KDD-2009 Workshop Report
DMMT'09: 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

ABSTRACT

We provide a summary of the Workshop on Data Mining Using Matrices and Tensors (DMMT'09) held in conjunction with ACM SIGKDD 2009, on June 28th in Paris, France. More than 50 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/kdd09-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

This DMMT'09 workshop is a continuation of the theme of SIGKDD 2008 Workshop on Data Mining using Matrices and Tensors (DMMT'08). DMMT'08 was 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 invited talks by Prof. Lenore Mullin from National Science Foundation and SUNY Albany; Prof. James Raynolds from SUNY Albany; Prof. Charles Elkan from University of California at San Diego; and Prof. Leiven De Lathauwer from Katholieke Universiteit Leuven, Belgium. There are also several research paper presentations. More than 50 people attended the workshop. The on-line proceedings of the workshop is available at `http://www.cs.fiu.edu/~taoli/kdd09-workshop/`.

4. INVITED TALKS

The workshop program is started by an invited talk entitled "Tensor Decompositions and Applications: a Survey" by Prof. Leiven De Lathauwer from Katholieke Universiteit Leuven, Belgium. Dr. Lathauwer is the developer of High-Order SVD (HOSVD). He gave a survey on tensor generalizations of the Singular Value Decomposition (SVD) and their applications. He also discussed some developments on Nonnegative Tensor Factorizations.

Prof. Charles Elkan from University of California at San Diego also gave an invited talk on factorizing matrices with missing entries. The known algorithms for approximate factorization of large matrices are so diverse that the multiple approaches have never been explained in one place, and have never been fully compared experimentally. Charles outlined the available alternatives, focusing especially on experimental results, on methods for factorizing matrices with unknown entries, and on methods based on stochastic gradient descent.

Prof. Lenore Mullin and Prof. James Raynolds gave a joint invited talk entitled "Tensors and n-d Arrays: Mathematics of Arrays, Psi-Calculus, and Composition of Tensor and Array Operations". Prof. Mullin is currently a program director for Algorithmic Foundations (AF) in Division of Computing and Communication Foundations (CCF) of Directorate for Computer & Information Science & Engineering (CISE) at National Science Foundation (NSF). They discussed the outer product/tensor product and a special case of the tensor product: the Kronecker product, along with optimal implementation when composed, and mapped to complex processor/memory hierarchies. They also demonstrated that how the use of "A Mathematics of Arrays" (MoA), and the psi-Calculus, (a calculus of indexing with shapes), provides optimal, verifiable, reproducible, scalable, and portable implementations of both hardware and software.

5. OVERVIEW OF THE RESEARCH PRESENTATIONS

The workshop program included several research presentations.

In their paper, Frank Nielsen (Ecole Polytechnique, France & Sony CSL, Japan) and Aurelien Serandour (Ecole Polytechnique, France) presented an empirical comparison of various distance metric-learning algorithms including optimization based metric learning and information theoretic metric learning using six UCI datasets. The study indicated that the performance results of the algorithms are largely dependent on the data characteristics (e.g., size, dimensions). It appears that no algorithm tends to dominate the other ones. Furthermore, results on well-defined sets may not represent the behavior on human-built ones. In summary, the study demonstrates the difficulty of distance learning and calls for more work to address various research challenges.

Eman Abdu and Douglas Salane from the City University of New York presented a spectral-based algorithm, SCCADDS (Spectral-based Clustering algorithm for CAtegorical Data using Data Summaries), for clustering categorical data that combines attribute relationship and dimension reduction techniques found in Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI). SCCADDS uses data summaries that consist of attribute occurrence and co-occurrence frequencies to create a set of vectors each of which represents a cluster and also utilizes spectral decomposition of the data summaries matrix to project and cluster the data objects in a reduced space. Comparing with existing spectral clustering methods, SCCADDS has several new features: 1) It uses the attribute categories similarity matrix instead of the data object similarity matrix and can scale well for large datasets since in most categorical clustering applications the number of attribute categories is small relative to the number of data objects; 2) It clusters the data objects directly by comparing them to candidate cluster representatives without the need for an iterative clustering method; and 3) Its complexity is linear in terms of the number of data objects. Experiments were conducted to demonstrate the effectiveness of the proposed method.

Mikhail Krivenko and Vitaly Vasilyev from the Institute of Informatics Problems of the Russian, Academy of Sciences, Moscow presented their work on sequential latent semantic indexing (SLSI). The main difference of the SLSI from the existing sequential algorithms is that the dimension of space is not fixed and dynamically changes to ensure a given level

of relative approximation error of a matrix of observations. The authors provided theoretical and experimental justification of the effectiveness of the proposed method. The experiments with the different collections of texts demonstrated that the SLSI algorithm could be seen as a tradeoff solution, which have a lower computational complexity and memory requirements compared to the standard LSI method and did not lead to a decrease of the quality of classification in contrast to other sequential algorithms.

In their work, Elisabeth Georgii (MPI for Biological Cybernetics/Friedrich Miescher Laboratory, Germany), Koji Tsuda (MPI for Biological Cybernetics, Germany), and Bernhard Scholkopf (MPI for Biological Cybernetics, Germany) described an an enumerative approach called DCE (Dense Cluster Enumeration) to identify dense clusters in tensors of arbitrary dimensionality. The density criterion is exploited in an effective way using a reverse search algorithm. In addition, DCE ranks the results by exact p-values and can deal with symmetry constraints. Compared to methods that co-analyze multiple networks, DCE is more general and flexible, allowing to analyze tensor data with an arbitrary number of dimensions, real or binary values, including symmetries or not. In addition, the size of the solution set of DCE can be controlled by tuning the density threshold and the weight distribution or sparsity of the tensor. The authors also discussed their future work on improving the efficiency of DCE for large-scale applications and on integrating background knowledge.

Mario Navas and Carlos Ordonez from University of Houston studied how to leverage a DBMS computing capabilities to solve Principal Component Analysis (PCA). They proposed a solution that combines a summarization of the data set with the correlation or covariance matrix and then solves PCA with Singular Value Decomposition (SVD). Deriving the summary matrices allow analyzing large data sets since they can be computed in a single pass. They introduced two solutions for solving SVD without external libraries: one based in SQL queries and a second one based on User-Defined Functions. Experimental evaluation demonstrated that their proposed method can solve larger problems in less time than external statistical packages.

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.

6. WORKSHOP ORGANIZATION

Work Co-chairs

Chris Ding, University of Texas at Arlington
Tao Li, Florida International University

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, Central University of Finance & Economics
Shenghuo Zhu, NEC Laboratories America

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/kdd09-workshop/`.