Proximus: A Methodology for Error-Bounded Compression and Categorization of Discrete Attribute Vector Sets.

Paul Ruth[1], Ananth Grama[1], Vipin Kumar[2], and Naren Ramakrishnan[3].

[1] Computer Sciences, Purdue University.

[2] AHPCRC, University of Minnesota.

[3] Computer Sciences, Virginia Tech.

Supported by generous support from the National Science Foundation and CERIAS.

Problem Formulation

- Given a set of discrete attribute vectors, determine a set of representative vectors (also discrete) such that every vector in the original set is within some bounded distance e from it.
- The method must scale to very large numbers of vectors and dimensions.

Problem Variants

- Clustering
- Vector Quantization
- Categorization
- Compression
- Pattern Extraction

Compressing Attribute Vectors

- Example: Consider the simple set of three attribute vectors:

$$\overbrace{\begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}}^{\textit{Attributes}}$$

This set of vectors can be simply represented as 2 of [0 1 1] and 1 of [1 1 1].

Compressing Attribute Vectors

⌘Consider the following rank-1 matrix:

$$r = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}$$

Since the order of vectors is not important, we can simply write this as 2 instances of [0 1 1 0].

But: Attribute vector sets are never rank-1!

Compressing Attribute Vectors

- Aha! But I could fix that for you..

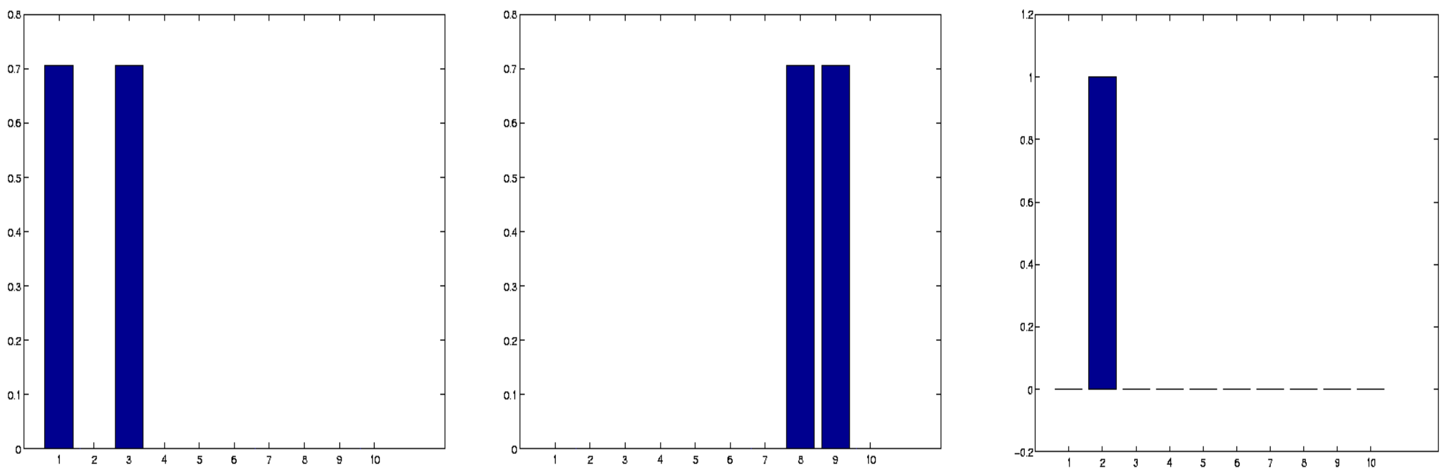
- Decompose the matrix into a sequence of rank-1 matrices using singular value decomposition.

Compressing Attribute Vectors

- But does this really solve the problem?
  - Remember, there are n vectors, each of which are of dimension m. n is typically much larger than m and the attribute set is sparse (that is, there are only O(n) non-zeros in the transaction set.
- We want to compress into something that takes much less than O(n) space.

Compressing Attribute Vectors.

