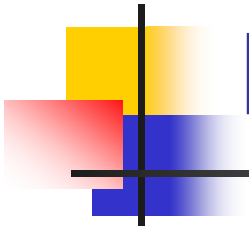
A Spectral-based Clustering Algorithm for Categorical Data Using Data Summaries (SCCADDS)

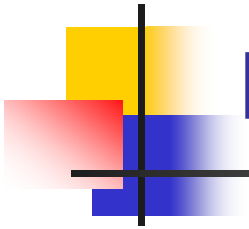
Eman Abdu
eha90@aol.com
Graduate Center
The City University of New York

Douglas Salane
dsalane@jjay.cuny.edu
Center for Cybercrime Studies
John Jay College of Criminal Justice
The City University of New York



Presentation Outline

- Motivation and Intended Applications
- Overview and Background
- SCCADDS Algorithm
- Experimental Evaluation and Comparative Analysis
- Conclusion and Future Work



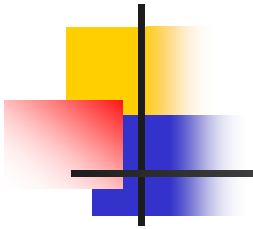
Problem

- The goal is to cluster a data set of “data objects” where each data object contains attributes (features).
- Each attribute is categorical – that is each attribute can have one value from a finite domain.
- The possible values for each categorical attribute are discrete nominal values with no pre-defined order or relationship.

Intended Applications

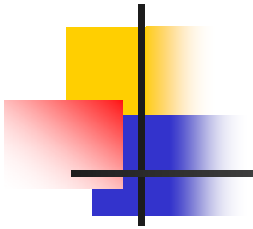
- Cluster Crime Data

- FBI's National Incident Based Reporting System (NIBRS)
  - Categorical Attributes
  - High Dimensional data
  - Large and Complex Database (incident, offense, victim, property, offender, arrestee)



Clustering Categorical Data Challenges

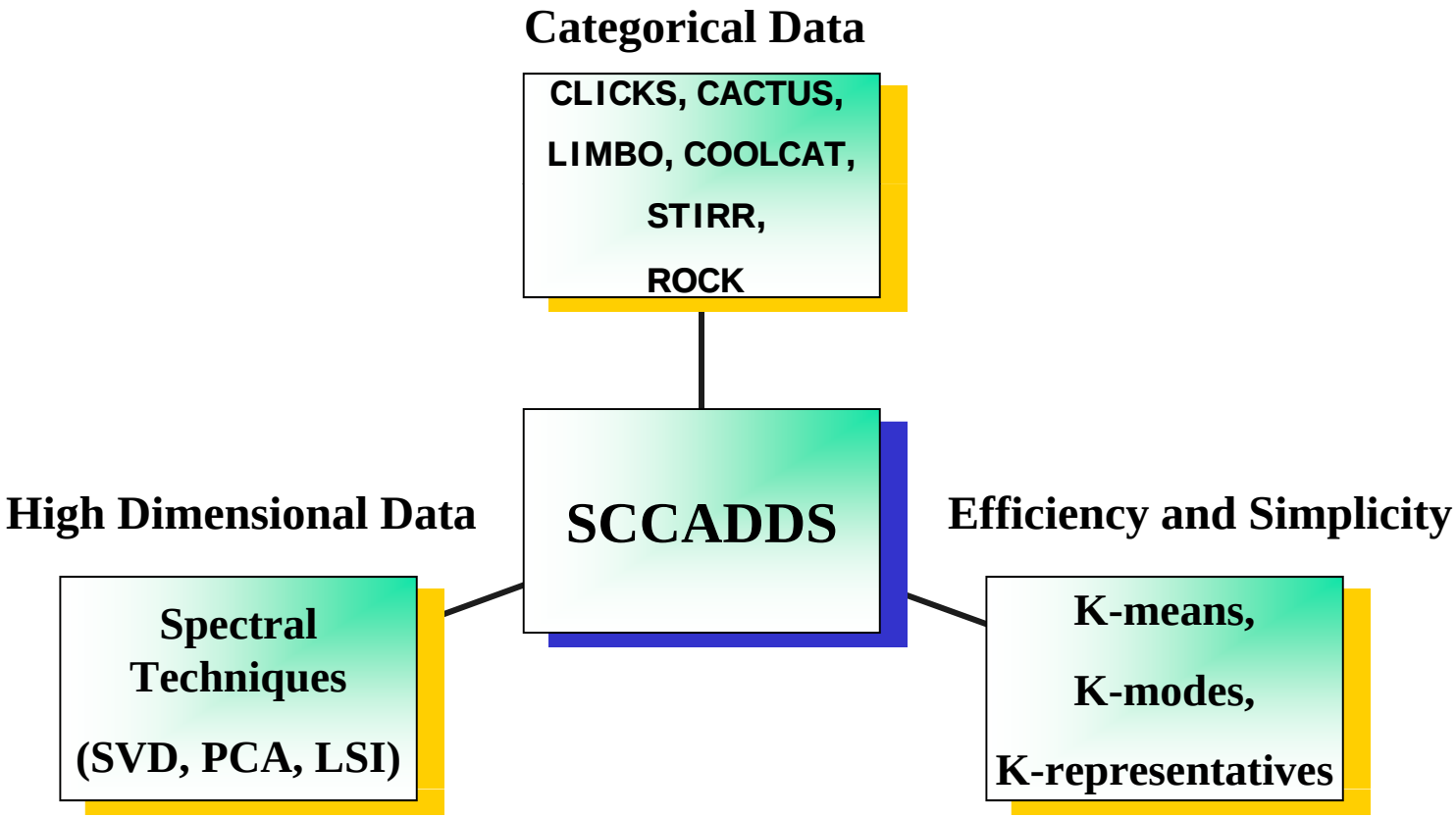
- Lack of geometric interpretation of data objects
- Difficulty in using traditional similarity measures such as cosine measure or Euclidean distance
- Difficulty in using other numeric techniques such as weighted average and other statistical measures (i.e., Standard of deviation, variance)
- High dimensionality



