Final Report for Period: 10/2009 - 09/2010
Submitted on: 05/03/2011
Principal Investigator: Park, Haesun .
Award ID: 0621889
Organization: Georgia Tech Research Corp
Submitted By:
Park, Haesun - Principal Investigator
Title:
CompBio: Collaborative Research: Development of Effective Gene Selection Algorithms for Microarray Data Analysis

Project Participants

Senior Personnel

Name: Park, Haesun
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Kim, Wooyoung
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Kim, Jingu
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Balasubramanian, Krishnakumar
Worked for more than 160 Hours: Yes
Contribution to Project:

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

Ding-Zhu Du: he is the collaborator and PI at the none-lead institute.

Lars Elden: Department of Mathematics, Linkoping University, Sweden
He has been collaborating in designing a tensor extension of the non-negative matrix factorization.

Activities and Findings

Research and Education Activities:

Some of the major problems that we studied extensively in this project include nonnegative matrix factorization (NMF), its extension to tensors, and the l1-regularized linear regression. We applied these methods to design automatic gene selection method and protein sequence motif information discovery. Due to the nonnegativity constraints in NMF, the factors of its lower-rank approximation provide a natural interpretation: each data item can be explained by an additive linear combination of physically meaningful basis components. In addition, NMF can work as a successful clustering method when an additional constraint of sparsity is imposed on the second nonnegative factor in NMF. Numerous successful applications of NMF were reported in areas including text mining, computer vision, and bioinformatics. In addition, the l1-regularized linear regression gives sparse solution and can be used in designing automatic gene selection methods.

One of many areas where we have made some significant contributions is development of efficient algorithms and theoretical study of their convergence properties. Several algorithms were already developed for NMF including the multiplicative updating method, the alternating least squares method, the active-set method, the projected gradient method, and the projected quasi-Newton method. However, none of these methods is fully optimized for the special characteristics of NMF computation. By exploiting the special characteristics, we designed a new algorithm for computing NMF. The algorithm is based on a fast active-set-type method called block principal pivoting method, which overcomes some limitations of traditional active-set methods. To evaluate the effectiveness and efficiency of the new algorithm, we performed extensive experimental comparisons of the new algorithm with previously developed ones.

NTF (Nonnegative Tensor Factorization) provides lower rank approximation of nonnegative tensors (either in PARAFAC-PARAllel FACtorization, or in Tucker decomposition forms) with nonnegative factors. NTF shares with NMF the property that its lower-rank approximation provides meaningful interpretations. Numerous data analysis algorithms have been designed for data sets which are typically represented as multidimensional arrays, i.e., matrices, tensors. Matrices and tensors provide a mathematical and algorithmic framework for analyzing multiscale, multidimensional data sets and extracting meaningful information from such data. Tensor factorization, a multilinear generalization of matrix factorization, is a powerful tool for multidimensional data analysis. It gives a meaningful lower rank approximation which can further be used for dimensionality reduction as well as visualization. Since modern data sets are multiscale and multidimensional, where each dimension is typically very high, fast algorithm for processing such data is essential. We developed an algorithm for NTF based on alternating nonnegative least squares (ANLS) and an active set type method called the block pivoting principle. While on the theoretical front, important and exciting problems arise with respect to rank of nonnegative factorization, from an application point of view, nonnegative tensor factorization plays an important role to analyze multidimensional, inherently nonnegative data. Our algorithm provides a faster way of computing such factorizations and extensions for regularized and sparse factorization are provided under the same framework.

We studied application of NMF to two problems in bioinforamtics. One of the applications was using sparse NMF for finding motifs from protein sequences. Finding motifs involves detecting groups of similar protein-segments from a large collection of protein sequences, and sparse NMF was successfully used to find such groups. Another application was using sparse NMF and its variant for clustering and semi-supervised clustering tasks in microarray analysis, respectively. We observed that sparsity constraints often substantially improve the clustering results obtained using NMF.

Another topic that we have explored is the L1 regularized linear regression, which is a linear regression problem in which the L1 norm of coefficient vector is constrained. The method is known to simultaneously avoid the over-fitting to training data and achieve sparsity in the computed solution. The sparsity has two important benefits; it improves the interpretation of the model by explicitly showing the relationship between the target variable and features, and it also allows computationally efficient model because only a small number of coefficients remain nonzero. We developed new models and efficient algorithms for toxic chemical agent detection applying L1 regularized regression with sparsity and nonnegativity constraints.

All of the above research involved graduate students supported by the grant. In addition to weekly research meetings, the students were provided with opportunities to learn presentation skills through the seminars they give throughout the academic year.

