L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

L_1 and L_2 approximation clustering for mixed data: Scatter decompositions and algorithms

Boris Mirkin

Rutgers University, Piscataway, USA and
Central Economics-Mathematics Institute, Moscow, Russia

Abstract: Clustering is considered usually an art rather than a science because of lacking comprehensive mathematical theories in the discipline. The major issue raised in this paper is that L_2 and L_1 approximation bilinear clustering can provide a theoretical framework for an extensive part of partitioning and hierarchic clustering concerning its algorithmical and interpretational aspects, which is supported with a theoretical evidence.

Key words: Partitioning, hierarchy, mixed data, approximation, contingency coefficients.

AMS subject classification: 62H30, 90C27, 05C50, 05C05.

1 Introduction

Clustering is considered usually an art rather than a science because of lacking comprehensive mathematical theories in the discipline. The major issue raised in this paper is that approximation bilinear clustering can provide a theoretical framework for a part of partitioning and hierarchic clustering concerning its algorithmical and interpretational aspects. Two approximation norms, L_1 and L_2, are considered and compared.

The remainder consists of two parts devoted respectively to partitioning (Sections 2 and 3) and hierarchic clustering (Section 4), and a conclusion (Section 5). In Section 2, a bilinear model relating data to a partition is considered. The model is introduced in Section 2.1 where two model-based principles for data standardization are suggested. In Section 2.2., an L_2 decomposition of the data scatter into explained and unexplained parts is