Information based clustering - a principled approach for cluster analysis

Noam Slonim, Gurinder Singh Atwal, Gašper Tkačik, and William Bialek

Joseph Henry Laboratories of Physics, and Lewis–Sigler Institute for Integrative Genomics
Princeton University, Princeton, New Jersey 08544
{nslonim, gatwal, gtkacik, wbialek}@princeton.edu

Theoretical Formulation. Imagine that there are N elements and N_c clusters and that we have assigned elements i to clusters C according to some probabilistic rules, $P(C|\mathrm{i})$. If we reach into a cluster and pull out elements at random, we would like these elements to be as similar to one another as possible. Similarity usually is defined among pairs of elements but as noted below we also can construct more collective measures of similarity among $r > 2$ elements; perhaps surprisingly we will see that that this more general case can be analyzed at no extra cost. Leaving aside for the moment the question of how to measure similarity, let us assume that computing the similarity among r elements $\mathrm{i}_1, \mathrm{i}_2, \cdots, \mathrm{i}_r$ returns a similarity measure $s(\mathrm{i}_1, \mathrm{i}_2, \cdots, \mathrm{i}_r)$. The average similarity among elements chosen out of a single cluster is

$$s(C) = \sum_{\mathrm{i}_1=1}^{N} \sum_{\mathrm{i}_2=1}^{N} \cdots \sum_{\mathrm{i}_r=1}^{N} P(\mathrm{i}_1|C)P(\mathrm{i}_2|C)\cdots P(\mathrm{i}_r|C)s(\mathrm{i}_1, \mathrm{i}_2, \cdots, \mathrm{i}_r), \tag{1}$$

where from Bayes' rule we have $P(\mathrm{i}|C) = P(C|\mathrm{i})P(\mathrm{i})/P(C)$, where $P(C) = \sum_{\mathrm{i}} P(C|\mathrm{i})P(\mathrm{i})$, and we further assume for simplicity, $P(\mathrm{i}) = 1/N$, although it is not essential. The intuition about the "goodness" of the clustering is expressed through the average similarity over all the clusters,

$$\langle s \rangle = \sum_{C=1}^{N_c} P(C)s(C). \tag{2}$$

The task then is to choose the assignment rules $P(C|\mathrm{i})$ that maximize $\langle s \rangle$, while, as in rate–distortion theory (1,2), simultaneously compressing our description of points as much as possible. To impose this we maximize $\langle s \rangle$ while constraining the information carried by the cluster identities,

$$I(C; \mathrm{i}) = \frac{1}{N} \sum_{\mathrm{i}=1}^{N} \sum_{C=1}^{N_c} P(C|\mathrm{i}) \log \left[\frac{P(C|\mathrm{i})}{P(C)} \right]. \tag{3}$$

Thus, our mathematical formulation of the intuitive clustering problem is to maximize the functional

$$\mathcal{F} = \langle s \rangle - T I(C; \mathrm{i}), \tag{4}$$

where the Lagrange multiplier T enforces the constraint on $I(C;\mathrm{i})$. Notice that, as in other formulations of the clustering problem, $\mathcal{F}$ resembles the free energy in statistical mechanics, where the temperature T specifies the tradeoff between energy and entropy like terms.

This formulation is intimately related to conventional rate–distortion theory. In rate–distortion clustering one is given a fixed number of bits with which to describe the data, and the goal is to use these bits so as to minimize the distortion between the data elements and some representatives of these data. In practice the bits specify membership in a cluster, and the representatives are prototypical or average patterns in each cluster. Here we see that we can formulate a similar tradeoff with no need to introduce the notion of a representative or average; instead, we measure directly the similarity of elements within each cluster; moreover, we can consider collective rather than pairwise measures of similarity.

Optimal Solution. Equating to zero the derivative of $\mathcal{F}$ with respect to the variables $P(C|\mathrm{i})$, yields after little algebra a set of implicit, self–consistent equations that any optimal solution must obey:

$$P(C|\mathrm{i}) = \frac{P(C)}{Z(\mathrm{i};T)} \exp\left\{\frac{1}{T}[rs(C;\mathrm{i}) - (r-1)s(C)]\right\}, \tag{5}$$

where $Z(\mathrm{i};T)$ is a normalization constant and $s(C;\mathrm{i})$ is the expected similarity between i and $r-1$ members of cluster C, $s(C;\mathrm{i}) = \sum_{\mathrm{i}_1=1}^{N} \cdots \sum_{\mathrm{i}_{r-1}=1}^{N} P(\mathrm{i}_1|C) \cdots P(\mathrm{i}_{r-1}|C) s(\mathrm{i}_1, \cdots, \mathrm{i}_{r-1}, \mathrm{i})$.

