CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling

George Karypis, Eui-Hong (Sam)Han, Vipin Kumar

Presented By: Jeff Antoniuk
Nov 20, 2000

Chameleon

- Introduction/Motivation
- Definitions –
  - K-nearest neighbor, relative inter-connectivity, relative closeness
- 2 Phase Algorithm
- Comparisons
- Conclusions
- Discussion

Introduction/Motivation

- Created to address short-comings of previous clustering algorithms
  - a. don't use internal information within clusters
  - b. merge based on only the min distance between representative points (ex CURE)

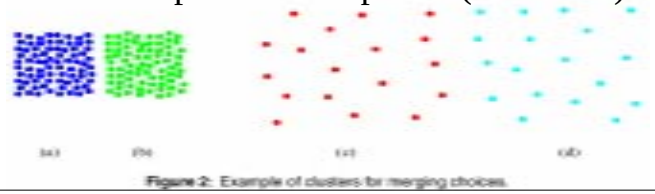


Figure 2: Example of clusters for merging choices.

Introduction/Motivation (Cont.)

  - c. merge based on aggregate interconnectivity between pairs of clusters (ex ROCK)

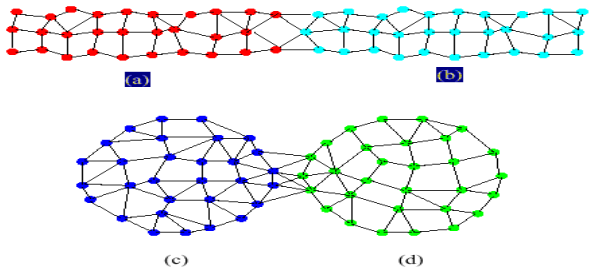


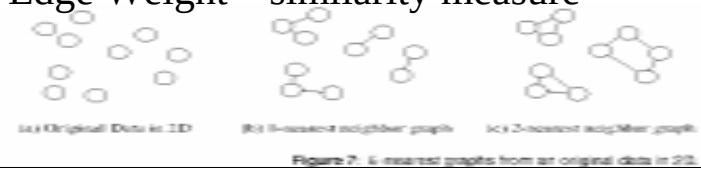
Figure 3: Example of clusters for merging choices.

Introduction/Motivation (Cont.)

- Merging decisions based on:
  - Relative interconnectivity
  - Relative closeness
- Dynamically adapts to differing internal characteristics of candidate clusters

K-nearest Neighbor

- Uses K-nearest neighbor sparse graph representation
- Vertex – data item
- Edge – between K most similar data items
- Edge Weight – similarity measure



Relative Inter-Connectivity

- Absolute inter-connectivity between C_i and C_j normalized wrt internal inter-connectivity of C_i and C_j

$$RI(C_i,C_j)=\frac{\left|EC_{\{C_i,C_j\}}\right|}{\frac{\left|EC_{C_i}\right|+\left|EC_{C_j}\right|}{2}}$$

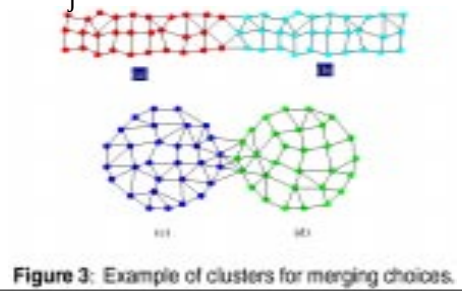
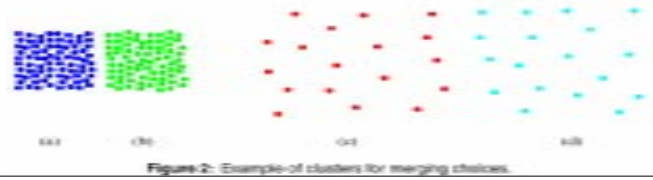


Figure 3: Example of clusters for merging choices.

