WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases

- Introduction
- Related Work
- WaveCluster Theory Overview
- WaveCluster Clustering Method
- Experimental Evaluations and Conclusions
- Questions and discussion

By Alex Strilets, December 1, 2000
CMPUT 695 Knowledge Discovery in Databases

Introduction

- Huge amount of spatial data accumulated from satellite images, medical equipment, GIS systems , etc.
- Characteristics of good clustering algorithm
  - good time efficiency
  - ability to identify clusters of arbitrary shapes (nested within one another, have holes, etc)

Introduction

- handling noise and outliers
- insensitive to the ordering of input data
- do not make any assumption about the number of clusters present
- ability to classify objects at a different level of accuracy

Introduction

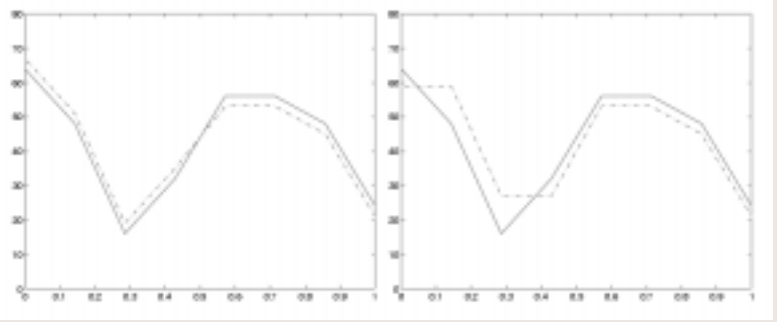
- Wave Cluster Algorithm Characteristics
  - capable of finding arbitrary shape cluster such as concave or nested clusters
  - does not assume any specific shape of clusters
  - prior knowledge about number of clusters is not required
  - not sensitive to outliers and ordering of input data
  - efficient on large databases O(N)

Related Word

- Partitioning Algorithms
  - PAM, CLARA, CLARANS
- Hierarchical Algorithms
  - BIRCH, CURE
- Density-Based Algorithms
  - DBSCAN
- Grid-Based Algorithms
  - STING, WaveCluster

WaveCluster Theory Overview

- Consider a set of following numbers:
  64, 48, 16, 32, 56, 56, 48, 24



WaveCluster Theory Overview

Example

64	48	16	32	56	56	48	24
56	24	56	36	**8**	**-8**	**0**	**12**
40	46	**16**	**10**	**8**	**-8**	**0**	**12**
43	**-3**	**16**	**10**	**8**	**-8**	**0**	**12**

67	51	19	35	53	53	45	21
59	27	53	33	**8**	**-8**	**0**	**12**
40	43	**16**	**10**	**8**	**-8**	**0**	**12**
43	**0**	**16**	**10**	**8**	**-8**	**0**	**12**

59	59	27	27	53	53	45	21
59	27	53	33	**0**	**0**	**0**	**12**
43	43	**16**	**10**	**0**	**0**	**0**	**12**
43	**0**	**16**	**10**	**0**	**0**	**0**	**12**

WaveCluster Theory Overview

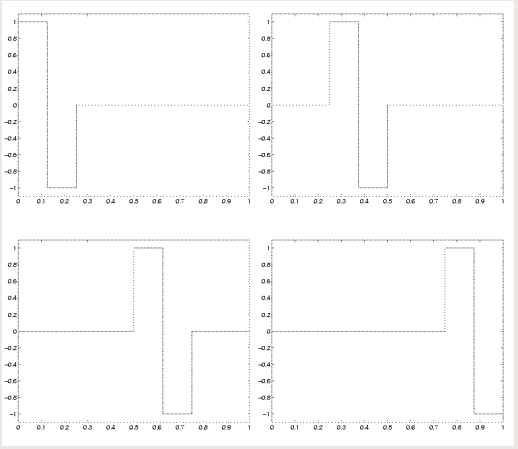
- Function *f(x)* is scaling function if it can be expressed as a liner combination of *f(2x-k)*
  - Example: Haar function: 1 on [0,1), 0 –elsewhere
    f(x) = f(2x) + f(2x-1)
- For each *j* let *Vj* be a vector space of 2^j funciton f^j_i defined as $f(2^j x-i)$
  - *V0 {f(x)}, V1{f(2x), f(2x-1)}, V2 {f(4x), f(4x-1), f(4x-2), f(4x-3)}*
  - *V0 (V1 (V2 …. (Vj*

WaveCluster Theory Overview

- Wavelet space W^j defined as an orthogonal complement of V^j in V^{j+1}
  - Lets *w(x)* be defined as 1 on [0,1/2), -1 on [1/2,1), 0 – elsewhere
  - *V1 = {f(2x), f(2x-1)} = V0 xW0*, were *V0={f(x)}* and *W0={w(x)}*, because *<f,w>=0* and every element of *V1* space can be represented as a liner combination of elements from *V0* and *W1* space:
    f(2x) = 1/2f(x) +1/2w(x), f(2x-1)=1/2f(x)-1/2w(x)

WaveCluster Clustering Algorithm: Quantization

- Example of W^4 wavelet space function



WaveCluster Clustering Algorithm: Quantization

- In our example above set of numbers can be represented via scale functions:

$64f^3_0 + 48f^3_1 + 16f^3_2 + 32f^3_3 + 56f^3_4 + 56f^3_5 + 48f^3_6 + 24f^3_7$

And decomposed into:

