

Daniel D. McCracken

Computational Sciences

Seminar Series

Wednesday 3:00 p.m., July 7, 2004
Black Hall, Room 152

Professor B. Mirkin

School of Computer Science and
Information Systems,
University of London, UK

Clustering for Data Mining: A Data Recovery Approach

Abstract: Data mining is a paradigm for finding patterns in databases, which typically cannot be considered as samples of independent observations so that traditional statistics modelling is not applicable. This, some say, deprives data mining from theoretical foundations and restricts computer scientists to methods for only finding patterns, while leaving most important questions of data pre-processing and result interpreting to the user. In the talk, I present the data recovery approach of statistics as applied to clustering, a discipline devoted to finding homogeneous groups of entities, and restricted to available data only. The corresponding Pythagorean decompositions of data scatter lead to a number of conclusions applied, in the talk, to K-Means clustering, a most popular clustering method. In particular, model-based advice will be given on such issues of current interest as (a) simultaneous clustering of quantitative and categorical data, (b) intelligent version of K-Means automatically determining the number of clusters and initial seeds, (c) a bunch of simple but original interpretation aids.

Prof. B. Mirkin specialises in mathematical models, computational algorithms and programs for visualisation and clustering of data in sociology, marketing research, molecular biology, ecology, and other applications. He published a number of papers and monographs on these subjects, including "Graphs and Genes" (Springer-Verlag, 1984, jointly with S. Rodin) and "Mathematical Classification and Clustering" (Kluwer Academic Publishers, 1996). He obtained his PhD in Computer Science (1966) and DSc in Systems Analysis and Statistics (1990) from Russian (then Soviet) Academy of Sciences. His latest work includes methods for mapping gene histories to evolutionary trees, exploring herpesvirus protein function, analysis of the structure of bribery through newspapers articles, comprehensive conjunctive learning, and modelling waiting time in British hospitals, etc.

Supported by: CWU Faculty Senate Dev't & Appropriation Committee
For more information, please contact Dr. Kovalerchuk 963-1438