KDD-2008 Workshop Report
DMMT'08: Data Mining Using Matrices and Tensors

Chris Ding
Comp. Sci. & Eng. Dept.
University of Texas at Arlington
chqding@uta.edu

Tao Li
School of Computer Science,
Florida International University
taoli@cs.fiu.edu

Shenghuo Zhu
NEC Laboratories America
zsh@sv.nec-labs.com

ABSTRACT

We provide a summary of the Workshop on Data Mining Using Matrices and Tensors (DMMT'08) held in conjunction with ACM SIGKDD 2008, on August 24th in Las Vegas, USA. About 100 people attended the workshop. We report in detail about the research issues addressed in the talks at the workshop. More information about the workshop can be found at `http://www.cs.fiu.edu/~taoli/kdd08-workshop`.

1. INTRODUCTION

The field of pattern recognition, data mining and machine learning increasingly adapt methods and algorithms from advanced matrix computations, graph theory and optimization. Prominent examples are spectral clustering, non-negative matrix factorization, Principal component analysis (PCA) and Singular Value Decomposition (SVD) related clustering and dimension reduction, tensor analysis such as 2DSVD and high order SVD, L-1 regularization, etc. Compared to probabilistic and information theoretic approaches, matrix-based methods are fast, easy to understand and implement; they are especially suitable for parallel and distributed-memory computers to solve large scale challenging problems such as searching and extracting patterns from the entire Web. Hence the area of data mining using matrices and tensors is a popular and growing area of research activities.

This workshop presents recent advances in algorithms and methods using matrix and scientific computing/applied mathematics for modeling and analyzing massive, high-dimensional, and nonlinear-structured data. One main goal of the workshop is to bring together leading researchers on many topic areas (e.g., computer scientists, computational and applied mathematicians) to assess the state-of-the-art, share ideas and form collaborations. We also wish to attract practitioners who seek novel ideals for applications. In summary, this workshop strives to emphasize the following aspects:

- Presenting recent advances in algorithms and methods using matrix and scientific computing/applied mathematics
- Addressing the fundamental challenges in data mining using matrices and tensors
- Identifying killer applications and key industry drivers (where theories and applications meet)
- Fostering interactions among researchers (from different backgrounds) sharing the same interest to promote cross-fertilization of ideas.
- Exploring benchmark data for better evaluation of the techniques

2. TOPIC AREAS

The Topic areas for the workshop include (but are not limited to) the following:

Methods and algorithms:

- Principal Component Analysis and Singular value decomposition for clustering and dimension reduction
- Nonnegative matrix factorization for unsupervised and semi-supervised learning
- Spectral graph clustering
- L-1 Regularization and Sparsification
- Sparse PCA and SVD
- Randomized algorithms for matrix computation
- Web search and ranking algorithms
- Tensor analysis, 2DSVD and high order SVD
- GSVD for classification
- Latent Semantic Indexing and other developments for Information Retrieval
- Linear, quadratic and semi-definite Programming
- Non-linear manifold learning and dimension reduction
- Computational statistics involving matrix computations
- Feature selection and extraction
- Graph-based learning (classification, semi-supervised learning and unsupervised learning)
- Matrix factorization for classification

Application areas:

- Information search and extraction from Web
- Text processing and information retrieval
- Image processing and analysis
- Genomics and Bioinformatics
- Scientific computing and computational sciences
- Social Networks

3. WORKSHOP OVERVIEW

The **2008 Workshop on Data Mining using Matrices and Tensors (DMMT'08)** is the first workshop on this theme held annually with the SIGKDD Conference. Through the workshop, we expect to bring together leading researchers on many topic areas (e.g., computer scientists, computational and applied mathematicians) to assess the state-of-the-art, share ideas and form collaborations. We also wish to attract practitioners who seek novel ideals for applications.

The program of the workshop included a keynote talk by Prof. Michael I. Jordan from University of California at Berkeley, three invited talks by Prof. Christos Faloutsos from Carnegie Mellon University, Prof. Haesun Park from Georgia Institute of Technology, and Prof. Lenore R. Mullin from NSF CISE CCF Theoretical Foundations Clusters. There are also seven research paper presentations. About 100 people attended the workshop. The on-line proceedings of the workshop is available at
`http://www.cs.fiu.edu/~taoli/kdd08-workshop/`.

