Autonomous Visualization for Clustering

Yun Ni, Zhangning Hu, Yang Lu
yunn, zhangningh, yangl

Motivation

Human Interpretability

Current clustering algorithms, from centroid-based clustering to density-based clustering, are suffering from a lack of human interpretability. As a result, many domain experts find it difficult to adopt clustering techniques in the research of their fields. In this project, we want to visualize clustering in consideration of human interpretability.

Human Interpretability of clustering has two aspects
- The axis are simple expressions that can be understood by humans
- We can easily tell the clusters apart in the visualization

Problem

Formal Definition

Given a dataset $\{d_1, d_2 \ldots d_m\}$, we want to search the function space F: D -> R^2 for the optimal function

$$\arg\max_{f \in F} Pr\{f(d_1), f(d_2) \cdots f(d_m) \text{ is multimodal distribution}\}$$

so that we can easily see the clusters by scatter-plotting $f(d_1)$,$f(d_2)$...$f(d_m)$.

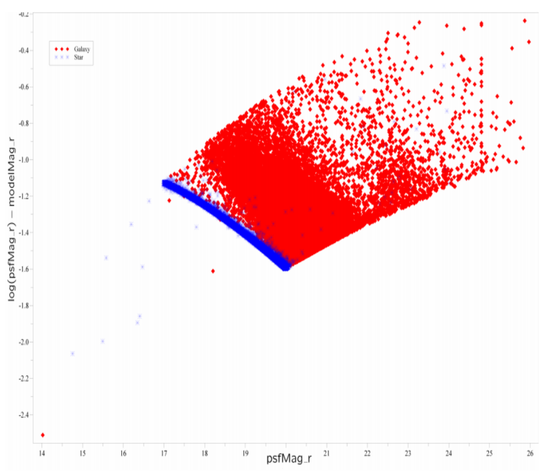
Function Space

Ideally, the function space F should be all the functions that are interpretable by humans. But to simplified the problem, we limit our functions to include at most 4 attributes in the dataset. That means we search over all the quadraple of the attributes and find the optimal function f*.

Related Works

Yes, Khalid El-Arini, Andrew W. Moore, and Ting Liu presents a novel approach to visualize classification in *Automonous Visualization, ECML/PKDD 2006*. The left scatter plot is one of the results in their work.

Concretely, they search the function space to maximize the distance between the centroids of positive instances and negative instances. By scatter-plotting by the optimal function, we can easily see the boundary of classification.



Methodology

Overview

In our algorithm, we score all scatter plots. High score necessarily but not sufficiently implies the scatter plot is multimodal distributed.
We can compute the score of scatter plot for every function in F and return the function with the best score.

Function Space Search

Rather than exhaustively searching all possible expressions, we use greedy search to find an approximation of the best pair of expressions:

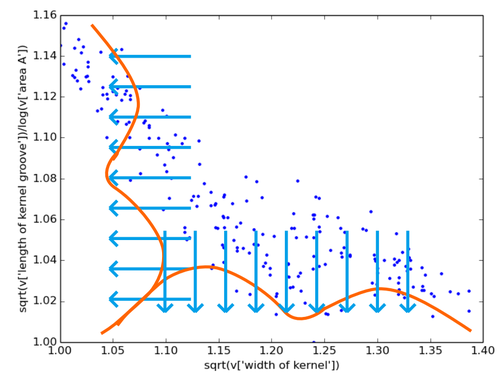
1. We exhaustively search all pairs of attributes with unary operations, recording the best pair.
2. Given the best pair, we consider all triples that contain the best pair, recording the best pair.
3. Given the best triple, we consider all quadruples that contain the best triple, recording the best pair.
4. We try to add unary operation to either expression. output the best pair.

Scoring Scatter Plot

This is the key step in our algorithm. We have tested three score functions.

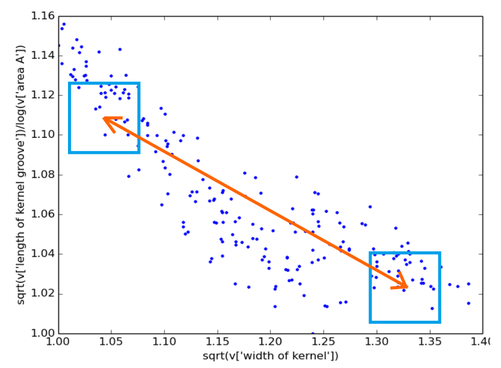
Score 1: Hartigan's Dip Test

We project all the data points onto x-axis and y-axis, and we use Hartigan's Dip Test to compute how likely the projection of data is multimodal distributed. Finally, we add up the dips of x-axis and y-axis as the score of the scatter plot.