Relative Closeness

- Absolute closeness between C_i and C_j normalized wrt the internal closeness of C_i and C_j

$$RI(C_i,C_j)=\frac{\bar{S}EC_{\{C_i,C_j\}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}EC_{C_i}+\frac{|C_j|}{|C_i|+|C_j|}\bar{S}EC_{C_j}}$$



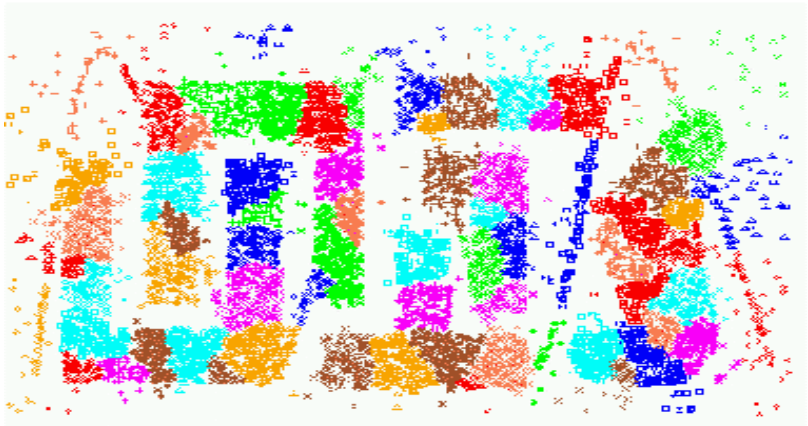
Algorithm – Phase I

- Purpose – partition data set into a number of sub-clusters to allow for dynamic modeling
- Algorithm
  - Starts all points belong same cluster
  - Repeatedly selects largest sub-cluster and bisects on min edge-cut where cluster size is > 25%
  - Terminates – largest sub-cluster contains less than MINSIZE vertices

Algorithm – Phase I (Cont.)

- MINSIZE – user specified parameter
- Smaller than the largest cluster expected to find in data set
- Sufficiently large to allow evaluation of relative closeness and relative inter-connectivity
- 1% - 5%

Algorithm – Phase I (Cont.)



http://www-users.cs.umn.edu/~karypis/publications/Talks/chameleon/

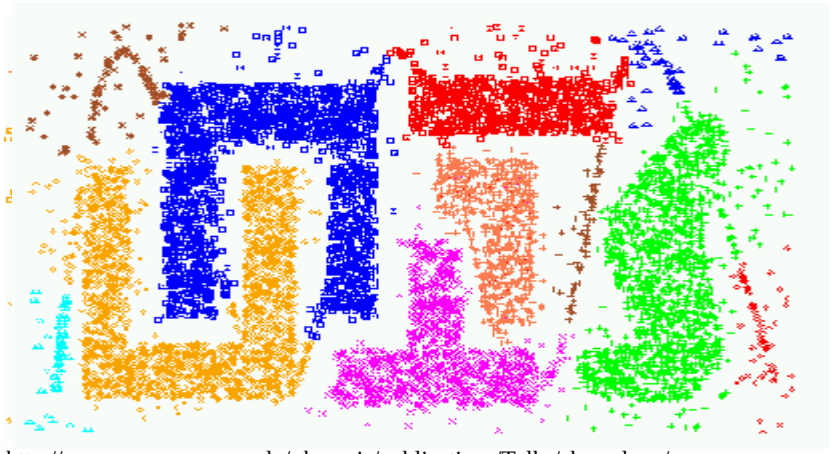
Algorithm – Phase II

- Purpose – merge sub-clusters using dynamic modeling
- Agglomerative hierarchical clustering
- Merges most similar pairs of sub-clusters based of relative inter-connectivity and relative closeness i.e. dynamic portion

Algorithm – Phase II (Cont.)

