Finding a Better k: A psychophysical investigation of clustering

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
josh@cogsci.ucsd.edu

Abstract

Finding the number of groups in a data set, k, is an important problem in the field of unsupervised machine learning with applications across many scientific domains. The problem is difficult however, because it is ambiguous and hierarchical, and current techniques for finding k often produce unsatisfying results. Humans are adept at navigating ambiguous and hierarchical situations, and this paper measures human performance on the problem of finding k across a wide variety of data sets. We find that humans employ multiple strategies for choosing k, often simultaneously, and the number of possible interpretations of even simple data sets with very few ($N < 20$) samples can be quite high. In addition, two leading machine learning algorithms are compared to the human results.

Keywords: machine learning; clustering; psychophysics.

1 Human Data

Eighteen undergraduate human subjects were recruited for this project, 11 female and 7 male, to determine the number of groups present in 50 distinct point light displays. Each point light display was presented at two different scales and two different rotations, for a total of four presentations per display and 200 trials per subject. Subjects were asked to determine the number of groups in each display and were encouraged to give more than one answer if appropriate. There was no time limit for response. Subjects were told to ignore answers above 20 and to focus on "the bigger picture" to find a reasonable answer less than 20. The sequence of trial presentations was structured into four blocks of 50 randomly ordered trials each, with each block consisting of a unique permutation of every point light display. After the subjects completed all 200 trials, they were interviewed in order to gain insight into their techniques. The human data were analyzed and will be presented collapsed across subjects, scales and rotations. The results are presented in normalized bar plots meant to represent a probability distribution over k, based on the number of responses at each particular k. For each display there are at least 72 responses represented, assuming one answer per subject per trial. The actual number of responses might be larger if subjects were inclined to give multiple answers.

The 50 point light displays used in this experiment were chosen to provide a mixture of depth and breadth within the extremely large space of possible point light displays. Sixteen of the displays consisted of various riffs on mixtures of Gaussians, while another three were mixtures of Gaussians overlaid with uniformly distributed random noise. Nine displays consisted solely of uniformly distributed random noise (with differing number of samples between eight and 10,000). Three displays depicted two-dimensional embeddings of real data. Eight displays contained lines, circles or a combination of the two. The final 11 displays consisted of other synthetic data transformed by a variety of nonlinear distortions. See Fig. 1 for thumbnails of all the displays used. Subjects always saw the

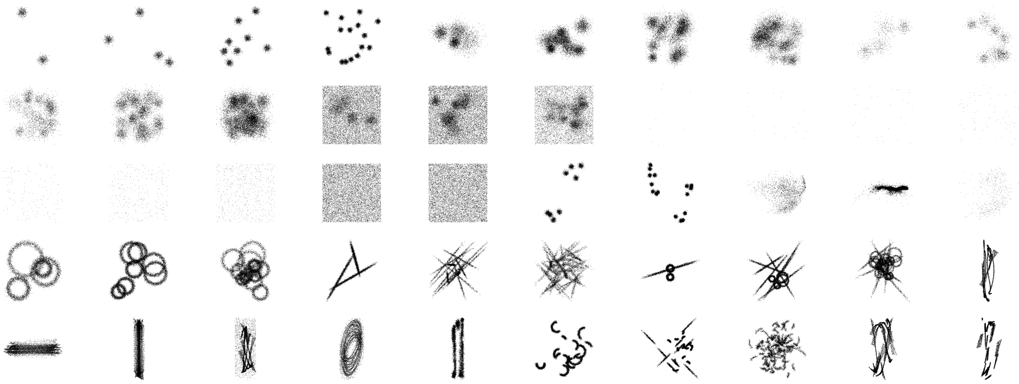

Figure 1: The stimuli.

displays as white points on a black background, but for the sake of presentation the displays in this paper are black on white and the points have been increased in size.

We focused heavily on mixture of Gaussian data sets due to the prominence of the Gaussianity assumption in the machine learning literature [4][5][6]. We also used several data sets with uniform noise in order to investigate how subject responses varied with sample size and to what extent subjects saw patterns where none were justified by the underlying distribution. Our shape-based and distorted displays were included for breadth and represent a case where the data are drawn from no standard underlying distribution.

