

CSI 661 – Data Mining

Concentration Areas and Papers

Applications to Unusual Data: Streaming Data

These papers look at applying the core algorithms learnt so far to the situation where the notion of a “data-set” does not exist, only a stream of data that cannot be stored.

Prediction

Street, Kim, A streaming ensemble algorithm (SEA) for large-scale classification, Knowledge Discovery and Data Mining Conference, 2001.

Available from: <http://portal.acm.org/citation.cfm?id=502568&dl=ACM&coll=portal>

P. Domingos and G. Hulten. Mining High-Speed Data Streams. In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining, pages 71--80, 2000.

Available from: <http://citeseer.nj.nec.com/domingos00mining.html>

Clustering

Barbara, D., Requirements for Clustering Data Streams, SIGKDD Explorations

Available from:

<http://www.acm.org/sigs/sigkdd/explorations/issue3-2/barbara.pdf>

Guha, Mishra, Motwani et al, Clustering Data Streams, FOCS.

Available from:

<http://citeseer.nj.nec.com/cache/papers/cs/17885/http://zSzzSztheory.stanford.eduzSz~sudiptoZSzmypperszSzdatastream.pdf/guha00clustering.pdf>

Combining Different Algorithms To Predict Complicated Structure: Bioinformatics

So far we have covered prediction in a simple context: TRUE, FALSE, POSTIVE, NEGATIVE. However, many interesting problems involve prediction in more complicated contexts. These papers introduce the problem prediction of what points a string will contact itself when it folds. This concentration deals with applying the data mining algorithms covered so far in **combination** to achieve the same aim.

G. Pollastri and P. Baldi ,Prediction of Contact Maps by Recurrent Neural Network Architectures and Hidden Context Propagation From All Four Cardinal Corners

Available from: <http://www.cs.albany.edu/~berg/lab/PollastriBaldi02.pdf>

Fariselli et al, Prediction of Contact Maps with Neural Networks and Correlated Mutations

Available from: <http://www.cs.albany.edu/~berg/lab/FariselliEtAl01.pdf>

Fast Clustering By Pre-Processing Data

So far we have predominantly covered the situation where we do not pre-process the data. If we pre-process the data into an efficient storage structure, such as a kd-tree, we can obtain remarkable algorithm speed-ups.

Pelleg and A. Moore. Accelerating exact k-means algorithms with geometric reasoning. In Knowledge Discovery and Data Mining, pages 277--281, 1999.

Available from:

http://www-dev.ri.cmu.edu:8080/pub_files/pub1/pelleg_dan_1999_1/pelleg_dan_1999_1.pdf

An Efficient k-Means Clustering Algorithm: Analysis and Implementation, Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu

Available from: <http://www.computer.org/tpami/tp2002/i0881abs.htm>

Sequence Mining

Ayres et al Sequential Mining Using Bitmaps

<http://www.cs.cornell.edu/johannes/publications.html>

Association Rules, Bayesian Probability and Belief Networks

Empirical Bayes screening for multi-item associations [ps]

DuMouchel W, Pregibon D (2001) Proc. KDD 2001, ACM Press, San Diego, CA, Best Paper Award KDD 2001.

<http://www.research.att.com/~dumouchel/papers/KDD2001.pdf>

Predicting and An Unknown Future: Steganalysis

Fridrich, et al, *Practical Steganography*

Available from: <http://citeseer.nj.nec.com/537736.html>

Data Quality

Heckerman, Microsoft Corporation, *Learning With Belief Networks*,

Available from: <http://citeseer.nj.nec.com/heckerman96tutorial.html>

Tayi, Ballou, Examining Data Quality, CACM, 1998 Feb, Vol 2.

Available from:

<http://portal.acm.org/citation.cfm?id=269012.269021&coll=portal&dl=ACM&idx=J79&part=magazine&WantType=magazine&title=Communications%20of%20the%20ACM&CFID=8481079&CFOKEN=63390935>

Ballou, Tayi, Methodology for Allocation Resources for Data Quality, CACM, 1989 March,

Available from:

<http://portal.acm.org/citation.cfm?doid=62065.62068>

Practical Applications of Anomaly Detection

[J. Dale Kirkland](#), Ted E. Senator, [James J. Hayden](#), [Tomasz Dybala](#), [Henry G. Goldberg](#), [Ping Shyr](#):
THE NASD Regulation Advanced-Detection System (ADS). [AI Magazine 20](#)(1): 55-67 (1999)

Or

[J. Dale Kirkland](#), Ted E. Senator, [James J. Hayden](#), [Tom Dybala](#), [Henry G. Goldberg](#), [Ping Shyr](#):
The NASD Regulation Advanced Detection System (ADS). [AAAI/IAAI 1998](#): 1055-1062

And

Ted E. Senator, [Henry G. Goldberg](#), [Jerry Wooton](#), [Matthew A. Cottini](#), [A. F. Umar Khan](#), [Christina D. Klinger](#), [Winston M. Llamas](#), [Michael P. Marrone](#), [Raphael W. H. Wong](#): The Financial Crimes Enforcement Network AI System (FAIS) Identifying Potential Money Laundering from Reports of Large Cash Transactions. [AI Magazine 16](#)(4): 21-39 (1995)

Visualizing Clustering Results

Davidson, I., "Visualizing Clustering Results", SIAM International Conference on Data Mining, 2002 [Color PDF](#)

Yin, K. and Davidson I., Visually Comparing Clustering Algorithms, Accepted for publication at 8th PAKDD 2004