- Merging decisions based on 2 possible schemes
  1. $RI(C_i, C_j) >= T_{RI}$ and $RC(C_i, C_j) >= T_{Rc}$
  2. Maximize
     - $RI(C_i, C_j) * RC(C_i, C_j)^{\alpha}$
     - Method used in experiments

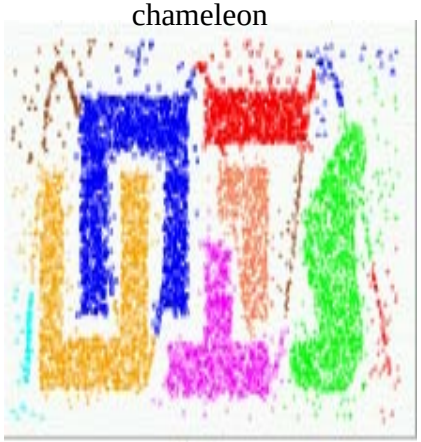
Algorithm – Phase II (Cont.)



http://www-users.cs.umn.edu/~karypis/publications/Talks/chameleon/

Comparison

chameleon

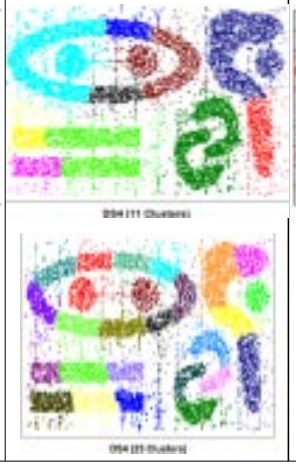


cure

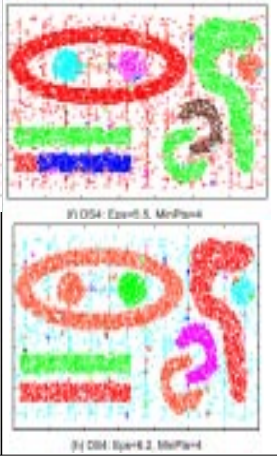
Comparison (Cont.)

chameleon

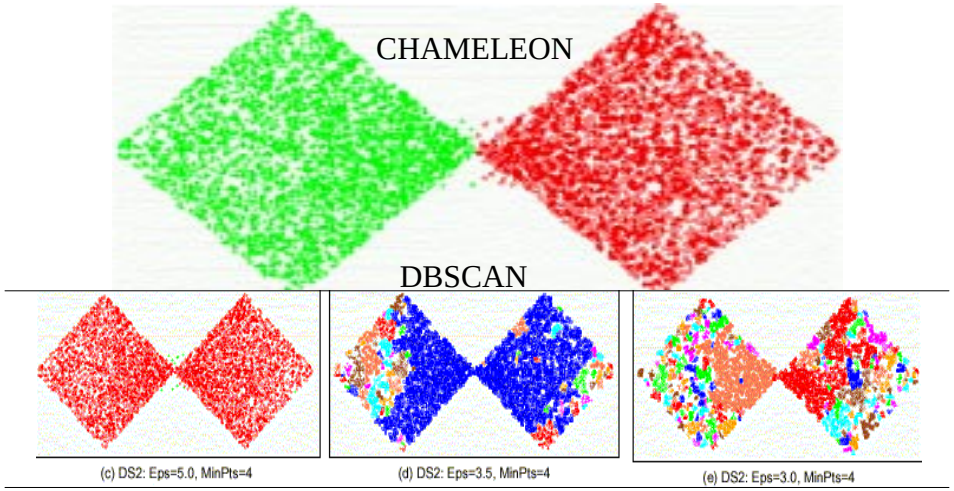
cure



dbscan



Comparison (Cont.)



Conclusion

- Advantages
  - Dynamic model reduces noise, poor merging decisions, considers shape of cluster
- Disadvantages
  - Graph must fit memory
  - Data item similarity measure required
  - cannot undo merge

Conclusion (Cont.)

- Major contributions
  - Agglomerative hierarchical clustering: Chameleon
  - Dynamic modeling
    - Relative inter-connectivity
    - Relative closeness

Bibliography

http://www-users.cs.umn.edu/~karypis/publications/Talks/chameleon/

George Karypis, Eui-Hong (Sam)Han, Vipin Kumar, *CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling*, Computer, Vol. 32, No. 8, August 1999