Current Techniques

- Parameters that are difficult to tune
- Not suitable for large data sets - Time complexity is often quadratic in terms of the number of data objects.
- Do not handle high dimensional data

Solution


Categorical Data
CLICKS, CACTUS, LIMBO, COOLCAT, STIRR, ROCK
SCCADDS
High Dimensional Data
Spectral Techniques (SVD, PCA, LSI)
Efficiency and Simplicity
K-means, K-modes, K-representatives

Categorical Data in Binary Format

- binary vector – each data object attribute represented by multiple components.
- Each component corresponds to one attribute domain value.
- K-means and spectral clustering algorithms can now be used on the binary records.

	color	shape
1	red	square
2	blue	circle

	Red	Blue	Square	Circle
1	1	0	1	0
2	0	1	0	1

Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) of an mxn matrix A is the decomposition

$$A = USV^t ,$$

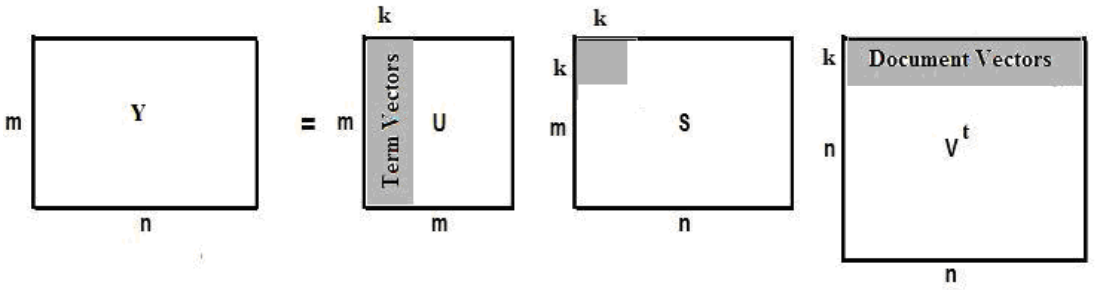
U is an m x m orthogonal matrix

V is an n x n orthogonal matrix

S is an m x n matrix with non zeros on the diagonal only.

