Scalable Cohort Construction for Patient-level Predictive Modeling

Hang Su, BS1, Sherry Yan, Ph.D2, Walter (Buzz) F. Stewart, PhD, MPH2, Jimeng Sun, Ph.D1
1Georgia Institute of Technology, Atlanta, GA, USA, 2Sutter Health, Walnut Creek, CA, USA

Summary
Cohort construction, which aims at finding suitable subjects for a study, is the essential first step for clinical predictive modeling as illustrated in below figure. Existing tools that leverage SQL and relational databases are suffering from significant performance issues especially when the underlying data volume is large. There are two main challenges in cohort construction: 1) flexible and intuitive programming interface to describe complex criteria for the cohort, and 2) efficient computation for extracting the cohort from observational data. To address these challenges, we proposed a flexible domain specific language (DSL) for defining cohorts and developed a simple and efficient intermediate patient representation for supporting parallel cohort construction. We demonstrated the expressive power of the DSL using cohort construction for epilepsy refractory patient prediction as an example.

Methods
Method Overview
1. Input patient events
2. Parallelize with Apache Spark
3. Group events per patient
4. Process isolated patients

Data Model
• Everything as event
• Increasing order in timestamp
• Mandatory concept field
• Optional begin, end timestamp
• Optional additional key-value attributes

Event Manipulation
Filter by attributes and timestamp
Project as one to one event mapping
Sub-list input events as ordered list
Group multiple events to form new event

Outcome & Eligibility
Outcome: manipulate patient events and predict
• Existence of event
• Attribute of event

Eligibility: transform then aggregate events in temporal or non-temporal way
• Aggregation: $\text{Aggregate(events)}[> | = | <] \text{value}$
i.e. total hospital stay should be more than 10 days
• Temporal: $\text{events}_A \ [\text{before}]/\text{during} /\text{after} \ \text{events}_B$
i.e. Diagnosis before medication no more than 2 days

• Composite: combine multiple criteria using Boolean logic

Results
Predictive Modeling Use Case
Outcome:
• Case: at least 4 failures
• Control: 1 failure
Index: First AED failure
Eligibility:
• Two v780.39 or one v345,* followed by AED
• At least 16

Scalability
• Compare with Spark SQL
• Measure running time

Table 1: Statistics of datasets used in this work.

Conclusions
A new cohort construction module for predictive modeling has been developed. This module takes flexible events as input and chained event transformation mechanism is applied to define prediction outcome, index date and eligibility criteria. Running on top of Apache Spark made the utility scalable to processing large healthcare observational data.

Contact: Dr. Jimeng Sun < jsun@cc.gatech.edu >