

**“Hypergraph Based Clustering in High-Dimensional Data Sets:
A Summary of Results”**

by Eui-Hong (Sam) Han, George Karypis, Vipen Kumar and Bamshad Mobasher

Presented by: Jianlan Song

Overview

- Introduction
- Hypergraph-Based Clustering
- Experimental Results
- Conclusion and Directions for Future Work

Introduction

- Clustering in data mining
- traditional clustering techniques
- Clustering based on a hypergraph model

Hypergraph-Based Clustering

- Hypergraph Modeling
 - Use the support as the weight of hyperedge
 - Use the confidence as the weight of hyperedge
 - Association-rule hypergraph

Hypergraph-Based Clustering (Continue)

- Finding Clusters of Items
 - Relationships presented by frequent items sets are “fine grain”
 - HMETIS
 - fitness function
 - $fitness(C) = \frac{\sum \{Weight(e) \mid e \subseteq C\}}{\sum \{Weight(e) \mid |e \cap C| > 0\}}$
 - connectivity function
 - $connectivity(v, C) = \frac{|\{e \mid e \subseteq C, v \in e\}|}{|\{e \mid e \subseteq C\}|}$

Experimental Results

- S & P 500 stock data
 - **Original data set**
 - binary table of size 1000×716
 - Each row corresponds to up or down movement indicator for one of the 500 stocks
 - Each column corresponds to a trading day from Jan. 1994 to Oct. 1996
 - **Clustering**
 - minimum support threshold of 3%
 - hypergraph of 440 vertices and 19602 hyperedges
 - **Partitions**
 - 40 partitions that 20 of them satisfy the fitness function
 - 16 out of these 20 partitions are clean clusters

Experimental Results (Continue)

- Protein coding database
 - **Original data set**
 - binary table of size 662×11986
 - Each row corresponds to an EST
 - Each column corresponds to a protein
 - **Clustering**
 - using a support of 0.02%
 - hypergraph with 407 vertices and 128,082 hyperedges
 - **Partitions**
 - 39 out of 46 partitions satisfied the fitness criteria
 - 12 of the 39 clusters are very good, 3 are bad and 2 are undetermined. Each of the remaining 22 clusters has subclusters corresponding to distinct protein families

Conclusion and Directions for Future Work

- Advantages
 - Does not require dimensionality reduction
 - Ability to control the quality according to the requirements of users and domains
 - Linearly scalable with respect to the number of dimensions of data and items
- Limitations
 - Suffers from the fact that right parameters are necessary to find good clusters
 - Does not naturally handle continuous variables as they need to be discretized