DIMACS
Center for Discrete Mathematics & Theoretical Computer Science

DIMACS EDUCATIONAL MODULE SERIES

MODULE 03-7
A Gentle Introduction to Mathematical Cluster Analysis
September, 2003

Harel Barzilai
Department of Mathematics and Computer Science
Salisbury University
Salisbury MD 21801
phone: 410-543-6472
email: hxbarzilai@salisbury.edu
harel@barzilai.org

Alexander Kheyfits
Department of Mathematics and Computer Science
Bronx Community College (CUNY)
University Avenue at W. 181st Street
Bronx, NY 10453
phone: 718-289-5616
email: alexander.kheyfits@bcc.cuny.edu

Kathy Andrews
Hillsdale College
Department of Mathematics
33 E. College
Hillsdale, MI 49242
email: kjandrews42@comcast.net

DIMACS Center, CoRE Bldg., Rutgers University, 96 Frelinghuysen Road, Piscataway, NJ 08854-8018
TEL: 732-445-5928 • FAX: 732-445-5932 • EMAIL: `center@dimacs.rutgers.edu`
Web: `http://dimacs.rutgers.edu/`

Founded as a National Science Foundation Science and Technology Center and a Joint Project of Rutgers University, Princeton University, AT&T Labs - Research, Bell Labs, NEC Laboratories America and Telcordia Technologies with affiliated members Avaya Labs, HP Labs, IBM Research, Microsoft Research.

Module Description Information

Title:

A Gentle Introduction to Mathematical Cluster Analysis

Author(s):

Harel Barzilai, Department of Mathematics and Ccomputer Science, Salisbury University

Alexander Kheyfits, Department of Mathematics and Computer Science, Bronx Community College (CUNY)

Kathy Andrews, Department of Mathematics, Hillsdale College

Abstract:

Mathematical Clustering or Cluster Analysis is a field which endeavors to "classify". More precisely, given a discrete data set, this set is classified into groupings, or clusters. The criteria for what makes a "good" set of such clusters vary, however in general we want similar data points to be made part of the same cluster, and data points assigned to different clusters should have some important features distinguishing them.

The basics notions and definitions of clustering are introduced, as well as key algorithms used in the field, including the notion of a "hierarchical" algorithm. The first chapter provides an overview of the field as well as an introduction to "Agglomerative" hierarchical algorithms; the second chapter describes "Divisive" hierarchical algorithms; and the third chapter introduces Sequential clustering algorithms.

A section of print References is provided for further exploration, as well as as a section of Web References Cited (with full URLs), and additionally an online page, "Online Resources for Further Exploration" is available from the DIMACS page with sections on Tutorials, Sample Applications, Researchers with Expertise, and Software.

Informal Description:

The types of algorithms surveyed in this module are listed in the Abstract.

Prospective readers may be interested to note that the wide variety of areas of application of this field include: business planning as well as public, city and regional planning such as choosing the locations for hospitals; taxonomic classification of species; classification of chemical molecules into types; and even web pages classified together as being of similar relevance to a search, or of similar interest to a reader.

Students who complete this module will have been introduced to the fundamentals of clustering through an informal and intuitive approach which takes the reader gradually into the full formal definitions and algorithms. They will also glimpse some of the applications of this area, and will be provided with ample references for further study of both applications, and the mathematical field of cluster analysis itself.

Target Audience:

This module is meant to be an "entry level" introduction for freshmen and sophomore students. It can be used in Finite Mathematics, Discrete Mathematics and Applied Mathematics courses, as well as in "Survey of Mathematics" courses for strongly motivated liberal arts students. It can also serve for guided self-study by mathematically inclined high school students. The module may also be of interest in beginning computer science courses, particularly courses with sections which focus on algorithms

Prerequisites:

The formal prerequisites include only high school mathematics through algebra and functions or precalculus. Graph theory is not assumed, and all but he most basic set-theoretic notation is defined within the module itself.

Mathematical Field:

Cluster analysis; Classification theory; Graph theory

Applications Areas:

Classification theory; Biological, Chemical, and Ecological taxonomy and classification; Data Mining; Internet analysis and search engine algorithms.

Mathematics Subject Classification:

Primary 91C20

Secondary 05C90, 90B80, 92B10

Contact Information:

Harel Barzilai
Department of Mathematics and Computer Science
Salisbury University
Salisbury MD 21801
410-543-6472
hxbarzilai@salisbury.edu, harel@barzilai.org

Alexander Kheyfits
Department of Mathematics and Computer Science
Bronx Community College (CUNY)
University Avenue at W. 181st Street
Bronx, NY 10453
718-289-5616
alexander.kheyfits@bcc.cuny.edu

Kathy Andrews
Hillsdale College
Department of Mathematics
33 E. College
Hillsdale, MI 49242
301-552-6429
keandrews@dmci.net

Other DIMACS modules related to this module:

Module 03-1 by Alex Kheyfits.

(There may be others of which we are not yet aware)