4. INVITED SESSION

4.1 Keynote Talk

The workshop program is started by a keynote talk entitled "sufficient dimension reduction" by Prof. Michael I. Jordan from University of California at Berkeley. The problem of "sufficient dimension reduction" (SDR) is that of finding a subspace S such that the projection of the covariate vector X onto S captures the statistical dependency of the response Y on X. Prof. Jordan first presented a general overview of the SDR problem, focusing on the formulation of SDR in terms of conditional independence. He also discussed some of the popular algorithmic approaches to SDR, particularly those based on inverse regression. Finally, He described a new methodology for SDR which is based on the characterization of conditional independence in terms of conditional covariance operators on reproducing kernel Hilbert spaces (a general characterization of conditional independence that is of independent interest).

4.2 Invited Talks

The workshop program also included three invited talks by Prof. Christos Faloutsos from Carnegie Mellon University, Prof. Haesun Park from Georgia Institute of Technology, and Prof. Lenore R. Mullin from NSF CISE CCF Theoretical Foundations Clusters.

Prof. Christos Faloutsos's talk is about surprising patterns in large graphs. He reviewed some 'laws' for static as well as evolving graphs (e.g., how do graphs look like? How do they evolve over time? How can we generate realistic-looking graphs?). He then provided some recent discoveries on blogs and influence propagation and described some tools to help us analyze large graphs (e.g., The 'Kronecker' generators, The CenterPiece subgraphs to spot central nodes in a community, and incremental tensor analysis to spot anomalies in Internet traffic data). Finally, he discussed emerging map/reduce approach and its impact on large graph mining.

Prof. Haesun Park talked about linear discriminant analysis (LDA) and its generalizations. Linear Discriminant Analysis (LDA) has been utilized as a method of choice for dimension reduction of clustered data. Prof. Park presented Linear Discriminant Analysis (LDA) has been utilized as a method of choice for dimension reduction of clustered data. The LDA/GSVD can be nonlinearized by using kernel functions. Some experimental results are presented in text classification, facial recognition, and fingerprint classification, to demonstrate the effectiveness of the proposed methods.

Prof. Lenore R. Mullin of National Science Foundation gave a full presentation about the new NSF/CISE initiatives in numerical computation, optimization, high performance computing. She especially emphasized tensor analysis as the rise of multi-linear arrays in data mining. Audience asked Dr. Mullin many questions about NSF funding opportunities.

5. OVERVIEW OF THE RESEARCH PRESENTATIONS

The workshop program included several research presentations.

Chris Ding from UT Arlington presented some recent theoretical progress in tensor clustering and error Bounds. He and his colleagues recently developed theoretical proof to show that the widely used ParaFac and HOSVD tensor decompositions are in fact performing simultaneous K-means data clustering and subspace factorization. This work extends the earlier development on the equivalence between K-means clustering and principal component analysis (PCA), and the equivalence between K-means clustering and non-negative matrix factorization (NMF). They also presented lower and upper bounds on the tensor reconstruction errors, similar to the Eckart-Young error formulation for Singular Value decomposition (SVD). Experiments on 3 image datasets are presented.

Evrim Acar from Sandia National Laboratories introduced their work on understanding Epilepsy seizure structure using tensor analysis. She introduced mathematical models based on multi-modal data construction and analysis with a goal of understanding epilepsy seizure dynamics and developing automated and objective approaches for the analysis of large amounts of scalp electroencephalogram (EEG) data. Seizure recognition aims to automatically differentiate between seizure and non-seizure periods. In their work, the multi-channel EEG signals were first rearranged as a third-order tensor with modes: time epochs, features and channels. Then a multilinear regression model, i.e., Multilinear Partial Least Squares (N-PLS), which is the generalization of Partial Least Squares (PLS) regression to higher-order

datasets, was used to model the tensor. The two-step approach facilitates EEG data analysis from multiple channels represented by several features from different domains. She showed that their approach gave promising results in terms of identifying seizure origins as well as marking seizure periods.

