Interactive Visual Clustering

Marie desJardins,[1] Julia Ferraioli,[2] and James MacGlashan[1]

[1]University of Maryland Baltimore County, [2]Bryn Mawr College

Motivation

Traditional automated clustering partitions data into clusters; however, the "best" clustering may depend on a user's goals.

Constrained clustering allows a user to provide some additional preconditions, but identifying the best set of constraints can be difficult.

Neither of these methods take into account data with relational attributes.

Proposed Solution

We propose an interactive method to allow a user to guide the clustering of data:

- We present the user with a visual display of the data using a spring-embedded layout.
- The user moves data nodes on the screen. Pairs of nodes that are moved close together receive must-link constraints; pairs moved far apart receive cannot-link constraints.
- After a node is moved, a constrained clustering algorithm is run using the specified constraints; new "cluster edges" are added to reflect the resulting clustering.
- Clustering of the data converges to the user's target clustering.

Data Sets

Circles - Synthetic data set with 120 instances. Contains two distinct clusters with attributes in the x-y plane.

Overlapping Circles - Synthetic data set with 100 instances; four overlapping clusters in the x-y plane.

Iris - From UC Irvine ML database. Originally 150 instances, reduced to 99 (33 randomly selected from each class). Clusters correspond to three species of Iris. The data contains 4 attributes.

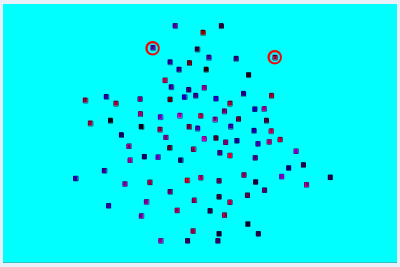
Amino Acid Indices - Measures chemical properties of amino acids; contains 20 attributes, one for each amino acid. Edges are present between instances with a high correlation. Two clusters are used: one for instances with alpha and turn propensities and another for hydrophobicity.

Amino Acid - Attributes and instances of Amino Acid Indices inverted. 20 instances with 25 attributes with high orthogonality selected by a domain expert.

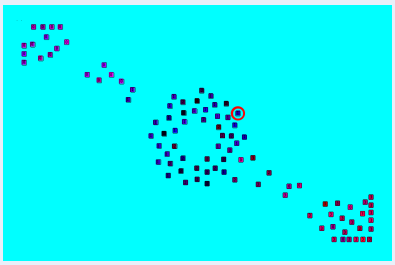
Evaluation

The effectiveness of the clustering algorithm is measured using the Adjusted Rand Index (ARI) metric. (An ARI of 0 means that instances are incorrectly clustered; an ARI of 1 means that all instances are correctly clustered.)
We simulate a user moving nodes, measuring the ARI after each movement, and compare IVC to other interactive methods: spring-embedded layout vs. manual; clustering vs. no clustering; and farthest-first vs. random node movements.

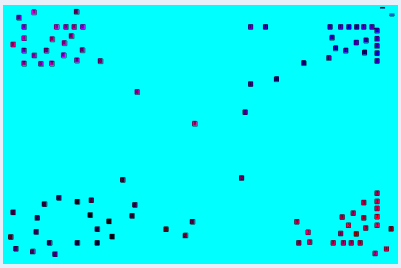
Screenshots



Initial display



After 2 nodes moved by user



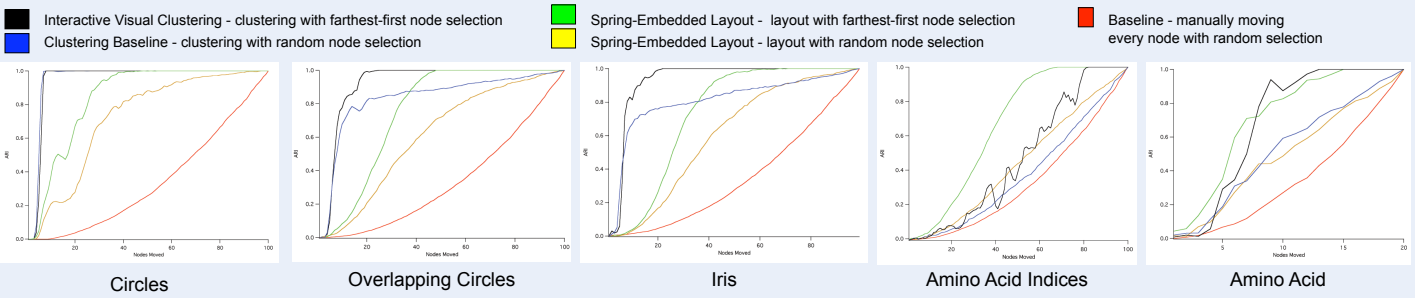
After 14 nodes moved by user

Results

IVC yields marked improvements for the Circles, Overlapping Circles, and Iris data sets. IVC performs worse than the spring-embedded layout for the Amino Acid Indices data set, and about as well as spring-embedding for Amino Acids, indicating that the constrained clustering is not leading to improved performance in these domains.

Future Work

Next steps include performing a user study and developing additional models of user behavior; incorporating additional types of user feedback; and working towards a mixed-initiative clustering paradigm.



Circles | Overlapping Circles | Iris | Amino Acid Indices | Amino Acid