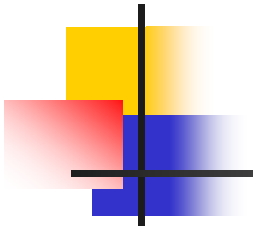
Truncated SVD of a Term/Document Matrix (Latent Semantic Indexing)

Y – Matrix of m terms (rows) by n documents (columns)



SVD of the term-document matrix

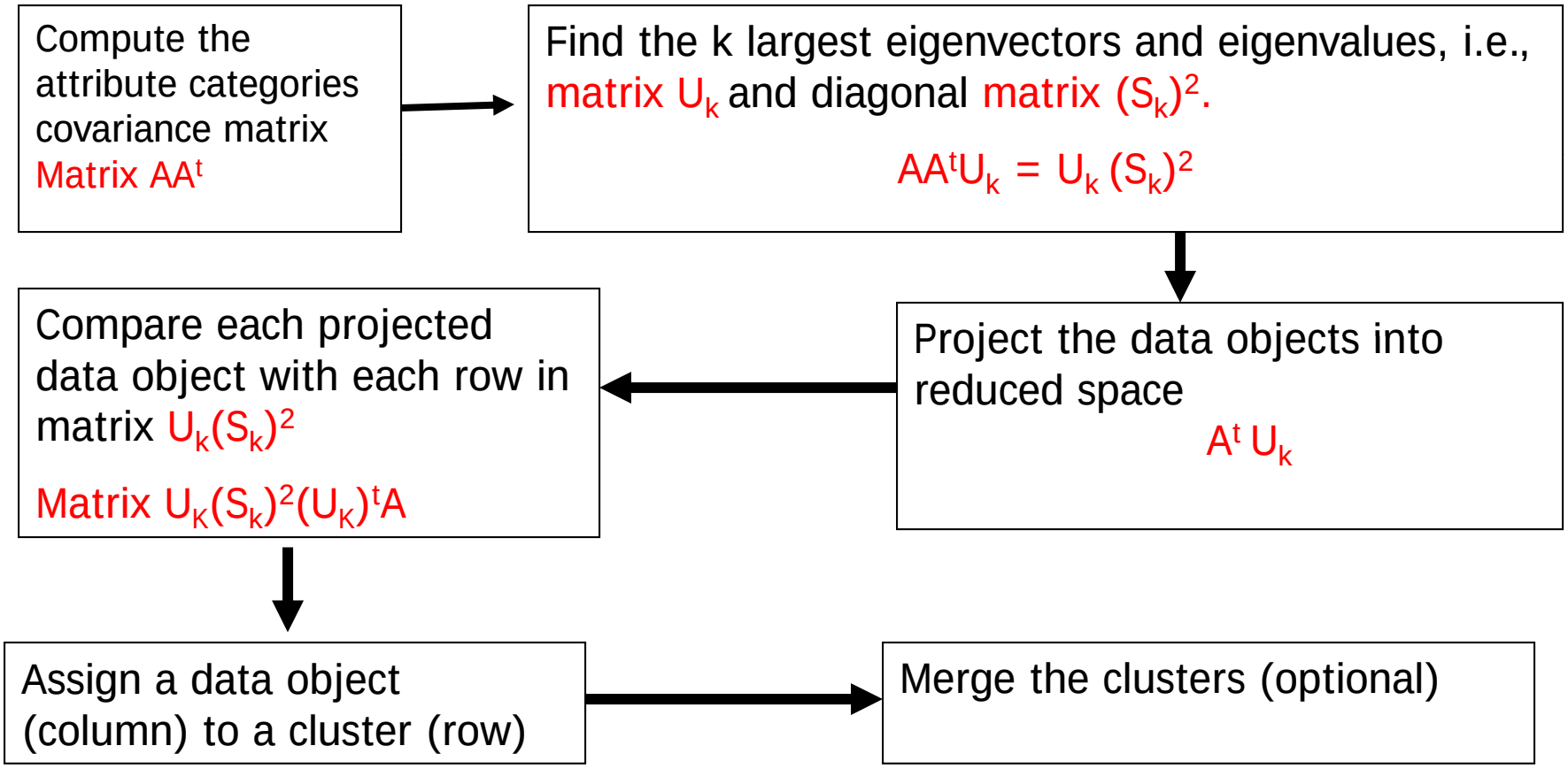
$$Y_k = U_k S_k V_k^t$$

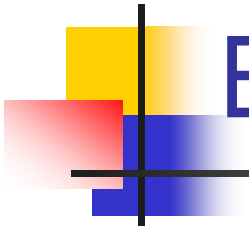


LSI

- Terms and documents are projected into a reduced space (k dimensions) – matrices U_k and V_k
- LSI allows the comparison of
  - Between terms (rows of $U_k S_k$)
  - Between documents (rows of $V_k S_k$)
  - Terms and documents (rows of U_K sqrt(S_k) and rows of V_K sqrt(S_k))
- LSI overcomes the problems of high dimensionality since all comparisons are performed in the reduced space.

