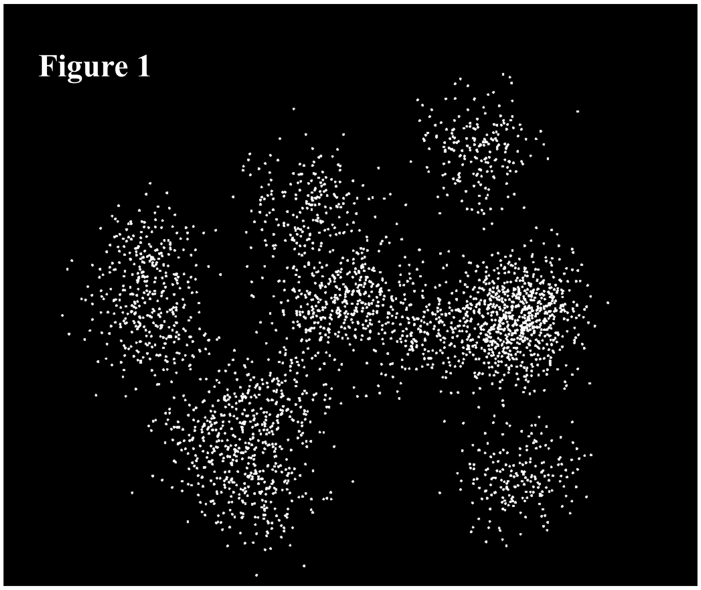
H-means: Using Human Judgments to Find the *k* in *k*-means

Abstract

This project describes a novel methodology for choosing the number of clusters to search for in an arbitrary data set. Human judgments concerning the number of clusters in several visual arrays are collected using psychophysical methods. I use these data to generate a probability distribution over the number of clusters present in each array. Leading algorithms, such as spectral eigengap methods, PG-means and X-means, are compared to the empirical data and then modified to better reflect human performance. Additionally, novel algorithms are proposed based on insights gleaned from post-task interviews of the human subjects.

Introduction

Humans are experts at categorization. In fact, categorization is often defined in terms of human performance. In computer science, splitting a data set into *k* separate groups is called clustering—a cornerstone of unsupervised machine learning with applications in computer vision, document classification and gene sequence analysis, among others. Clustering is like categorization without the semantics—pure numerical similarity alone is used to choose appropriate clusters; no top-down reasoning is involved.



Most clustering algorithms focus on how to best group all the points in a data set given a desired *k*. In many applications, however, *k* is not obvious, and several attempts have been made to design an algorithm that will automatically determine *k* [1, 2, 3]. There is no mathematical optimization that can universally choose *k*, and in many situations multiple values of *k* may be appropriate. In this situation, mining human judgment and tolerance for ambiguity is extremely valuable.

I propose to leverage the innate categorization expertise of humans in order to create a better *k*-choosing algorithm. I will present human subjects with a series of visual stimuli (Figure 1) representing several different data sets and query them about the number of clusters they perceive and the strategies they use to come to their conclusions. This human data will generate a probability distribution over *k* for each of the data sets, and give an informal idea of the criteria generally employed when making abstract clustering decisions.

I will then compare current computational methods for choosing *k* to my experimental results. The methods compared will include: Projected Gaussian (PG)-means [1], which projects both the data and the proposed model into one dimension in order to more effectively perform model fitness tests while iteratively adding cluster centroids, eigengaps in spectral clustering [2], which uses a random walk technique to find a high order transition matrix whose number of surviving (non-zero) eigenvalues indicate a suitable *k*, and X-means [3], which iteratively splits clusters and compares the Bayesian Information Criterion score of both the original clusters and the child clusters to decide which ones to keep. I then intend to enhance and combine the most successful of these methods, as well as introduce my own methods, resulting in a *k*-choosing algorithm more effective than the current state-of-the-art. This general approach of using psychological data to develop better machine learning algorithms has proved fruitful in other problem domains [4, 5].

I hypothesize that humans have multiple methods for visually grouping abstract data. These different methods will give rise to probability distributions over *k* that contain multiple peaks. These peaks may also be the result of a hierarchical interpretation of the data. I anticipate that hierarchical methods, such as the hierarchical spectral eigengap method [2], will be the best models for fitting human data.

Method

Human subjects will be presented with two-dimensional displays consisting of points of light on a dark background. These displays will be drawn both from real-world data as well as synthetic data generated from known distributions (a form of control, since I will know the exact mapping between underlying distribution and human judgment). Subjects will report the number of different groups (clusters) of light they see in the display, and rank their answers if they perceive more than one possible grouping. At the end, I will ask them to report any methods or techniques they felt they were using in order to make their determinations.

I will generate a probability distribution over *k* for each data set. I will then investigate whether the current computational methods mentioned above find clusterings that match peaks in the probability distributions generated from the human data. Using insights garnered from this process I intend to combine and modify these existing algorithms or introduce new algorithms to better replicate human performance.

Expected Results

For simple synthetic stimuli, I expect essentially unanimous agreement among subjects concerning *k*. As the synthetic stimuli become more complex and ambiguous through the addition of noise and data drawn from multiple overlapping distributions with high variance, there will be a wider range of subject responses. Responses to real data sets will likely vary between these extremes. The entropy of the resulting probability distribution will be a good measure of the difficulty of performing a particular clustering.

Subjects will sometimes report more than one perceived clustering possibility. Even with the very simple mixture of Gaussians stimulus shown in Figure 1, one might see four clusters (with the triangular center region being one large blob) or seven clusters. Finally, I expect that subjects will use multiple heuristics, such as finding dense regions and empty regions, and following paths that connect data points to one another.

For simple stimuli (those resulting in low entropy probability distributions) I expect agreement between humans and the algorithms, as well as between algorithms. As the stimuli become more complex I expect these correspondences to diminish. Particularly, I expect those algorithms that attempt to provide one best guess for *k* to fail, since the human data might be of the form “anywhere from 4 to 6 clusters.”

The psychophysical data I collect will be useful for other machine learning researchers by providing a measure of ambiguity as well as a desirable range of *k* for several real and synthetic data sets. The novel clustering algorithms I produce will help move the vital problem of choosing *k* forward.

The results of this study, while being very important for the field of computer science, will also find application across many different scientific domains. In fields as diverse as ecology, economics, and physics, methods are needed to accurately cluster large bodies of data without prior knowledge of an appropriate number of clusters. I am currently participating in research with an oceanographer to cluster the ocean into many ecologically relevant biomes using a high dimensional data set consisting of ocean measurements such as salinity, temperature and phosphate concentration. It would be very useful to have a data driven method of determining the number of biome types, without having to rely on expert human analysis.

References

[1] Feng Y, Hamerly G (2006). NIPS
[2] Azran A, Ghahramani Z (2006). CVPR
[3] Pelleg D, Moore A (2000). ICML
[4] Tenenbaum J (1999). NIPS
[5] Martin D, Fowlkes C, Tal D, Malik J (2001). ICCV