A Graph–Theoretic Clustering Algorithm: Finding Dense Regios in a Graph

Selim Aksoy
Intelligent Systems Laboratory
Department of Electrical Engineering
University of Washington
Seattle, WA 98195-2500
aksoy@isl.ee.washington.edu

November 15, 1998

This note summarizes the graph–theoretic clustering algorithm described in the paper by Shapiro and Haralick [1]. A clique of a graph is a subset of nodes that are connected to each other. The major clique is the clique that has the largest number of nodes. The goal of this algorithm is to find regions in a graph, i.e. sets of nodes, which are not as dense as major cliques but are compact enough within user specified thresholds.

1 Definitions

- (S, R) represents a **graph** where S is the set of nodes and $R \subseteq S \times S$ is a symmetric binary relation on S (if R is not symmetric, (S, R) is called a digraph).
- If M is a subset of S such that $M \times M \subseteq R$, then M is a **clique** of the relation R.
- The largest cliques are called **major cliques**.
- $(X, Y) \in R$ means Y is a **neighbor** of X.
- The set of all nodes Y such that Y is a neighbor of X is called the **neighborhood** of X and is denoted by Neighborhood(X).
- **Conditional density** $D(X|Y)$ is the number of nodes in the neighborhood of Y which have X as a neighbor;

$$D(X|Y) = \#\{N \in S \mid (N, X) \in R \text{ and } (Y, N) \in R\}.$$

- If R is symmetric, $D(X|Y)$ is the number of neighbors common to both node X and node Y;

$$D(X|Y) = D(Y|X) = \#\{\text{Neighborhood}(X) \cap \text{Neighborhood}(Y)\}.$$

- Given an integer K, a **dense region** Z around a node $X \in S$ is defined as

$$Z(X, K) = \{Y \in S \mid D(Y|X) \geq K\}.$$

- If M is a major clique of size L, then $X, Y \in M$ implies that $D(X|Y) \geq L$. Thus $M \subseteq Z(X, L)$ and $K \leq L \leq \#Z(X, K)$.

- $Z(X) = Z(X, J)$ is a **candidate dense region** around X where $J = \max\{K \mid \#Z(X, K) \geq K\}$.

- **Association** of a node X to a subset B of S is defined as

$$A(X|B) = \frac{\#\{\text{Neighborhood}(X) \cap B\}}{\#B}\ , \qquad 0 \leq A(X|B) \leq 1.$$

- **Compactness** of a subset B of S is defined as

$$C(B) = \frac{1}{\#B} \sum_{X \in B} A(X|B)\ , \qquad 0 \leq C(B) \leq 1.$$

- A **dense region** B of the graph (S, R) should satisfy

  1. $B = \{N \in Z(X) \mid A(N|Z(X)) \geq \text{MINASSOCIATION}\}$ for some $X \in S$,
  2. $C(B) \geq \text{MINCOMPACTNESS}$,
  3. $\#B \geq \text{MINSIZE}$

  where MINASSOCIATION, MINCOMPACTNESS and MINSIZE are thresholds supplied by the user.

- **Clusters** of the relation R can be determined by taking the union of the dense regions that have a high enough overlap. Define the **dense-region relation** F as

$$F = \{(B_1, B_2) \mid B_1, B_2 \text{ are dense regions of } R, \\ \frac{\#B_1 \cap B_2}{\#B_1} \geq \text{MINOVERLAP or } \frac{\#B_1 \cap B_2}{\#B_2} \geq \text{MINOVERLAP}\}$$

  where MINOVERLAP is a threshold supplied by the user.

2 Algorithm for finding dense regions

To determine the dense region around a node X,

1. Compute $D(Y|X)$ for every other node Y in S.
2. Use the densities to determine a dense–region candidate set for node X by finding the largest positive integer K such that $\#\{Y \mid D(Y|X) \geq K\} \geq K$.
3. Remove the nodes with a low association from the candidate set. Iterate until all of the nodes have high enough association.
4. Check whether the remaining nodes have high enough average association.
5. If the candidate set has a high enough compactness, check its size.

3 Algorithm for graph theoretic clustering

Now we have a dense–region for some (or all) of the nodes. To find the clusters of the graph,

1. Merge regions that have enough overlap if all of the nodes in the resulting set after merging have high enough associations.
2. Iterate until no regions can be merged. MINOVERLAP can be decreased by a small amount (e.g. %2) at each iteration.
3. Any node X which was not included in any cluster can be added to one of the clusters, B_i, by a best fit criterion if B_i satisfies $A(X|B_i) \geq A(X|B_j)$ for all clusters B_j.

Result is a collection of clusters in the graph.

References

[1] L. G. Shapiro and R. M. Haralick. Decomposition of two-dimensional shapes by graph-theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):10–20, January 1979.