Though all of the data sets are two-dimensional, we anticipate that insights gained from this study will lead to algorithmic improvements even in high-dimensional spaces. Certain algorithms (such as the Eigengap algorithm discussed below) operate over affinity matrices that are insensitive to the underlying dimensionality of a data set. Thus, improvements in these algorithms as measured by similarity to human performance in two dimensions will likely scale to high-dimensional data.

2 Results

Several interesting trends emerge in the human responses. In the interview section of the study, subjects predominantly report two central strategies: looking for areas of greatest density, perhaps separated by empty space ($N = 13$), and counting shapes or blobs ($N = 11$). Many of those subjects report using both strategies ($N = 9$). The latter strategy can be interpreted as a model fitting strategy, where subjects see a collection (mixture) of objects (e.g. arcs or Gaussians) and then explicitly count the number of those objects regardless of overlapping density. Rarer strategies include grouping by shape orientation ($N = 1$), and grouping by shape type (if there are both circles and lines in a display, there are two groups, $N = 1$). Finally, one subject explicitly mentions a hierarchical strategy, where he or she searches for small clusters first, and then groups them into larger clusters.

In concordance with interview responses indicating two primary strategies, we find several examples of bimodal responses for displays where these two strategies would diverge. Some examples are shown in Fig. 2.

3 Strategies

Based on the observations discussed above, humans follow at least two broad strategies when choosing k, density strategies and model fitting strategies. In this section, two algorithms from recent work in the field that represent these two strategies will be briefly described and their performance compared to the human data.

3.1 Density Strategies

Density strategies discover clusters by looking for regions of low density between groups of points, following density within groups to find all the points that belong to them, and attempting to ignore low density noise. A recent algorithm [8], which this paper will refer to as the Eigengap algorithm, brings these strategies for finding k under the spectral clustering umbrella. The Eigengap algorithm treats each data point as a node on a graph, and then performs a random walk between the nodes, with the probability of transitioning between any two nodes weighted by the distance between them. If two nodes are close together then the probability of transitioning from one to the other will be high and if two nodes are far apart then the probability of transitioning from one to the other will be low. Thus, if a group of points is separated by a large distance from the rest of the data, a random walk will be unlikely to transition across that gap. In this case, all the points within the group will have a high probability of ending up on other points in the group and little probability of ending up outside the group.

A matrix, P, representing the probability of any point ending up at any other point in the data set will therefore be block diagonal if there are distinct groups within the data set that are separated by sufficient distance. This block diagonal structure is represented by the n largest eigenvalues of P, and eigenvalues greater than the nth will generally be much smaller than the first n eigenvalues. By finding the largest difference between neighboring eigenvalues sorted in descending order, one can find a useful estimate of the number of groups in the data. For example, if the difference between the third and fourth eigenvalues is 0.4 and that distance is greater than the distance between all other adjacent eigenvalues, then there are likely to be three groups in the data.

As the random walk progresses the Eigengap algorithm naturally finds groups of coarser and coarser structure. Over an increasing number of steps, a random walk will become more and more likely to cross over low density sections of the data set, and thus two groups that initially might be separated will over time merge and lower values for k will be discovered. In this way the Eigengap algorithm can respond well to hierarchical data given a sufficiently long random walk.

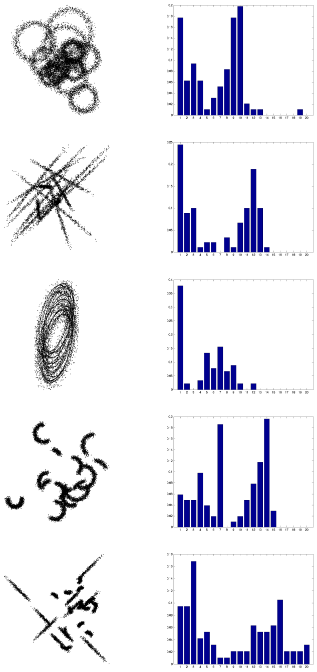

Figure 2: A sample of displays that elicited bimodal responses from subjects.

