

Broad Overview

Machine learning can be described as the discipline of automatically “learning” concepts from data either with, or without human guidance. The goal of most machine learning applications is to discover patterns within data so that properties of hitherto unseen data can be predicted reliably. However, the task of learning from data can often be very challenging (naturally contingent upon the concept that one is trying to learn), more so in the presence of vast amounts of data, where most approaches rapidly become infeasible. Therefore, it is imperative to either develop algorithms that are highly scalable, or to obtain approximate representations of data that are amenable to more sophisticated techniques.

In my research, I deal with both the abovementioned aspects. My Ph.D. thesis is on *Matrix Nearness Problems in Data Mining*, wherein I develop various structured or pattern revealing approximations to the input data along with efficient algorithms for their computation. Some examples of frequently arising matrix nearness problems are dimensionality reduction, sparse approximations, nonnegative matrix approximation, clustering, co-clustering, and kernel matrix approximation. Each of these problems has numerous applications in a diverse variety of fields ranging from bioinformatics, signal processing, text analysis, to astronomy. By developing new algorithms and efficient implementations for important nearness problems, my work enables several applications across all these domains, especially in the face of large amounts of data.

In addition to my doctoral research, I have also worked on obtaining highly scalable algorithms and implementations for fundamental machine learning problems such as classification and regression. A simple example of my recent work on these basic problems is given by support vector regression; consider a dataset with several million high-dimensional vectors as input. None of the commonly available software implementations for support vector regression scale to such large data sizes, a problem that my approach overcomes.

In what follows below, I have summarized some of my Ph.D. research work, as well as my non-doctoral research, the latter indicated by a ‘★’ suffixed to the section name.

Efficient Large Scale Classification and Regression ★

Automatic classification of objects into distinct categories is the most fundamental problem of machine learning. The method of support vector machines (SVMs) has enjoyed enormous success towards this goal. However, despite the intense efforts by the research community, efficient algorithms for really large-scale problems have been lacking. What started as a diversion for me, has turned into the fastest SVM software, which outperforms all well-known SVM software, including the recent award winning SVM^{perf} software by [Joa06]. I attribute the success of my method to the philosophy of keeping things simple with an emphasis on an efficient implementation [Sra06]. Too often, one sees overly complicated approaches to solve problems where a simple and efficient approach would suffice.

Encouraged by the performance of my SVM, I have extended it to handle the problem of support vector regression too, once again obtaining perhaps the fastest implementation. I would like to remark that my work on SVMs was a byproduct of my effort to develop large-scale optimization software for solving linear and quadratic programs (this software of mine can handle problems with tens of millions of constraints with hundreds of thousands of variables, something that can be remarkably achieved on a simple desktop computer!).

I am currently collaborating with Stefanie Jegelka (Max-Planck Institute for Biological Cybernetics, Tübingen, Germany) to develop new decomposition based optimization algorithms, that once again, will not only benefit SVM implementations, but will also be applicable to general quadratic programming problems. In fact, within the next few months, I plan to convert my

SVM implementation into an academically and commercially licensable software, which should establish itself as the leading implementation for linear SV-classification and regression.

Nonnegative Matrix Approximation

Modeling data is central to numerous applications, both in machine learning and statistics. Non-negative matrix approximation (NNMA, also known as nonnegative matrix factorization) is a powerful matrix decomposition technique that helps to model data in an efficient and interpretable manner, providing an alternative to the widely employed method of Principal Components Analysis (PCA). NNMA has been applied to numerous applications, and continues to gain further interest. The potential applications are too numerous to summarize here and I refer the reader to the short survey in [SD06] for more details. Furthermore, the increasing importance of NNMA and other matrix approximation problems in data mining can be gauged by the fact that 4 days were devoted at the massive datasets workshop [MMD06] to such problems.

I have been studying the NNMA problem for the past few years, and in addition to formulating important generalizations for it, I have also developed efficient algorithms based on multiplicative updates for solving it [SD06]. The generalizations that I study consider the NNMA problems where the error of approximation is measured by some convex distortion function (such as the Frobenius norm, KL-Divergence or a Bregman Divergence). The algorithms that I have developed are generic, and many of the techniques that I have formulated therein can be carried over to the solution of other related problems too. The optimization problems that must be solved for NNMA are quite difficult on account of their non-convexity. However, this lends considerable challenges to the problem, making the quest for solutions more rewarding.

More recently I have worked with a colleague on developing theoretically sound and practically feasible algorithms for NNMA that improve upon the state of the art. However, despite our improvements, the methods and implementations thereof are still not scalable to really large-scale problems. It remains a goal of mine to develop methods that can solve matrix approximation problems such as NNMA for very large-scale data. Preliminary work towards that goal will appear in my Ph.D. dissertation.

As before, I strive to provide efficient implementations of the algorithms that I derive. To that end, I have authored a high-performance software implementation of many NNMA algorithms to complement my theoretical work.