In their paper, S.K. Tasoulis (University of Patras) and D.K. Tasoulis (Imperial College London) proposed an improvement of the Principal Direction Divisive Partitioning algorithm from three perspectives: (1) how to split a cluster, (2)which cluster to split, and (3) stopping criterion. Their proposed algorithm merges concepts from density estimation and projection-based methods towards a fast and efficient clustering algorithm, capable of dealing with high dimensional data. Experimental results showed improved partitioning performance compared to other popular methods. They also explored the problem of automatically determining the number of clusters.

Vaclav Snasel (Technical University of Ostrava) et al. proposed two methods for boolean matrix factorization: an artificial neural network based boolean factorization and a genetic algorithm for boolean matrix factorization. The neural network boolean factorization is based on the Hopfield-like neural network model while the genetic algorithm based boolean factorization makes use of a constructive algorithm for suggesting base vectors. Experiments were also conducted to evaluate the two proposed algorithms.

Motivated by the observation that simple reformulation of Gaussian processes can lead to much faster execution times on graphs, Thomas Gartner and Shankar Vembu from Fraunhofer Institute IAIS, Germany presented several strategies for efficient implementations of kernel methods with graph kernels. In particular, regularized least squares and support vector machine were discussed in detail to illustrate these strategies. The authors also showed how to combine these strategies with other popular algorithms for graphs, including graph ranking algorithms and low dimensional embedding algorithms. A toolkit is implemented in python for regularized least squares and support vector machine.

In their work, Shipeng Yu (Siemens Medical Solutions), Jinbo Bi (Siemens Medical Solutions) and Jieping Ye (Arizona State University) introduced the probabilistic higher-order PCA (PHOPCA), a family of probabilistic models for 2D (and higher-order) data. They showed that PHOPCA recovers the optimal solutions of several PCA-style algorithms under mild conditions. Efficient EM-type algorithms were derived for learning, with less time complexity than the non-probabilistic counterparts. Several extensions of PHOPCA were also discussed. Some empirical results were presented using face images, USPS handwritten digits and a real application in cardiac view recognition of echocardiogram.

It has been shown that many SVM models can be formulated into quadratic programming while the path-tracing for SVMs (e.g., the task of tracing the regularized piecewise linear solution path for SVMs) can be attacked by parametric quadratic programming (PQP). Zhili Wu (Hong Kong Baptist University) et al. considered the relation between path-tracing for SVMs and the generalized mean-variance portfolio optimization from a PQP view. The relation allows the path-tracing task to be handled by tailoring the critical line algorithm (CLA) originally proposed for mean-variance portfolio optimization. The CLA algorithm systematically utilizes the equality and bounding constraints in the PQP formulation and leads to a robust one-per-iteration approach based on Karush-Kuhn-Tucker conditions.

6. WORKSHOP ORGANIZATION

Work General Chair
Hongyuan Zha, Georgia Institute of Technology

Work Co-chairs
Chris Ding, University of Texas at Arlington
Tao Li, Florida International University
Shenghuo Zhu, NEC Laboratories America

Committee Members
Tammy Kolda, Sandia National Labs
Jesse Barlow, Penn State University
Michael Berry, University of Tennessee
Yun Chi, NEC Laboratories America
Lars Elden, Linkping University, Sweden
Christos Faloutsos, Carnegie Mellon University
Estratis Gallopoulos, University of Patras
Joydeep Ghosh, University of Texas at Austin
Ming Gu, University of California, Berkeley
Michael Jordan, University of California, Berkeley
Yuanqing Lin, University of Pennsylvania
Huan Liu, Arizona State University
Michael Ng, Hong Kong Baptist University
Haesun Park, Georgia Tech
Wei Peng, Xerox Research
Robert Plemmons, Wake Forest
Alex Pothen, Old Domino University
Yousef Saad, University of Minnesota
Horst Simon, Lawrence Berkeley National Laboratory
Fei Wang, Florida International University
Jieping Ye, Arizona State University
Kai Yu, NEC Laboratories America
Hongyuan Zha, Georgia Tech
Zhongyuan Zhang, Chinese Academy of Sciences

Most submissions were reviewed and discussed by two reviewers and workshop co-chairs. We are very indebted to all program committee members who helped us organize the workshop and reviewed the papers very carefully. We would also like to thank all the authors who submitted their papers to the workshop; they provided us with an excellent workshop program. More information about the workshop can be found at `http://www.cs.fiu.edu/~taoli/kdd08-workshop/`.