SCCADDS


Compute the attribute categories covariance matrix
Matrix AA^t
Find the k largest eigenvectors and eigenvalues, i.e., matrix U_k and diagonal matrix $(S_k)^2$.
$AA^tU_k = U_k (S_k)^2$
Project the data objects into reduced space
$A^t U_k$
Compare each projected data object with each row in matrix $U_k(S_k)^2$
Matrix $U_K(S_k)^2(U_K)^tA$
Assign a data object (column) to a cluster (row)
Merge the clusters (optional)



Experimental Evaluation

- Standard data sets
  - Quality testing
  - Comparative results
- Synthetic data sets
  - Clustering quality
  - Comparative results
  - Scalability – Data set size, dimensions

Standard Data sets
UCI Machine Learning Repository

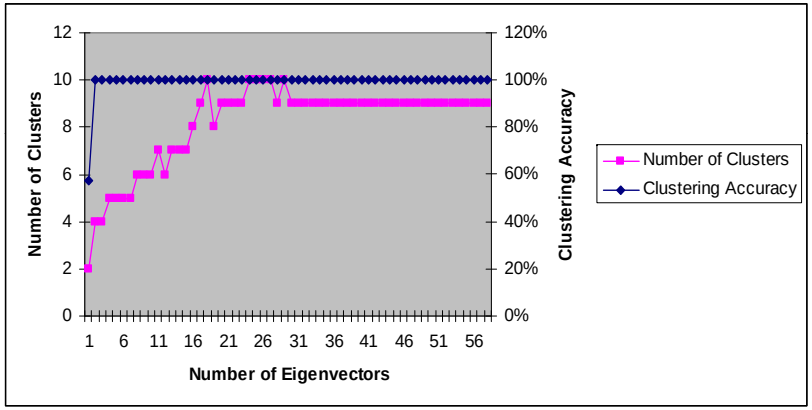
Dataset Name	Description	Number Of Records	Number Attributes	Number of Known labels
Soybean	Soybean diseases	47	21	4
Congressional	Voting dataset- Two parties democrats and republican	434	16	2
Mushroom	Mushroom species classified into poisonous and edible	8,124	22	2

SCCADDS
Quality Testing
Standard Data Sets

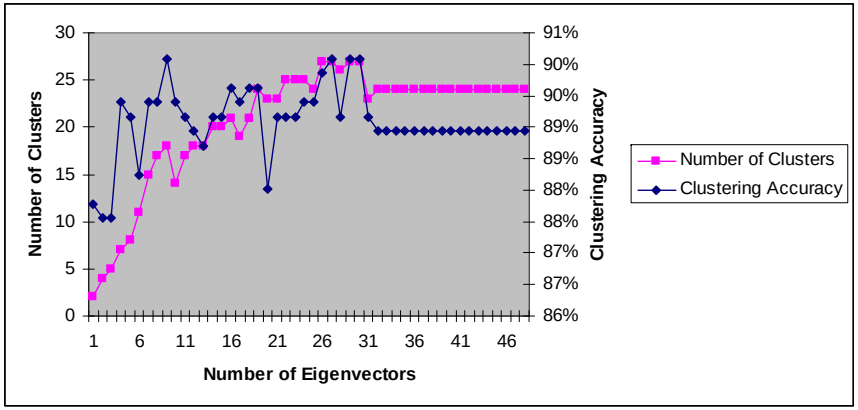
Data Set Name	Number of data objects	Number of eigen vectors	Number of Clusters in the data set	Number of Clusters Found by SCCADDS	Accuracy
Soybean	47	**2**	4	4	**100%**
Congressional Votes	434	**1**	2	2	**88%**
Mushroom	8,124	**1**	2	2	**89%**