Findings:

The effectiveness of each work described above was examined through implementation and experiments. The new algorithm that we developed for NMF has been confirmed to be significantly faster than other existing ones. In experiments, we observed how several algorithms reduce the objective function value with respect to computation time. The results showed that the new algorithm provided the lowest objective function value with any amount of computation time. In the work on the algorithm for NTF, the new algorithm also showed the fastest performance among all existing algorithms. Its theoretical convergence property has been studied and we discovered that every limit point produced by the

algorithm is a stationary point, which is the best one can expect in general due to non-convex nature of NMF.

Applying NMF for protein motifs discovery and clustering, successful results was obtained.
The problem of discovering motifs from protein sequences is a critical and challenging task in the field of bioinformatics. The task involves clustering relatively similar protein segments from a huge collection of protein sequences and culling high quality motifs from a set of clusters. A granular computing strategy combined with K-means clustering algorithm was previously proposed for the task, but this strategy requires a manual selection of biologically meaningful clusters which are to be used as an initial condition. This manipulated clustering is undisciplined as well as computationally expensive. In our work, we utilize sparse non-negative matrix factorization (SNMF) to cluster a large protein data set. We show how to combine this method with Fuzzy C-means algorithm and incorporate bio-statistics information to increase the number of clusters whose structural similarity is high. Our experimental results show that an SNMF approach provides better protein groupings for similar secondary structures while maintaining similarities in protein primary sequences.

Regarding the l1-norm regularized linear regression, our new algorithm has been confirmed to be significantly faster than existing methods. In the comparison, several types of existing algorithms were included: active-set-type methods (such as least angle regression and the feature-sign search algorithm), projected gradient methods, and coordinate descent methods. Under various conditions of test problems, the new algorithm outperformed these methods.

Regarding nonnegativity constrained tensor factorizations, experimental results conclude that indeed nonnegative tensor factorization of inherently nonnegative multidimensional data provided better solutions with better interpretation capability. Compared to existing algorithms for NTF, our algorithm computes the factors in less time. In order to verify that indeed only the right factors are recovered, we adopted a visual approach where we formed a tensor based on three images (as a multilinear combination of outer product of vectors) and used our algorithm for recovering the factors in the presence of noise. It was seen that our algorithm was able to extract the factors with surprisingly very limited loss of information only.

Training and Development:

Outreach Activities:
The PI has lead an effort to involve the undergraduate students by supervising two summer interns from North Carolina A&T university who are African Americans and a female student from Harvey-Mudd college during the summer of 2009. This effort continued in the summer of 2010 with multiple students participating from the US and India.

Journal Publications

H. Kim and H.Park, "Sparse non-negative matrix factorization via alternating non-negativity constrained least square for microarray data analysis", Bioinformatics, p. 1495, vol. 23-12, (2007). Published,

H. Kim, H. Park, and B. Drake, "Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations", BMD Bioinformatics, p. , vol. 8, (2007). Published,

H. Kim and H. Park, "Cancer class discovery using non-negative matrix factorization based on alternating non-negativity-constrained least squares", Lecture Notes in Computer Science, Bioinformatics Research and Applications, p. 477, vol. 4462, (2007). Published,

H. Kim and H. Park, "Nonnegative Matrix Factorization based on Alternating Nonnegativity Costrained Least Squares and Active Set Method", SIAM Journal on Matrix Analysis and Applications, p. 713, vol. 30-2, (2008). Published,

C.Park, M. Jeon, P. Pardalos, and H. Park, "Quality assessment of gene selection in microarray data", Optimization Methods and Software, p. 145, vol. 22, (2007). Published,

S. Lee, S. Soak, M. Jeon, and H. Park, "Statistical properties analysis of real world tournament selection in genetic algorithms", Applied Intelligence, p. 195, vol. 28-2, (2008). Published,

W. Kim, B. Chen, J. Kim, Y. Pan, and H. Park, "Sparse Nonnegative Matrix Factorization for Protein Sequence Motif Discovery", Expert Systems and Applications, p. , vol. , (2011). Accepted,

Books or Other One-time Publications

M. Mallick, B. Drake, H. Park, et al., "Comparison of Raman spectra estimation algorithms", (2009). Conference Proceedings, Accepted
Collection: Proceedings of the 12th International Conference on Information Fusion
Bibliography: pp. 2239-2246

J. Kim and H. Park, "Toward faster nonnegative matrix factorization: a new algorithm and comparisons", (2008). Conference Proceedings, Published
Collection: Proceedings for the eighth IEEE International Conference on Data Mining
Bibliography: pp. 353-362

B. Drake, J. Kim, M. Mallick, and H. Park, "Supervised Raman Spectra Estimation based on Nonnegative Rank Deficient Least Squares", (2010). Book, Published
Collection: Proceedings of the 13th International Conference on Information Fusion
Bibliography: Not known yet.