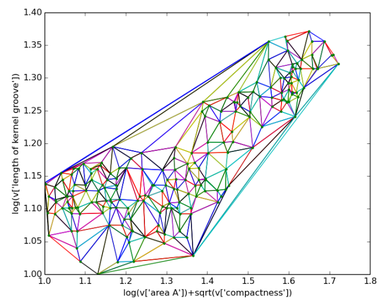
Score 2: Modes Distance

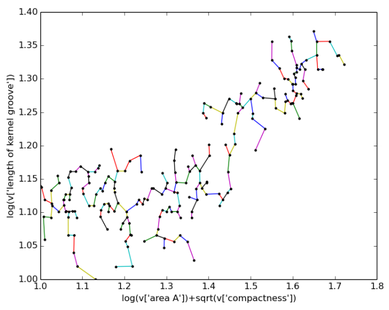
We scan over the plot using a moving window, and we count the number of points within each window. Having the number of points in each window, we compute the distance between the top two windows (in fraction of the plot size) as the score of the scatter plot.



Score 3: Minimum Spanning Tree

The scatter plot that humans see is actaully a 2D Euclidean Space. So we do Delaunay triangulation and draw a minimum spanning tree. We will try to score the scatter plot by random walks on the minimum spanning tree.

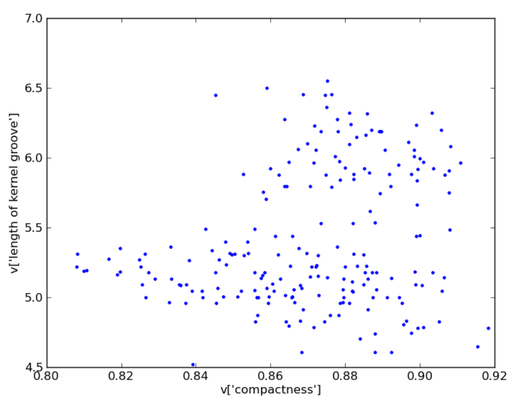




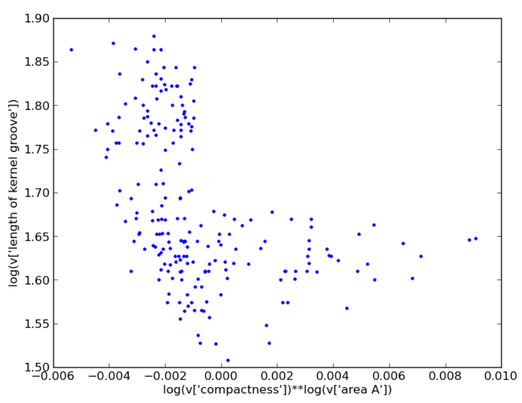
Results

The following figures show the optimal scatter plots by different scores. We can see that both plots are more interpretable than the origina data. In our test data, Score 1 performs better than Score 2. But we believe that in some cases, Score 2 measures "Interpretability" better than Score 1. We have not come up with a good way to score using MST yet. But we believe MST performs even better as it takes the structure of scatter plots into account.

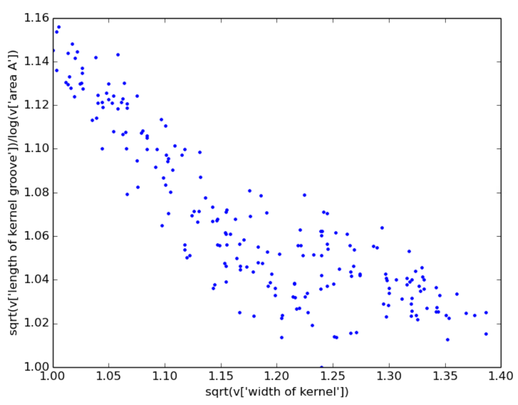
Original Data



Score 1: Hartigan's Dip Test



Score 2: Modes Distance



Score 3: Minimum Spanning Tree