SCCADDS
Quality Testing
Standard Data Sets

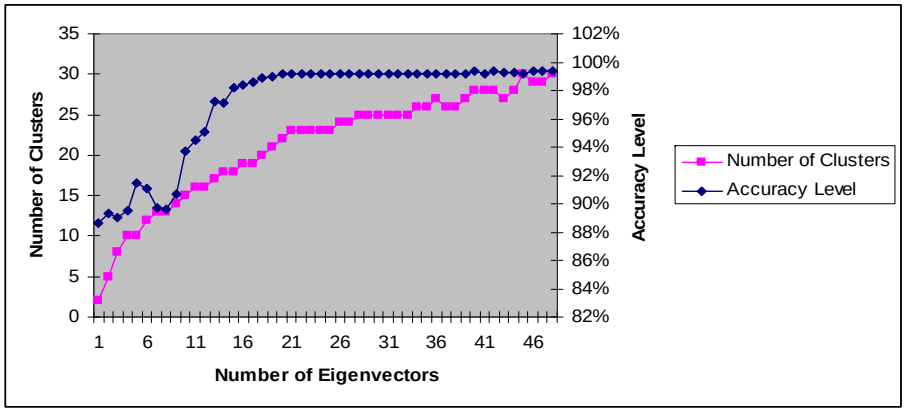
Soybean Data Set



Congressional Votes Data Set



Mushroom Data set



Comparative Results

Congressional Votes Data Set

	Accuracy	Entropy	Remarks
SCCADDS	88%	0.45	This is the result for 2 clusters. Higher accuracy levels may be achieved with a higher number of clusters
K-representatives	88%		This is average of 10 executions of the algorithm using random partitioning.
Traditional Hierarchical Clustering Algorithm	86%		
ROCK	79%	0.499	
LIMBO	87%		
COOLCAT	85%	0.498	
CLICKS	Not available for 2 clusters	Not available for 2 clusters	An accuracy level of 96% is achieved with 13 clusters.
Eigencluster		0.48	

Comparative Results

Mushroom Data Set

	Accuracy	Remarks
SCCADDS	89%	This is the accuracy level for 2 clusters. Higher accuracy levels may be achieved with a higher number of clusters.
K-representative	78%	This is average of 10 executions of the algorithm using random partitioning.
ROCK	57%	
LIMBO	89%	
COOLCAT	73%	
CLICKS	Not available for 2 clusters	An accuracy level of 97% is achieved with 19 clusters (Zaki et al., 2007).
Eigencluster	81%	

Synthetic Data Sets Quality Testing

SCCADDS

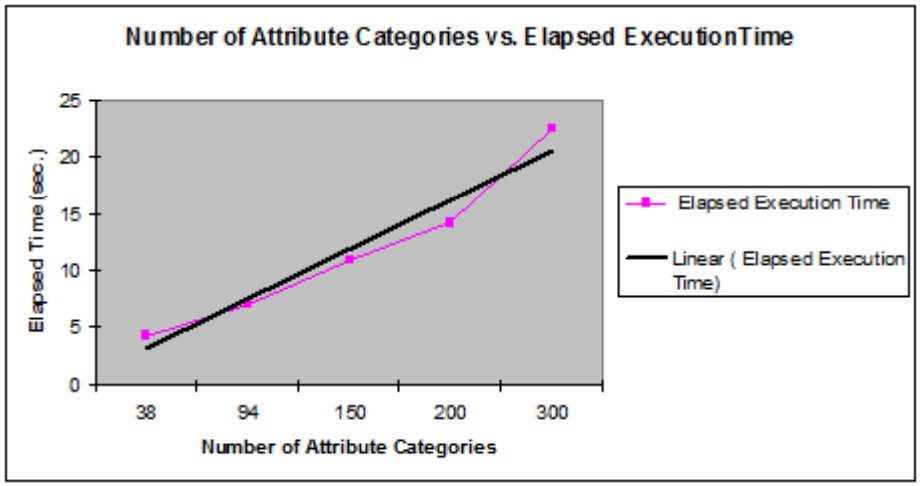
Data Set Name	Domain size of attributes	Number of Attributes	Number of attributes that participate in a rule	Number of Clusters	Data set Size	Noise Ratio
DS1	10-20	10	4	5	1000	0
DS2	10-20	10	4	10	5000	0
DS3	10-20	10	5	10	5000	2%
DS4	10-20	10	5	10	5000	10%
DS5	10-20	20	10	10	5000	10%