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

Our research on NMF and NTF provided efficient algorithms, better insights, and successful applications. The algorithms newly developed for NMF and NTF provide more efficient ways to compute lower rank approximation. The new algorithms were validated through experimental comparisons, so they can be used in many practical applications. In addition, the observation on the special characteristics of NMF and NTF computations that we utilized to develop the efficient algorithms will provide better insights on the computational aspect of each problem. The problem of discovering motifs from protein sequences is a critical and challenging task in the field of bioinformatics. The task involves grouping relatively similar protein segments from a huge collection of protein sequences. A granular computing strategy combined with K-means clustering algorithm was previously proposed for the task, but, the strategy requires a manual selection of biologically meaningful clusters, which are to be used as an initial condition. This process is undisciplined as well as computationally expensive. We utilize sparse NMF to cluster large protein datasets. We show how to combine the method with Fuzzy C-means algorithm and incorporate secondary structure information to increase the percentage of protein segments clusters with high structural similarity. Our experimental results show that sparse NMF approach provides better protein groupings in protein structures maintaining the similarities in protein sequences.

Contributions to Other Disciplines:

The NMF involves the nonnegativity constrained least squares (NNLS) problem with multiple right hand sides. The NNLS problems also arise in many different science and engineering areas including chemometrics and bioinformatics, and the algorithm could benefit those areas as well. The work on the L1 regularized linear regression provides a novel algorithm not only to the problem itself but also to related problems. In the literature on machine learning and statistics, the L1 regularized linear regression is used as a subroutine for many other L1 regularized learning tasks such as the logistic regression and sparse dictionary learning. Those induced tasked can also benefit from the new algorithm. Computations on tensors are becoming ubiquitous in many applications. Nonnegative tensor factorization methods have been used for face recognition and computation of human motion signatures and for hyper spectral data analysis. Projection to a low-dimensional subspace is a common technique in many areas of data mining and pattern recognition. Allowing more dimensions, the document collection can be represented in a way to reveal multi-dimensional and multiscale relationships such as those among terms, documents, authors, and genre, in the data simultaneously. Recently new methods have been developed, where one couples structural information, in the form of one or more graphs, to textual information. In the case of collaborative filtering, instead of a matrix representing ratings of users and ratings, allowing multidimensional representation with details about genre of movie, time of viewing of movie will help in better prediction of user's interest. In numerous applications, data sets are more naturally represented as tensors than matrices including nuclear astrophysics, climate modeling, chemometrics, genome signal analysis, and biometric recognition. Tensor-based methods can be utilized for data compression, modeling, and regression, fusing information obtained from different sources and scales. In particular it has been used for analysis of enzymic activity in vegetables, influence of temperature on vibrational spectra, semiconductor metal etching and other such applications in the field of

chemometrics. The NMF involves the nonnegativity constrained least squares (NNLS) problem with multiple right hand sides. The NNLS problems also arise in many different science and engineering areas including chemometrics and bioinformatics, and the algorithm could benefit those areas as well. The work on the L1 regularized linear regression provides a novel algorithm not only to the problem itself but also to related problems. In the literature on machine learning and statistics, the L1 regularized linear regression is used as a subroutine for many other L1 regularized learning tasks such as the logistic regression and sparse dictionary learning. Those induced tasked can also benefit from the new algorithm. Computations on tensors are becoming ubiquitous in many applications. Nonnegative tensor factorization methods have been used for face recognition and computation of human motion signatures and for hyper spectral data analysis. Projection to a low-dimensional subspace is a common technique in many areas of data mining and pattern recognition. Allowing more dimensions, the document collection can be represented in a way to reveal multi-dimensional and multiscale relationships such as those among the term, document, authors, and genre, in the data simultaneously. Recently new methods have been developed, where one couples structural information, in the form of one or more graphs, to textual information. In the case of collaborative filtering, instead of a matrix representing ratings of users and ratings, allowing multidimensional representation with details about genre of movie, time of viewing of movie will help in better prediction of user's interest. In numerous applications, data sets are more naturally represented as tensors than matrices including nuclear astrophysics, climate modeling, chemometrics, genome signal analysis, and biometric recognition. Tensor-based methods can be utilized for data compression, modeling, and regression, fusing information obtained from different sources and scales. In particular its used for analysis of enzymic activity in vegetables, influence of temperature on vibrational spectra, semiconductor metal etching and other such applications in the field of chemometrics.

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Conference Proceedings

Categories for which nothing is reported:

Organizational Partners
Activities and Findings: Any Training and Development
Any Web/Internet Site
Any Product
Contributions: To Any Human Resource Development
Contributions: To Any Resources for Research and Education
Contributions: To Any Beyond Science and Engineering
Any Conference