

PROFILE

Skilled research scientist with a doctorate in computer science. Strong background in **machine learning, clustering, statistics, and social networks**. Experienced in independent and group projects, both in research and implementation.

EDUCATION

Doctorate of Philosophy, Computer Science August 2010

Rensselaer Polytechnic Institute

Dissertation topic: Theory and application of an improved covariance matrix approximation.

Applications to accent identification and clustering in social networks.

Master of Science, Computer Science 2005

Rensselaer Polytechnic Institute

Thesis topic: Application of fitting parameterized distributions to data of stars in Milky Way galaxy.

Bachelor of Engineering, Dual degree 2003

Computer & Systems Engineering and Computer Science

Rensselaer Polytechnic Institute GPA 3.6

PUBLICATIONS

Jon Purnell, Malik Magdon-Ismail, and Heidi Jo Newberg "A Probabilistic Approach to Finding Geometric Objects in Spatial Datasets of the Milky Way", ISMIS, 2005 (Master's Thesis)

Jon Purnell, Malik Magdon-Ismail "Learning American English Accents using Ensemble Learning with GMMs", ICMLA, 2009

Jon Purnell, Malik Magdon-Ismail "Approximating the Covariance Matrix with Low-rank Perturbations", IDEAL, 2010

Jon Purnell, Malik Magdon-Ismail "Approximating the Covariance Matrix of GMMs with Low-rank Perturbations", IJDM, 2011 (accepted)

EXPERIENCE

Post-doctoral Research October 2010 – October 2011

Computer Science, RPI

Developing open source code based on thesis research, assembling additional results for journal papers, training graduate student. Collaborating with Army, BBN Tech under SCNARC

Teaching Assistant September 2006 – May 2008

Computer Science, RPI

Assisted professor in grading, supervising lab hours, and office hours to help students understand the material. Taught Computer Science II, Data Structures & Algorithms, and Programming Languages.

Research Assistant May 2008 - Present

Computer Science, RPI

Researched problem through current literature, developed algorithms, conducted experiments and gathered results under the supervision of Prof. Magdon-Ismail.

Research Consultant October 2007 - March 2008

Centennial Group

Provided a research report to compliment the company's overall project in e-commerce security.

- Researched the field of e-commerce security: dangers and problems of e-commerce, successes and pitfalls of current security technologies, emerging security technologies.
- Created a metric to compare security systems against each other taking into account such factors as security strength, infrastructure maintenance, and end-user needs.
- Communicated with management on goals of the project, feedback on preliminary reports, and fine-tuning the final report.

Programming Consultant March 2007 - Present

American Heart Association

Designed and implemented software to convert medical records between hospital and AHA formats.

- Communicated with AHA directors on project goals and terminologies.
 - Programmed application's functionality to convert records and ensure proper execution.
 - Developed and implemented interface, based on feedback from testers and users.
-

RELEVANT SKILLS

Programming Languages: C/C++, Perl, Matlab, MFC, Unix shell
Additional Coursework: Discrete Time Systems, Stochastic Processes, Statistics

HONORS

Scholarships/Awards

- Rensselaer Alumni Scholarship (September 1999 – May 2003)
 - Philip H. Parthesius Fellowship (September 2005)
-

EXTRACURRICULAR

Leadership Roles

- Elder of Third Presbyterian Church of Troy and Committee Chairperson
- Leadership Team and Small Group Leader for Graduate Christian Fellowship at RPI
- President and VP of Finance for Sigma Phi Epsilon, New York Delta chapter at RPI

Community Service

- Capital City Rescue Mission in Albany, NY
 - American Red Cross Blood Donor
 - Eagle Scout Project in Smithtown, NY (1998)
-

REFERENCES

Malik Magdon-Ismail *Advisor at RPI*
Room 207 Lally, CS Department, RPI,
110 8th Street, Troy, NY 12180
Work: 518-276-4857
Email: Magdon@cs.rpi.edu
Webpage: <http://www.cs.rpi.edu/~magdon/>

Mukkai S. Krishnamoorthy *Prof. of DSA while I was a TA*
Associate Professor
Room 207 Lally, CS Department, RPI,
110 8th Street, Troy, NY 12180
Work: 518-276-6911
Email: moorthy@cs.rpi.edu
Webpage: <http://www.cs.rpi.edu/~moorthy/>

M.V. Muhsin *Contact at Centennial Group*
Director & COO
Africa & Middle East
Centennial Group
The Watergate Office Building
2600 Virginia Ave, NW, Suite 201
Washington, D.C. 20037
Work: 1-301-537-8350
Email: mohamed@mvmuhsin.com
Webpage: <http://www.centennial-group.com>