$48f^0_0 - 3w^0_0 + 16w^1_0 + 10w^1_1 + 8w^2_0 - 8w^2_1 + 0w^2_2 + 12w^2_3$

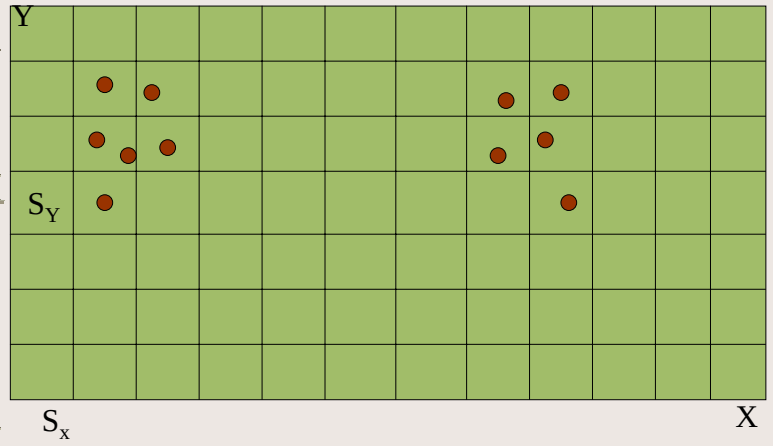
WaveCluster Clustering Algorithm

- Given a set of spatial objects o_i, $1<=i<=$ N detect cluster and assign labels to the objects based on the cluster they belong to
  - Quantize feature space, then assign object to the units
  - Apply wavelet transform on the feature space
  - Find connected components in the transformed feature space at different levels
  - Assign labels to the units
  - Make the lookup table
  - Map each object to the clusters

WaveCluster Clustering Algorithm: Quantization

- Divide each dimension d into m equal intervals, let s_i be the size of each unit in i dimension
- An object o_k corresponding to the feature vector $F_k=(f_1, f_2, \ldots f_d)$ will be assign to the unit $M_j=(v_1, v_2, ..v_d)$ if
  for all i: $1<=i<=d$ $(v_i-1)s_i <= f_i < v_i s_i$

WaveCluster Clustering Algorithm: Quantization



WaveCluster Clustering Algorithm: Transform

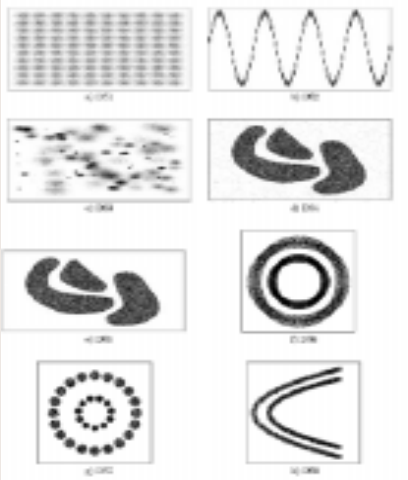
- Applying wavelet transform on each unit M_i result in new transformed feature space T_k
- Each connected component is a set of units from T_k and is considered a cluster
- Corresponding to each resolution level r there will be a set of clusters C_r
- Use some other well-know algorithms to find connected components in transformed feature space

WaveCluster Clustering Algorithm: Label and Lookup Table

- Assign each point in the transformed feature space to one cluster
- Clusters found in the transformed feature space are based on the wavelet coefficients
- WaveCluster algorithm makes lookup table to map units in the transformed feature space to the units in the original feature space
- Label each point in the original feature space with label of the unit that it belongs to

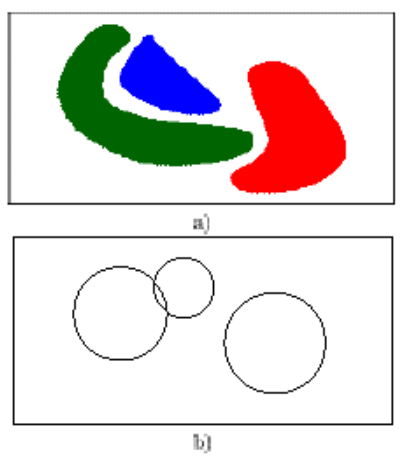
Experimental Evaluation and Conclusions

- Synthetic Datasets
- Each datasets contains 100,000 points



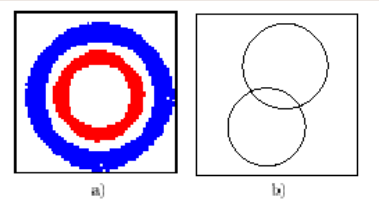
Experimental Evaluation and Conclusions

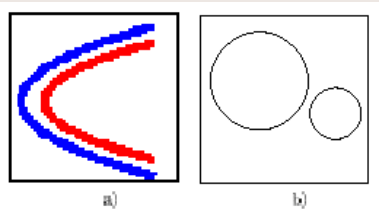
- Clustering Arbitrary Shapes
  a) – WaveCluster
  b) - BIRCH



Experimental Evaluation and Conclusions

- Clustering Nested and Concave Patters
  a) – WaveCluster
  b) - BIRCH





Experimental Evaluation and Conclusions

- Compare WaveCluster, BIRCH and CLARANS
- 8-10 times faster then BIRCH
- 200-300 times faster then CLARANS
- Total time = I/O time + Processing time
- I/O time > Processing time