



DEPARTMENT OF COMPUTING
THE HONG KONG
POLYTECHNIC UNIVERSITY



IEEE HONG KONG SECTION
COMPUTER CHAPTER

Pattern Clustering and Data Grouping

Andrew K.C. Wong
Pattern Intelligence Inc.
and

Systems Design Engineering, University of Waterloo
akcwong@pami.uwaterloo.ca

Date: 11 December 2006 (Monday)
Time: 2:30 p.m. – 3:30 p.m.
Venue: PQ703, Department of Computing,
The Hong Kong Polytechnic University

Abstract

A basic task of machine learning and data mining is to automatically uncover patterns that reflect regularities in a data set. Today, data contents become more complex and diverse. They usually contain both nominal and ordinal data. Interesting information and relevant patterns might be scattered, entangled and spanning in various data subspaces. In the past, we have developed an algorithm known as *Pattern Discovery* to discover statistically significant patterns effectively wherever they are in the database. Recently, we have developed a new method known as *Pattern Clustering and Data Grouping* (PCDG) which is able to cluster similar patterns into *pattern clusters* while grouping pattern-induced data into *data groups*. In that sense data associated with similar statistical patterns or rules are automatically organized into groups. Then the probabilistic characteristics of patterns in each data group and the relationship among them can be revealed.

PCDG takes a data set and the association patterns discovered by either pattern discovery or association rule mining as inputs. It induces a data group located in the data set for each association pattern. Using a similarity measure between data groups, it applies a clustering algorithm to simultaneously cluster similar association patterns and group the respective pattern-induced data into data groups. Data groups induced usually span different subspaces and could overlap each other if they contain common parts of the discovered patterns. Clustering terminates based on a stopping criterion. After pattern clustering and data grouping, the probabilistic characteristics of each data groups can be explicitly displayed. PGDG can further discover and represents the structural relations among data groups in the form of attributed hypergraphs, revealing the similarity and differences among data groups. Hence, knowledge trapped in the data can be unveiled and organized for interpretation and understanding. In the talk, supporting experiments on synthetic and real-world data, including automatic discovery of gene expression subgroups from microarray data of cancerous tissues and DNA splicing classes in DNA sequences will be reported.

About the Speaker

Dr. Wong is a Distinguished Professor Emeritus (Systems Design Engineering) at the University of Waterloo where he is also an Adjunct Professor of the School of Computer Sciences and the Electrical and Computer Engineering Department. He was the Founding Director of the renowned Pattern Analysis and Machine Intelligence Laboratory (PAMI Lab) at the University of Waterloo and a Distinguished Chair Professor at the Hong Kong Polytechnic University (00-03).

Dr. Wong is a co-founder, and retired director (93-03) of Virtek Vision International Corporation. Virtek is a publicly traded company and is a leader in laser vision technology. He was president of the organization from 1986 to 1993 and later served as Chairman from 1993 to 1997. In 1997, he co-founded Pattern Discovery Software Systems Ltd. and has served as Chairman ever since. This year, he founded Pattern Intelligence, a technology holding company with its technologies supporting several high tech companies.

Dr. Wong holds a Ph.D. from Carnegie Mellon University; and a B.Sc (Hons) and M.Sc. from the Hong Kong University. He is an IEEE Fellow (for his contribution in machine intelligence, computer vision, and intelligent robotics). He is the FCCP Winner of the 1991 Award of Merit.

Enquiries:

In case of any questions, please contact Dr. George Baciu (Phone: 2766-7272, Email: csgeorge@comp.polyu.edu.hk).