

KAUSTAV DAS

Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Phone: (412) 736-6431
Email: kaustav@cs.cmu.edu
Web: <http://www.cs.cmu.edu/~kaustav>

RESEARCH INTERESTS

- ◇ Machine Learning and its applications to real-world problems.
- ◇ Unsupervised Anomaly and Pattern detection in large datasets.
- ◇ Developing efficient algorithms for mining large sets of data by combining various statistical and machine learning methods

EDUCATION

- ◇ **Ph.D. in Machine Learning** August 2003 to present
CARNEGIE MELLON UNIVERSITY
Thesis topic: *Anomaly Detection in Large Categorical Datasets*
Advisor: Prof. Jeff Schneider (GPA: 4.0/4.0)
- ◇ **M.S. in Knowledge Discovery and Data Mining** August 2003 to December 2006
CARNEGIE MELLON UNIVERSITY
(GPA: 4.0/4.0)
- ◇ **B.Tech. in Computer Science and Engineering** July 1999 to April 2003
INDIAN INSTITUTE OF TECHNOLOGY (IIT), KHARAGPUR
(GPA: 9.49/10)

ACADEMIC EXPERIENCE

- ◇ **At Carnegie Mellon University** August 2003 to present
 - ◇ **Detection of Anomalies in Categorical Datasets**
We consider the problem of detecting anomalies in high arity categorical datasets. We are interested in the problem of unsupervised anomaly detection, where we use the unlabelled data for training, and detect records that do not match the training distribution well. We present a novel definition of anomalies, and propose an approach of comparing against marginal distributions of attribute subsets [3]. We have also investigated the detection of groups of anomalies that are generated by some process and have certain similarities between them in [1] and [2].
 - ◇ **Detection and Visualization of Anomalies in Multivariate Time-series Data**
I am investigating a useful way to search through all the possible combinations (additive and multiplicative) of the multiple variables. The objective is to identify a visually striking anomaly in the dataset which is most significant in one of the composite time series.
 - ◇ **Modeling Oil Fields**
This work is in collaboration with British Petroleum. The objective is to develop a predictive model for their oil fields based on time series sensor data from the wells. We investigated potential oil well connectivity from the data using auto-regressive and hidden markov models.
 - ◇ **Alarming in Surveillance Systems**
Surveillance systems are being increasingly used in a variety of fields including public health monitoring and anti-terrorist applications. Typically such systems are required to signal an alarm when some anomaly is detected. The objective is to optimize with respect to various cost functions, eg. the false-

positive and false negative costs. We formulated a POMDP based approach to solve the problem optimally. It performs better than commonly used approaches such as CUSUM [5]. We have also investigated the use of Discriminative Random Fields for detecting disease clusters in bio-surveillance [4].

◇ **Senior Thesis on “Efficient Multiple Gene Sequence Alignment using A^* Search” at IIT Kharagpur** July 2002 to April 2003

We also created a reasoning framework for efficiently extracting information regarding the possibility of synthesizing a specified sequence from a genome database, along with the restriction enzymes and DNA Ligase that may be required for its synthesis [6].

◇ **Unsupervised Image Segmentation and Labeling**

Development and implementation of an unsupervised image segmentation and labeling method under the guidance of Prof. Anjan Sarkar, IIT Kharagpur. My implementation was used to obtain the experimental results mentioned in the paper “A MRF Model-Based Segmentation Approach to Classification for Multispectral Imagery” published in IEEE Transactions on Geo-Science and Remote Sensing, Vol 40, No 5, May 2002.

PROFESSIONAL EXPERIENCE

◇ **Summer Internship at Microsoft Research Lab, Seattle** May-July 2005

Worked on a user interface tool to automate similar repetitive graphical interface tasks across different applications.

◇ **Summer Internship at Microsoft India Development Centre, Hyderabad** May-July 2002

Worked on Interix (Unix environment subsystem in Windows) process kernels to emulate certain posix commands.

AWARDS AND HONORS

- ◇ Ranked 100th in the entrance examination to Indian Institutes of Technology (IITs). Over 100,000 students participated.
- ◇ Selected to write the Indian National Mathematical Olympiad 1998.

PUBLICATIONS

1. Anomaly Pattern Detection in Categorical Datasets.
To appear in the Proc. of the 14th ACM intl. conf. on Knowledge Discovery and Data Mining (KDD 2008).
Kaustav Das, Jeff Schneider and Daniel Neill.
2. Detecting Anomalous Groups in Categorical Datasets.
Submitted to Advances in Neural Information Processing System (NIPS 2008).
Kaustav Das, Jeff Schneider and Daniel Neill.
3. Detecting Anomalous Records in Categorical Datasets.
In Proc. of the 13th ACM intl. conf. on Knowledge Discovery and Data Mining (KDD 2007).
Kaustav Das and Jeff Schneider
4. Disease Outbreak Detection using Discriminative Random Field.
In Proc. Fifth Annual International Society for Disease Surveillance Conference (ISDS 2006).
Kaustav Das, Robin Sabhnani and Eric Xing
5. Belief State Approaches to Signaling Alarms in Surveillance Systems.
In Proc. of the 10th ACM intl. conf. on Knowledge Discovery and Data Mining (KDD 2004).
Kaustav Das, Andrew Moore and Jeff Schneider

6. An Algorithm for extraction of DNA fragments using restriction Enzymes.
In Sixth International Conference on Information Technology (CIT 2003).
Kaustav Das, Pallab Dasgupta and P. P. Chakrabarti.

TEACHING EXPERIENCE

- ◇ Teaching assistant for Machine Learning (15-781), Spring 2005. Involved a weekly recitation, setting homeworks and exams, grading.
- ◇ Teaching assistant for Multimedia Databases (15-826), Spring 2006. Involved setting homeworks and exams, grading.

PROFESSIONAL SERVICE

- ◇ Reviewed papers for ICML 2007 and the Machine Learning Journal

RELEVANT GRADUATE LEVEL COURSES

Machine Learning, Artificial Intelligence, Intermediate Statistics, Statistical Foundations of Machine Learning, Probabilistic Graphical Models, Randomized Algorithms, Time Series Analysis, Graduate Algorithms, Machine Learning Theory, Multimedia Databases and Data-mining, Discrete Multivariate Analysis

SYSTEMS EXPERIENCE

- ◇ Languages: C++ with STL, C.
- ◇ Languages I use occasionally: Java, Perl, Python, HTML, L^AT_EX.
- ◇ Used runtime profiler (`gprof`) and debugger (`gdb`).

REFERENCES Available on request.