Incremental aspect models ★

Automatically discovering and tracking topics from streaming data (e.g., news articles) is a challenging problem, which is a part of the Topic Detection and Tracking (TDT) initiative. In fact for the current generation web-related problems some form of automatic topic detection and topic based machine learning systems are becoming increasingly important (e.g., in the CALO project).

I worked with Arun Surendran at Microsoft Research (Knowledge Tools group) during the summer of 2005, and we developed a new method to discover and track topics from data streams [SS06]. The research approach that I took was to rapidly prototype several existing methods, while figuring out their deficiencies in order to simultaneously develop my own solution to the problem. We developed a theoretical base for the task and programmed a practical solution, eventually filing two patents for the associated work. This work later motivated me to develop a general theoretical framework for incrementally learning low-rank approximations, which connects directly to my dissertation research.

Work on Clustering and Co-clustering

Clustering is the central problem of data mining, wherein one aims to automatically group together similar objects. I begun my foray into the field of data mining with some work on clustering text data by modeling it using directional distributions instead of the standard multivariate

Gaussian or multinomial distributions. Directional distributions generalize the concept of cosine similarity, and just as the Euclidean k-means algorithm may be considered as a limiting case of mixture modeling using multivariate Gaussians, the cosine similarity based k-means algorithm (spherical k-means) may be considered as the limiting case of mixture modeling using multivariate von Mises-Fisher distributions (which are the *natural* distributions for data distributed on the unit hypersphere).

My work on modeling data using directional distributions led to a successful collaboration resulting in two papers [BDGS03, BDGS05]. Recently, I have been looking at extending this work to richer probabilistic models that promise to yield improved experimental results and provide more extensive data modeling ability. Furthermore, the spectrum of mathematical techniques required is very diverse and elegant, ranging from asymptotic analysis for the solution of difficult non-linear equations to the basic mixture modeling methods like Expectation Maximization (EM). The clustering procedures that we have developed based on directional distributions seem particularly promising for text and biological data [BDGS05].

In addition to working on clustering, I have also worked on co-clustering [CDGS04], wherein one essentially aims to group together not only the individual data points, but also their features. Co-clustering has found application in numerous domains such as language modeling, gene expression data analysis, recommender systems, among others, and continues to increase in importance as a standard data analysis technique. My work on nonnegative matrix approximations has a direct connection to both clustering and co-clustering, since one can obtain relaxed solutions to both problems by an appropriate use of NNMA, highlighting the fact that clustering, co-clustering, and NNMA are all matrix nearness problems.

Sparse Signal Approximation *

In a collaboration with Joel Tropp (Univ. Mich. Ann Arbor), I have also worked on the problem of reconstructing a sparse signal given a suitable number of random measurements. Our work appeared in [ST06], wherein my main contribution was the design and implementation of an efficient algorithm to solve the therein studied sparse approximation problem. The problem of sparse signal reconstruction is of considerable importance, and it continues to be actively pursued by both the signal processing and machine learning communities. As a part of my future research, I plan to devote some of my efforts towards investigating important applications and algorithmic challenges in this area.

Other Matrix Approximations and Optimization

Beyond the problems mentioned above, in my Ph.D. thesis, I also study a variety of other matrix approximation problems, especially with the goal of obtaining efficient and scalable algorithms for their solution. The most notable amongst these are the problems of low-rank approximations, where one seeks to approximate an input matrix A by the product BC , where both B and C are of low-rank and may additionally be required to satisfy some constraints. The underlying application determines which objective function one uses to measure the error of approximation, leading to a different optimization problem in each case. In my thesis I develop efficient generic methods for particularly important special cases (ℓ_1 -norm based sparse approximations, KL-Divergence or Frobenius norm based problems, etc.) of such approximation problems by drawing upon methods from convex optimization and analysis.

Another interesting problem that I have studied is called *Metric Nearness*, wherein given a set of interpoint dissimilarity measurements, one aims to construct a set of interpoint distances that satisfy the triangle-inequality. This problem has interesting connections to the well-known all pairs shortest paths (APSP) problem from graph theory. It also has important applications to metric based indexing in biological databases, in addition to immediately yielding approximation algorithms for combinatorial problems that are more easily solved for metric data.

Software Developed

As a part of my research efforts, I have also developed efficient software that implements many of the algorithms that I have derived. These software packages are summarized below.

