An Introduction to Distributed Data Mining

Course Description

Many distributed computing environments store and process massive amounts of data. Examples include the internet, intranets, local and wide area networks, adhoc wireless and sensor networks. Learning patterns from these distributed data sources involves development of scalable algorithms for mining and knowledge discovery.

This course will present the state-of-the-art in mining distributed data and discuss open research problems and challenges in this area. We will discuss data models and communication protocols typically used for learning in distributed environments and present techniques for distributed classification (such as decision tree construction, ensemble learning and construction of support vector machines), association rule, bayesian network and unsupervised learning (such as clustering using multiple rounds of communication and ensemble-based models). Application of distributed mining algorithms on cloud computing environments, data grid and peer-to-peer networks will be discussed. Real-world applications and case studies in the areas of complex networks and graph mining, social networks, scientific computing (medical, astronomy, ecology) and industrial applications (such as from the energy domain) will be presented to stimulate the need for research in this area.

The class format will be a combination of lectures given by the instructor, students and occasional guest speakers, as well as discussions of recent papers on selected topics. These topics will be posted on the class website and updated based on participant interest. Students are expected to give one (or two) short presentations. There will be no exams but a semester long project, comprising of four main assignments: a project proposal, an intermediate report, the final submission, and a class presentation on the project. Some of the good projects will be considered for publication in conferences depending on the interest of students.

Text Books

Pre-requisites

One of the following courses: Introduction to Artificial Intelligence (4701) or Machine learning (4771). An equivalent course taken at another university can also substitute. Auditing would require permission of the instructor.

Biographical Sketch of the Instructor

Haimonti Dutta is an Associate Research Scientist at the Center for Computational Learning Systems (CCLS), Columbia University, NY. She received her Ph.D. in Computer Science and Electrical Engineering (CSEE) from the University of Maryland, Baltimore County (UMBC) in 2007, MS in Computer and Information Science from Temple University (2002) and Bachelor of Computer Science and Engineering (BCSE) from Jadavpur University (Kolkata, India), 1999. Her research interests include machine learning, data mining and pattern recognition; distributed optimization; data intensive computing; distributed and scientific data mining. She has been on the program committee for many conferences including Knowledge Discovery and Data Mining Conferences (KDD), International Conference on Data Mining (ICDM), SIAM Data Mining Conference (SDM), European Conference on Machine Learning (ECML) and has presented/published research papers at many prestigious venues including ICDM, SIAM Data Mining Conference, ICML, HiPC and ICMLA. She is a recipient of the Dr B. C. Roy Scholarship for academic excellence and the UMBC Graduate Dissertation Fellowship, and was nominated for the Best Paper Award at the International Conference on Machine Learning and Applications (ICMLA) in 2008.

Tentative Course Schedule

1. Distributed Data Mining: An Introduction
   - Motivation
   - Data Models in DDM
   - Synchronous vs Asynchronous Algorithms
2. Communication Protocols
   - Gossip-based Communication
   - Randomized Gossip Algorithms
   - Coverge-cast, Up-Cast and Down-Cast
3. Distributed Tree Constructions
   - Distributed Breadth First Search
• Distributed Depth First Search
• Distributed Minimum Spanning Tree Construction

4. Distributed Classification
   • Ensemble-based Learning – Bagging, Boosting, Random Forests, Stacking and Rotation Forests
   • Classifiers built on homogeneously partitioned data – The Meta-learning Framework
   • Classifiers built on heterogeneously partitioned data

5. Distributed Support Vector Machines

6. Distributed Association Rule Learning

7. Distributed Bayesian Network Learning.

8. Distributed Clustering
   • Multiple Communication Round Algorithms
   • Centralized Ensemble-based Methods

9. Distributed Algorithms for Mining on Large Scale Systems
   • Mining on the Grid
   • Algorithms for Mining on Peer-to-Peer Systems

10. Real World Applications and Case Studies
    • Medicine – Mining EEG Data on the Cloud
    • Industrial Application – Distributed Computing for Energy Applications
    • Link Analysis – Citation Graph Mining
    • Scientific Computing – Mining Astronomy Sky Surveys
    • Learning on Peer-to-Peer Networks such as Kazaa, eMule

11. Future Directions
    • Open Research Issues
    • Potential Applications for the Grid, Cloud and Peer-to-Peer Networks

12. Resources
    • References
    • Web Resources
    • Projects
A list of upto 20 most significant publications that will be covered in the course

References


