates the original set $P$.
    - Solving the problem for the coreset provides an approximate solution for the problem on $P$. 
streamDM

- First Release
  - 31 December 2015
- Second Release
  - 30 June 2016
- Third Release
  - 31 December 2016
Who?
Contributors

- 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)
Summary
streamDM

- **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.
Epilogue
Open Source Software Projects 2015
Huawei Noah’s Ark Lab

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

- Supporting:
  - Apache SAMOA
  - MOA
Thanks!

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