1. **SSLIB**. Sparse matrix manipulation library written from scratch in C++. It includes various formats of sparse matrix storage and forms the backbone of all my software that uses sparse matrices. Additionally, by interfacing with ARPACK via my wrapper library, SSLIB offers ability to compute eigenvectors and eigenvalues for large sparse matrices.
2. **NNMA**. Fast, high-performance implementation of various NNMA algorithms built on LAPACK, BLAS, and SSLIB. It is also written in C++ and is easily extensible to include new NNMA algorithms.
3. **SSVD**. My wrapper library (C++) around SVDPACKC. This permits the easy inclusion of sparse matrix singular vector computation into any C++ code, and is available from Netlib.
4. **FSOLVER**. My implementation of large scale linear and quadratic programming, including specialized codes for non-negative least squares (NNLS), regularized least squares, ℓ_1 -norm minimization, and iterative linear system solving.
5. **BLITZ^{SVM}**. My highly scalable implementation of ℓ_1 - and ℓ_2 -norm Support Vector Machines (SVMs) for classification and regression. This implementation also exploits SSLIB for its sparse matrix computations, and to my knowledge, is the fastest SVM implementation available.
6. **EMENGINE**. An easily extensible C++ code that can be used to implement the EM algorithm for arbitrary probability distributions. Naturally, the user needs to override the default methods to obtain a fine tuned version for a particular distribution. This software is under construction, and currently it includes routines for directional distributions (which require significant engineering due to numerical difficulties).
7. **Misc**. In addition to the above software packages or libraries, I have written many other smaller routines or programs (e.g., regression routines for extremely large scale data, KNN code for large-data etc.). I have also contributed to some of the clustering software developed in our group.

Research Vision

I plan to continue working on fundamental problems in machine learning and data mining, with the aim of developing theoretically sound and highly scalable algorithms and implementations. My interests within the field are broad ranging because I enjoy working on almost all types of machine learning problems. However, some specifically interesting problems/domains to me are: (i) problems in bioinformatics, especially those related to clustering, pathway analysis, gene and protein interaction network analysis, and other graph based algorithms and models for gene and protein data, (ii) dimensionality reduction, both linear and non-linear methods, especially robust alternatives to Principal Components Analysis (PCA) with applicability to massive data sets, (iii) large-scale transductive (semi-supervised) SVMs and transductive regression problems, (iv) sparse PCA, sparse matrix approximations and data modeling for both matrices and tensors, (v) Independent Components Analysis (ICA), sparse and nonnegative ICA, (vi) regularized approximations of fundamental problems such as least-squares, nonnegative least squares, generalized linear models, ℓ_1 -norm regularization, etc., (vii) large-scale optimization methods for important problems in machine learning, such as maximum entropy based learning, information-theoretic methods, non-linear models with kernels, etc., (viii) semidefinite programming and second order

cone programming, both from an optimization as well as a machine learning perspective, (ix) auxiliary function techniques such as EM, CCCP, variational methods, etc., (x) learning metrics and kernels from examples for classification or clustering, (xi) Gaussian processes, (xii) graphical Models, Markov Processes, and Bayesian methods, (xiii) online learning, (xiv) topic discovery and tracking, (xv) search, query-processing, and information retrieval problems, and (xvi) most importantly, applications to various types of data such as text, biological, social networks, graphs, or images, to name a few.

I have experience with some of the abovementioned problems and topics, but naturally not with all—preliminary work related to some of them will appear in my dissertation, and I am collaborating with colleagues on some others.

My broad vision is to continuously expand my knowledge and expertise in many of the problems/domains mentioned above, with the underlying goal of producing world-class research that is of importance not only to the organization where I work, but to the broader research community. Furthermore, by not restricting my problem domains to mere toy or concept level datasets, I aim to produce efficient and scalable software implementations for many of the problems that I study. My vision is to thus generate both research and software that brings us not only academic, but also industrial respect and numerous technology advances via patents.

I foresee myself working with an active team of researchers on several real-life data mining problems and applications (user modeling, social networks, targeted advertising and marketing, search enhancements, etc., to name a few). In a few years time, I would like to be at the helm of a group of highly motivated researchers and engineers with the aim of constructing actual deployable systems (after helping technology transfers to the software engineers) that will have a much wider impact than research alone can bring.

I would like to conclude by stating that I look forward to an exciting and challenging research and development career.

References

- [BDGS03] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Generative model-based clustering of directional data. In *ACM SIGKDD*, pages 19–28, 2003.
- [BDGS05] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *JMLR*, 6:1345–1382, Sep 2005.
- [CDGS04] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum Sum Squared Residue based Co-clustering of Gene Expression data. In *Proc. 4th SIAM International Conference on Data Mining (SDM)*, pages 114–125, Florida, 2004. SIAM.
- [Joa06] T. Joachims. Linear SVMs in linear time. In *SIGKDD*, 2006.
- [MMD06] MMDS 2006. Workshop on Algorithms for Modern Massive Data Sets. URL: <http://www.stanford.edu/group/mmds/>, June 2006.
- [SD06] S. Sra and I. S. Dhillon. Multiplicative Update Algorithms for NNMA with Generalized Loss Functions. *Submitted*, November 2006.
- [Sra06] S. Sra. Efficient large scale linear programming support vector machines. In *ECML*, pages 767–774, September 2006.
- [SS06] A. Surendran and S. Sra. Incremental Aspect Models for Mining Document Streams. In *PKDD*, pages 633–640, September 2006.
- [ST06] S. Sra and J. A. Tropp. Row-action methods for compressed sensing. In *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2006.