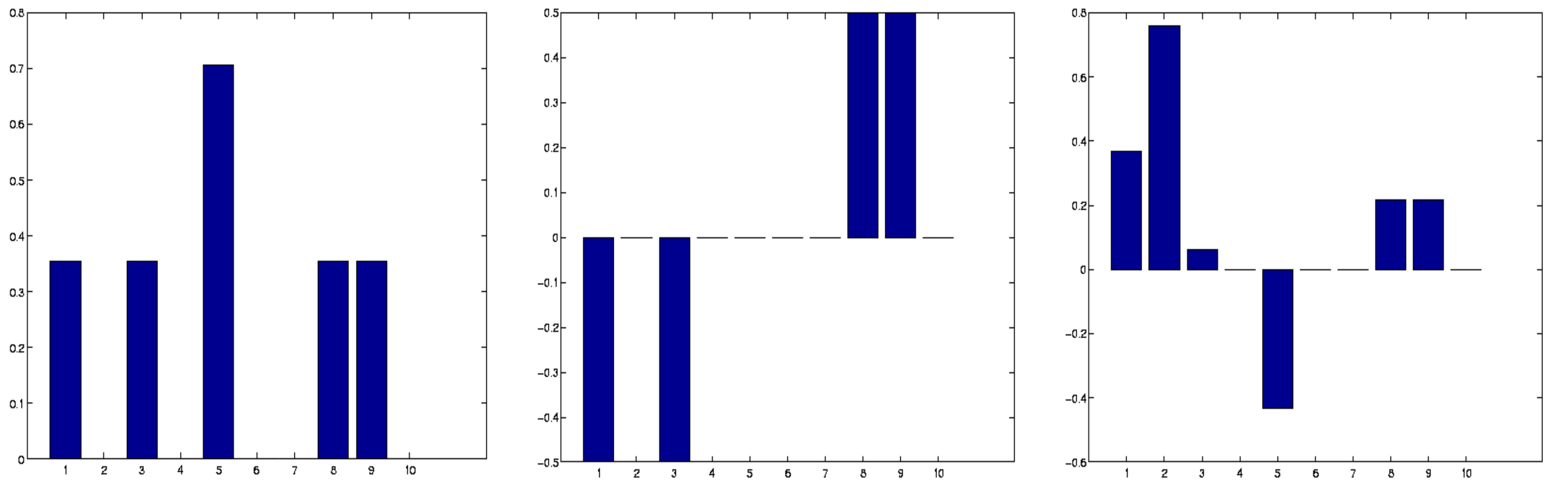
- Consider the singular vectors of a sample binary attributed transaction set:



This one worked rather nicely!

Compressing Attribute Vectors.

- Watch what happens here though!



Compressing Attribute Vectors.

- Singular vectors are orthogonal. There is no physical interpretation of the negatives.
- Singular vectors are only defined w.r.t prior singular vectors. Reconstruction is the only known way to query original data-set.
- Non-integral values for discrete attribute sets do not have physical interpretations.
- Non-integral column values do not have any physical interpretation either.

Compressing Attribute Vectors.

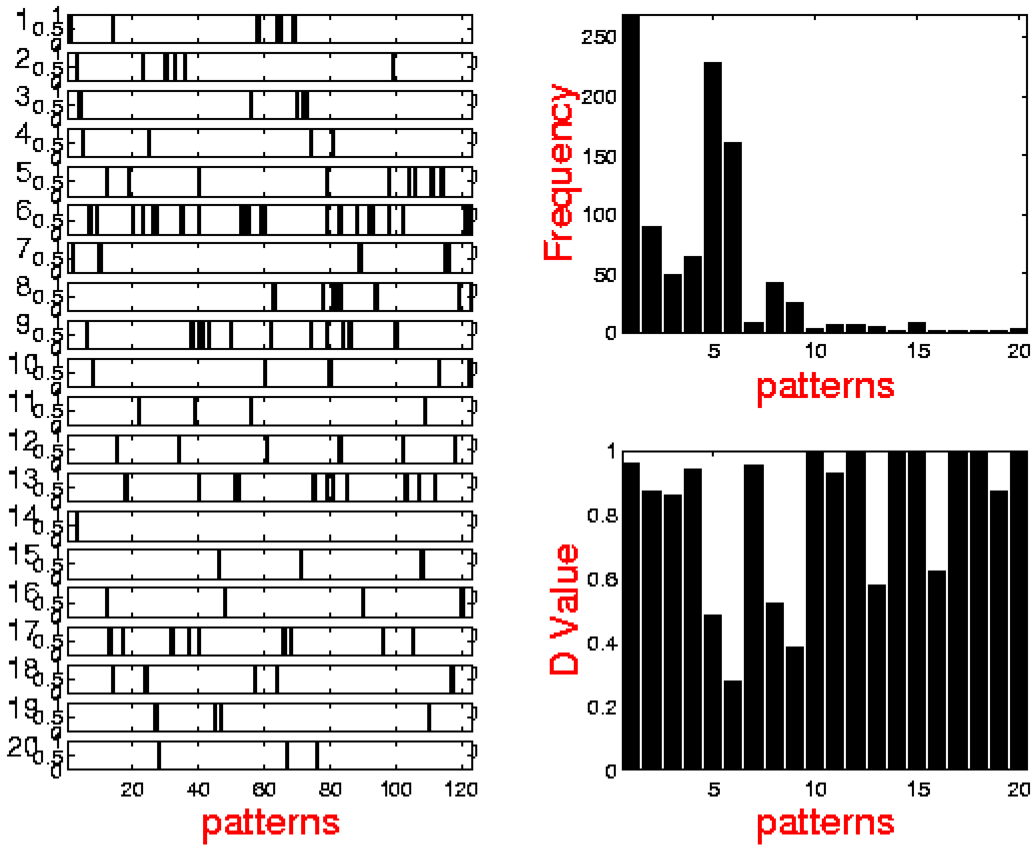
- Use discrete transforms for solving problems related to non-integral values.