Synthetic Data Sets Quality Testing

SCCADDS

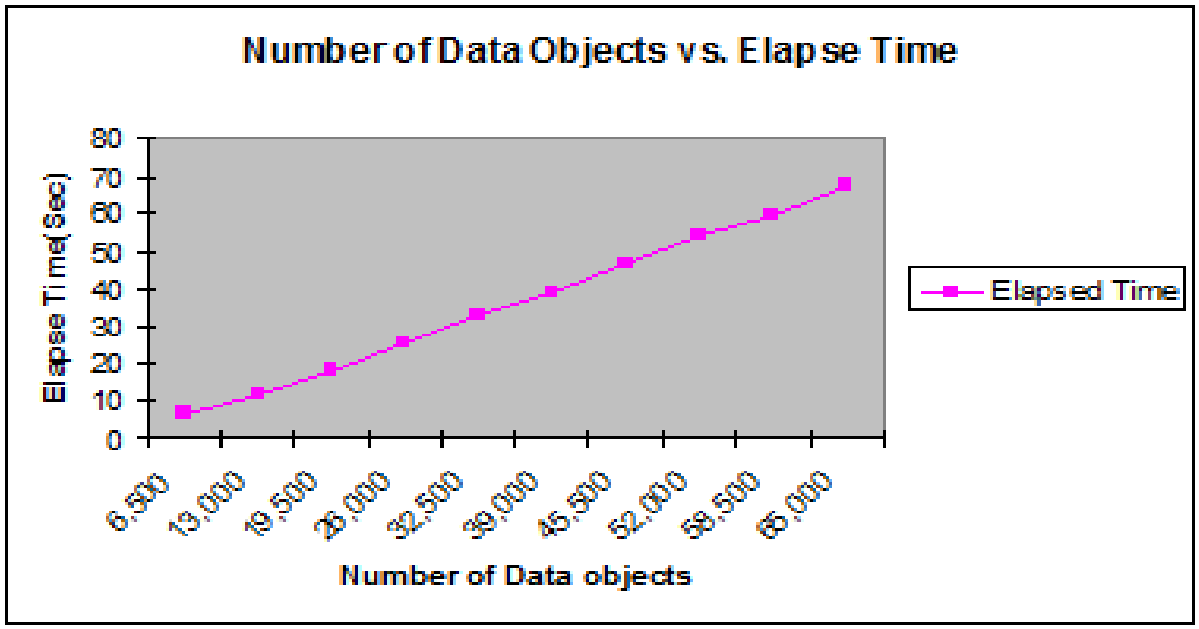
	SCCADDS			Eigencluster			K-representative
Data Set Name	Number of eigen vectors	Number of clusters found	Accuracy	Alpha	Number of clusters	Accuracy	Accuracy
DS1	4	5	100%	0.4	5	93%	96%
DS2	9	10	99%	0.4	10	91%	91%
DS3	9	10	100%	0.4	10	91%	91%
DS4	9	10	99%	0.4	10	85%	96%
DS5	7 9	10 10	99% 100%	0.2 0.35	9 11	82% 95%	90%

Synthetic Data Sets
Scalability Testing
Number of Attributes (Dimensions)

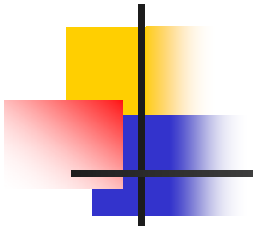


				SCCADDS			
Data Set Name	Number of Data Objects	Number of Attributes	Domain	Number of Attributes Categories	Elapsed Execution Time (sec.)	% change in Number of attributes	% of Change in Elapsed Execution Time
DS1A	5000	5	10	38	4.2	0	0
DS2A	5000	10	10	94	7	147.37%	66.67%
DS3A	5000	15	10	150	10.9	59.57%	55.71%
DS4A	5000	20	10	200	14.1	33.33%	29.36%
DS5A	5000	30	10	300	22.5	50.00%	59.57%

Synthetic Data Sets
Scalability Testing
Data Set Size



SCCADDS



Conclusion

• SCCADDS is a spectral-based algorithm for clustering categorical data

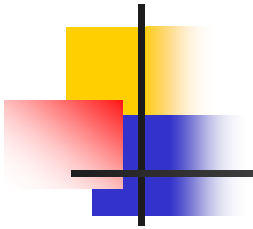
• Uses spectral techniques and data summaries

• Linear in terms of data objects and as such scalable to large data sets (assuming number of attributes and categories is not large.)