Christine Rutan *Contact at AHA*
Senior Director, Quality Improvement Initiatives
American Heart Association
Founders Affiliate
440 New Karner
Albany, NY 12205
Work: 518-869-4048
Email: Christine.Rutan@heart.org
Webpage: <http://www.americanheart.org>

Research Statement

Executive Summary: I specialize in clustering and its applications to large and high-dimension relational datasets. I have and continue to develop an approach that not only allows for overlapping clustering but also runs in linear time with respect to the number of samples and the dimensions of the data. I have applied this approach to speech processing, for the purpose of accent identification, and social networks, for the purpose of identifying communities. I plan on continuing to work on my clustering algorithm and its application to social networks.

I began my PhD with a desire to study machine learning. At the time I was intrigued by speech processing and recognition. After surveying the literature, I found that there was a growing need for unsupervised algorithms in accent identification. The modeling of accents required a larger amount of speech than for speaker identification, so the task of transcribing the speech for use by Hidden Markov Models became overwhelming. I investigated the use of Gaussian Mixture Models (GMMs) as a way to model processed speech data without using any transcripts. I discovered that this was both a difficult problem, in that both computers and humans cannot consistently identify accents correctly, and a promising problem, in that our approaches suffered a small penalty in accuracy without the *a priori* knowledge of transcripts.

During our work with accent identification I came across problems which lead me to the core of my thesis. The speech data I worked with had 39 dimensions and a few hundred thousand samples. Calculating the covariance matrices for the GMM was time consuming, but suffered greatly in accuracy when we approximated them by their diagonal. It was while optimizing the GMM code that a property of the covariance matrices was discovered that could be exploited. After working through the theory and practical proofs of concept, I found an iterative method that would allow the covariance matrix to be approximated by a rank-1 perturbation of a diagonal matrix. In addition, the approximation could be done in linear time with respect to the dimension of the data, which was an improvement over calculating the full covariance matrix in quadratic time. Although the approach was developed for the GMM, it remains an applicable approximation to any sample-based covariance matrix.

With this improved GMM in hand, I began to look for an interesting application. My advisor suggested that we apply this model to clustering in social networking. Working on the DBLP, a database of publications in Computer Science, we sought to identify communities of authors given their past collaborations on published papers. A graph was constructed using this information and then embedded into Euclidean space using Sampled Spectral Distance Embedding (SSDE), which provides a mapping of the graph vertexes that preserves the dis-

tance between nodes in the graph by their distance in Euclidean space. The embedded data was then clustered using the improved GMM. In addition to identifying different subtopics of computer science (confirmed by observing paper title keywords and conference venues), I had an approach that could handle social networks on the order of one million nodes. Further, we were able to do this in less time than other existing overlapping clustering algorithms.

There is still more work to be done on this approach. The GMM only provides the probability that a paper came from a mixture (aka. Cluster), but not an assignment. A heuristic was developed to determine to which clusters a paper should be assigned. Initially, papers were assigned to the mixture they had the greatest probability of being within. This was then extended by assigning a paper to a second cluster if the corresponding probability was relatively close to the greatest probability. Papers were added to additional clusters based upon the ratio of the corresponding probability to the paper's greatest probability. This was continued until we reached a predetermined overall average of clusters a paper was assigned to.

Current Research I am exploring other heuristics for identifying overlapping clusters using the GMM. The threshold we used for adding papers to additional clusters was simple and good for a tentative measure, but a new threshold is needed that can be applied to any dataset. We are using various social network properties (e.g. power-law, small-world, betweenness) to evaluate the benefit of assigning papers to additional clusters. This helps us to evaluate our various approaches to identifying overlapping clusters.

I am also exploring improvements in other aspects of our approach. Both the initialization of the GMM parameters and finding the optimal number of clusters are important to performance. By applying some approaches from other clustering algorithms, I anticipate improving the robustness and accuracy of the GMM. Further, the linear time complexity of the GMM allows us to increase the dimensionality of the space into which the SSDE embeds the graph. With high dimensional data, we retain more of the initial information and can therefore improve the quality of clusters.

Future Research I plan to continue my research in clustering algorithms, particularly for the applications of social networks. The main thrust would be in developing improved approaches to finding overlapping clusters/communities. In addition to metrics for defining a community, there would be a benefit in defining metrics and properties of agents/nodes/people in the overlap between clusters. The field has yet to define an optimal definition for communities, so there is a great opportunity to explore and innovate. As social networking and clustering improves, so will the related fields of biology, sociology, and marketing.