- Discrete transforms are variants of semi-discrete decompositions (SDD) in which the outer product vectors can only take the values 0 or 1. Singular values can take arbitrary values though.

Discrete Transforms for Compressing Vectors.

- Relaxing Orthogonality:
  - Compute first discrete singular vector.
  - Eliminate all attribute vectors that are well approximated by the singular vector.
  - If no vectors match, remove the best few vectors and reinsert them at the end.
  - Repeat until patterns are statistically insignificant.

P r o x i m u s !



Proximus: Applications - Stock Market Data.

- 103 stocks selected at random.
- Data corresponds to high, low, open, close, and volume over two years.
- Discretize the stock data using standard indicators.
- Results in 103 vectors, each of length 5800 (15 attributes, 520 trading days).

Proximus: Technical Analysis of Stock Data.

- Error tolerance can be adjusted. At Hamming distance of 40% of vector length, we get following groups:

  - egrp, msft, sape, tecd, vcom, vsea
  - amxn, atvi, bosa, eftd, intc, mcicp, trid, vias, vshp, vtss
  - elnk, ifmx, lcos, stmp
  - coke, lnce
  - cost, naut, safc
  - aapl, bnbn, dell, ebay, hits, ibm, mcaf, mqst, novl, psft

  (and others).

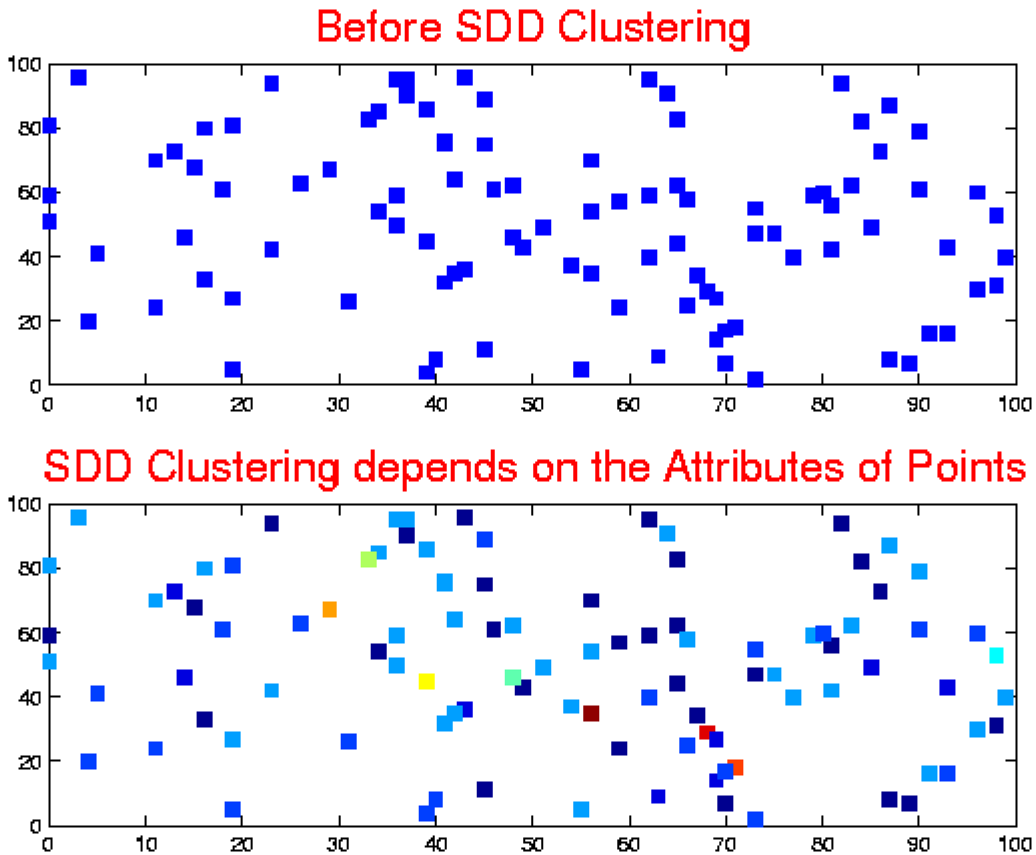
Notes on Proximus for technical analysis.

- It is not about identifying stocks in the same sector; rather about identifying stocks that exhibit same behavior with respect to selected indicators.
- Grouping is only as good as the indicators.
- In addition to groupings, we also get dominant behavior.
- Proximus also tells what which indicators are significant and which are not.

Application: Document Classification and Retrieval.

- Using vector space model of documents we can use the Proximus framework to classify.
- Searches can then be performed with respect to the dominant vectors corresponding to each category.

Application: Classifying Point-sets.



.. So where are we now?

- How do we analyze representative vectors?
- Applications of the Proximus framework.
- Theoretical bounds on the optimality of the representative vector set.