3.2 Model Fitting Strategies

Several model fitting strategies based on an assumption of mixture of Gaussian distributed data have been proposed in the past [4] [5]. This section describes a recent variant called PG-means [6]. PG-means searches for Gaussian clusters in a data set using an iterative process. The algorithm is initialized with $k = 1$ and it attempts to find an appropriate centroid and covariance matrix for a single Gaussian cluster given the data using the Expectation-Maximization (EM) algorithm. PG-means then randomly projects the data set and the Gaussian model down to one dimension n times (we used $n = 10$). The Kolmogorov-Smirnov (KS) test is applied to each projection and if every KS test indicates a sufficiently good fit (as measured by a parameter α that was set to 0.001) then the current value for k is accepted. Otherwise, k is incremented by one and the entire process is repeated.

If PG-means did not find an answer less than $k = 20$, the algorithm was halted and its response considered to be $k = 1$. Note that unlike the Eigengap algorithm, PG-means will only give one possible value for k.

3.3 Comparison with Human Data

To broadly compare Eigengap and PG-means performance with human performance, both the human results and the algorithmic results are interpreted as probability distributions over k. The sum Kullback-Leibler (KL) divergence is then calculated between the human results and both Eigengap and PG-means over all 50 data sets. The human results are considered the true distribution and the algorithmic results are considered the model distributions for purposes of calculating KL.

Unsurprisingly, given its ability to return multiple values of k and discover hierarchical organization, Eigengap outperforms PG-means with a sum KL divergence of 269.1 compared to 316.2 for PG-means. A simple unweighted combination of the two, however, performs better than either algorithm on its own with a sum KL divergence of 245.8 (an improvement of 8.7 percent over the Eigengap algorithm). See Fig. 3 for some sample comparisons of this combined result to human responses.

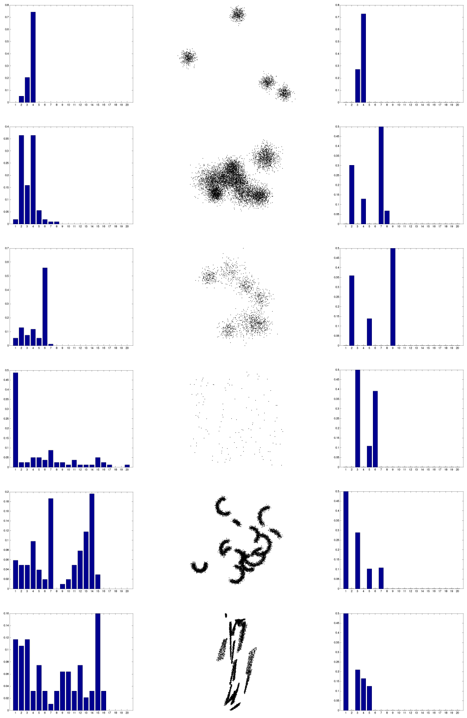

Figure 3: Sample human (left) versus combined Eigengap and PG-means (right) probability distributions over k.

4 Conclusion

Finding reasonable values for k is an important and difficult problem in unsupervised machine learning. As one can see from the samples in Fig. 3, current algorithms do well in certain situations and very poorly in others. By further investigating human performance and attempting to apply the insights garnered from such investigation, substantial progress can be made in developing new algorithms to tackle this thorny problem.

References

[1] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, Cambridge, MA, 2002.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[3] J.M. Santos and J. Marques de Sá. Human clustering on bi-dimensional data: An assessment. Technical Report 1, INEB Instituto de Engenharia Biomédica, Porto, Portugal, 2005.

[4] Dan Pelleg and Andrew Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proc. 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, San Francisco, CA, 2000.

[5] Greg Hamerly and Charles Elkan. Learning the k in k-means. In *Advances in Neural Information Processing Systems*, volume 17, 2003.

[6] Yu Feng and Greg Hamerly. Pg-means: learning the number of clusters in data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 393–400. MIT Press, Cambridge, MA, 2007.

[7] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[8] A. Azran and Z. Ghahramani. Spectral methods for automatic multiscale data clustering. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:190–197, 17-22 June 2006.