Large Scale Sequence Analytics

Vladimir Pavlovic
I/UCRC Center for Dynamic Data Analytics
Department of Computer Science
Rutgers University

Project Definition

• Algorithms for rapidly & scalable analysis of large dynamic sets of sequences:
  - Music Genre
  - Artist, etc
  - Biological data
  - Financial indicators

Project Definition (cont’d)

• Unique focus: string representation for asserting similarity of sequences

• Benefits:
  – Scalability
  – Efficiency
  – Data-driven analytics & knowledge building
Previous Related Work

• Large body of work. String kernels:
  – Vishwanathan & Smola. Fast kernels for string and tree matching, NIPS 2002
  – Leslie & Kuch, Fast string kernels using inexact matching, JMLR 2004
  – Rosu & Thoma: Efficient computation of gapped substring kernels on large alphabets, JMLR 2005
  – Lodhi et al. Text classification using string kernels, JMLR 2002

Previous Related Work (cont’d)

• String spectrum kernel concept: dealing with non-Euclidean sequence spaces

Our prior work

  – Efficient, alphabet-independent combinatorial algorithms
  – Multiresolution (sparse spatial sample) embedding
  – Extraction of motifs (common string features)
Proposed Work

• Scalable string methods for data-driven market segmentation
  – Data-driven dynamic grouping for discovery of market segments
  – Distributed implementation

Data-driven Segmentation (cont’d)

• Obtain dynamic grouping of companies in “market segments” from daily market indicators (closing price, returns, volume, etc)
• Extract cluster prototypes, analyze boundary points, and transition of points (companies) between clusters
• Contrast to GICS (Global Industry Classification Standard) static grouping

• Challenges:
  – Impact on trading?
  – Domain-specific symbolic representation

Distributed Implementation

• Distributed combinatorial algorithms for string kernels
• Exploit data and algorithmic parallelization
  – Efficient combinatorial algorithms are based on multiple (possibly concurrent) rounds of counting sort – ideal setting for distributed implementation
  – Alphabet-size independence assures scalability to domains large alphabets
• Implementation on Amazon Elastic Compute Cloud (Amazon EC2)
Implication of Project

- **Possible Applications**
  - Rapid & scalable sequence analytics for data-driven automated trading (e.g., pairs/segment trading)
  - Large-scale text analysis
  - DNA barcoding analytics for biodiversity assessment
- **Impact on Industry**
  - Increase portfolio of advanced, tailored data & knowledge extraction methods

Deliverables

- **First 6 months**
  - Fast algorithms and symbolic sequence representation
- **12 months and associated knowledge transfer**
  - Distributed implementation
  - Analysis of data-driven groupings
- ** Longer term plans and ideas**
  - Information-theoretic sequence embedding: sequences are points on probabilistic density manifolds. How to efficiently compute similarity?
  - Distance-preserving bitstring embedding: find symbolic representations that reflect distances in original spaces (e.g., Hamming distance ≈ Euclidean distance)
Project Name: Scalable String Methods for Data-Driven Market Segmentation

Project Investigators: Dr. V. Pavlovic (PI), C. Hendahewa, S.-H. Yi (students)

Description:
Segmentation of markets into sectors, the distinct subset whose components share similar characteristics, is vital for understanding and acting upon the changing market dynamics. Trading companies make critical predictions, in part, based on co-analysis of trends within homogeneous sectors. Most current sectors are static and unrelated to trading objectives, reducing their utility for making accurate market decisions.

Segmentation of trading markets involves comparative analysis of short and long-term temporal patterns exhibited by a variety of financial indicators, such as the trading frequency, stock price, spread, etc. In this project we will assert dynamic segment using data-driven symbolic representations of sequences, where each sequence segment is represented as a symbol in a (potentially large) alphabet. Similarity between pairs of sequences will be computed using a family of scalable and efficient string kernels developed in our prior work. Based on this similarity our goal will be to, for different temporal windows (e.g., 3mo, 6mo, 1yr) construct clusters of companies that exhibit similar patterns as well as follow the behavior of these clusters over time (cluster centers, boundary points, cluster point exchanges, and spill-overs).

We will evaluate our models by contrasting the learned segments with the traditional (e.g., manufacturing) sectors and by assessing the virtual “trading” gain obtained using the predicted sectors. The evaluation data will be provided by our industrial partners and the Wharton WRDS database.

Experimental Plan:
- Sept. 10: Data collection and symbolic representation
- Oct. 10: Development of Similarity and Dynamic clustering algorithms
- Spring 11: Testing of algorithms / distributed implementation
- Summer 11: Performance evaluation

Related Work Elsewhere:
- Market-segment models based on stochastic PDEs (typically in-house/unpublished)

How Ours Is Different:
- Symbolic string pattern representation
- Highly computationally efficient and scalable

Related Work in Center:
- String kernel methods for text and biological sequence analysis

Milestones:
- 2010-2011: Data-driven symbolic representation and dynamic clustering
- 2011: Extend work by distance-preserving bitstring embedding of sequences

Deliverables:
- Technical demonstration along with a technical report resulting in a publication.

Budget: $60,000

Potential Benefits to Member Companies:
- Increase portfolio of analytic methods for dynamic temporal/sequential pattern analysis
- Distributed, large-scale implementation of sequence analysis algorithms