• Produces quality clusterings, better than most clustering algorithms in its class

•Few parameters, reasonably easy to tune

•Not highly sensitive to number eigenvectors

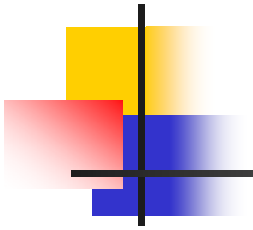


Future Work

- Optimize the implementation of SCCADDS
- Test SCCADDS on NIBRS data sets with multiple segments
- Extend SCCADDS to do fuzzy clustering (each data object belongs to more than one cluster, intensity measure provided.)

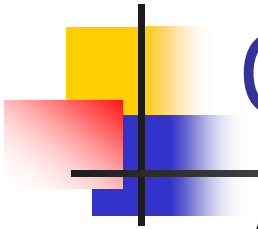
NIBRS

- An incident-based reporting system that collects detailed information regarding crime incidents
- Contains data from 1994 – 2005
- Contains data for over 29 million incidents
- Large and complex database with direct and indirect relationships between data segments



SCCADDS

- It uses the framework of Principle Component analysis to find the principle components for terms and documents
- Ding et al (2004) showed that principle components are a relaxed solution of the cluster membership indicators in k-means clustering algorithm.
- It uses the attribute categories covariance matrix
- Relation to LSI



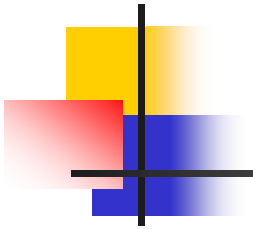
Clustering Algorithms

Algorithms specifically designed for Categorical Data

- K-modes Clustering Algorithms (Huang 1998)
- K-representatives Clustering Algorithms (San et al., 2004)
- CACTUS (Ganti et al., 1999)
- CLICKS (Zaki et al., 2007)
- ROCK (Guha et al., 2000)
- COOLCAT (Barbara et. al., 2002)
- STIRR (Gibson et al., 1998)
- LIMBO (Andritsos et al., 2004)

Spectral Algorithms

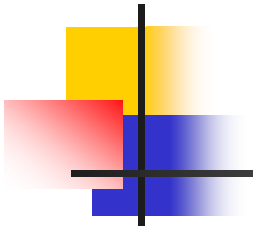
A Divide-and-Merge Methodology for Clustering (Eigencluster; Cheng, Kannan, Vempala and Wang, 2006)



Spectral Based Methods

- An effective technique for data reduction
- A technique that has successfully been used in clustering large and high dimensional data sets
- Theoretical foundation – graph partitioning (Shi and Malik, 2000), SSE (Ding and He, 2004), Variance based clustering (Drineas et al., 2004)
- Is not a greedy algorithm and not vulnerable to local optimum

Data Objects Matrix A



- A is matrix of data objects where the data objects are the columns
- For efficiency, we use the eigenvector decomposition of AA^t to arrive at the U (attribute values) and V (Data points) matrix.
- Mathematical relationships between eigenvectors decomposition and SVD

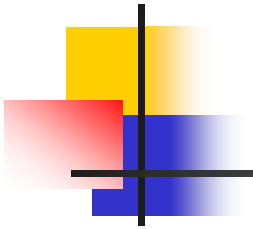
$$A = USV^t$$

$AV = US$ and $A^tU = VS$

$AA^t = USS^tU^t$ and $A^tA = VSS^tV^t$

$AA^tU = USS^t$

Note that AA^t is the attribute values similarity matrix and A^tA is the document-document similarity matrix.



Types of Clustering Algorithms

- Partitional
  - The clusters are disjoint and are not related to each other.
  - Each record belongs only to one cluster
- Hierarchical
  - The clusters form a tree structure
  - Divisive hierarchical vs. agglomerative hierarchical
- Fuzzy
  - Each record is associated with a weight that defines the extent of its membership to each cluster

SCCADDS
(**S**pectral-Based **C**lustering Algorithm for **Ca**tegorical **D**ata Using **D**ata **S**ummaries)

- Efficient – runs in linear time (number of data objects) – unlike most spectral-based clustering algorithms
- Scalable - uses the attribute-value similarity matrix instead of the data objects similarity matrix and as such is scalable to very large datasets (number of records)
- Produces quality clustering better than most algorithms in its class
- Easy to implement and tune