An obvious feature of Eq. (5) is that element i should be assigned to cluster C with higher probability if it is more similar to the other elements in the cluster. Less obvious is that this similarity has to be weighed against the mean similarity among all the elements in the cluster. Thus, our formulation automatically embodies the intuitive principle that "tightly knit" groups are more difficult to join.

The probability $P(C|\mathrm{i})$ in Eq. (5) has the form of a Boltzmann distribution, and increasing similarity among elements of a cluster plays the role of lowering the energy; the temperature T sets the scale for converting similarity differences into probabilities. As we lower this temperature there are a sequence of "phase transitions" to solutions with more distinct clusters that achieve greater mean similarity in each cluster. For a fixed number of clusters, reducing the temperature yields more deterministic $P(C|\mathrm{i})$ assignments.

Although Eq. (5) is an implicit set of equations, we can turn this self–consistency condition into an iterative algorithm that finds an explicit numerical solution for $P(C|\mathrm{i})$ that corresponds to a (perhaps local) maximum of $\mathcal{F}$. See (3) for details.

Information as a Similarity Measure. So far, in our formulation, the similarity measure s remained arbitrary. Is it possible to use the generality of information theory to address this issue as well? To be concrete, consider the case where the elements i are genes and we are trying to measure the relation between gene expression patterns across a variety of conditions $\mu = 1, 2, \cdots, M$ where gene i has expression level $e_{\mathrm{i}}(\mu)$ under condition μ. Then, for each gene we can define the probability density of expression levels, $P_{\rm i}(e) = \frac{1}{M}\sum_{\mu=1}^{M}\delta(e - e_{\rm i}(\mu))$, which should become smooth as $M \to \infty$. Similarly we can define the joint probability density for the expression levels of r genes $i_1, i_2, \cdots, i_r$,

$$P_{\rm i_1 i_2 \cdots i_r}(e_1, e_2, \cdots, e_r) = \frac{1}{M}\sum_{\mu=1}^{M}\delta(e_1 - e_{\rm i_1}(\mu))\delta(e_2 - e_{\rm i_2}(\mu))\cdots\delta(e_r - e_{\rm i_r}(\mu)). \quad (6)$$

Given this joint distribution, information theory provides natural measures of relatedness. For $r = 2$, we consider the mutual information between the expression levels,

$$s(\rm i,j) = I_{\rm i,j} = \int de_1 \int de_2\, P_{\rm ij}(e_1, e_2)\log_2\left[\frac{P_{\rm ij}(e_1, e_2)}{P_{\rm i}(e_1)P_{\rm j}(e_2)}\right] \text{ bits}, \quad (7)$$

and this is naturally extended to measure the multi–information among multiple variables, or genes: $I^{(r)}_{\rm i_1, i_2, \cdots, i_r} = \int d^r e\, P_{\rm i_1 i_2 \cdots i_r}(e_1, e_2, \cdots, e_r)\log_2\left[\frac{P_{\rm i_1 i_2 \cdots i_r}(e_1, e_2, \cdots, e_r)}{P_{\rm i_1}(e_1)P_{\rm i_2}(e_2)\cdots P_{\rm i_r}(e_r)}\right]$ bits.

We recall that the mutual information is the *unique* measure of relatedness between a pair of variables that obeys several simple and desirable requirements independent of assumptions about the form of the underlying probability distributions (1). In particular, the mutual (and multi–) information are independent of invertible transformations on the individual variables. For example, the mutual information between the expression levels of two genes is identical to the mutual information between the log of the expression levels: there is no need to find the "right" variables with which to represent the data. The absolute scale of the information measure also has a clear meaning. For example, if two genes share more than one bit of information then the underlying biological mechanisms must be more subtle than just turning expression on and off. In addition, the mutual information reflects any type of dependence among variables while ordinary correlation measures typically ignore nonlinear dependences.

While these theoretical advantages are well known, in practice information theoretic quantities are notoriously difficult to estimate from finite data. Although there is no general solution to these problems, we have found that in practice the difficulties are not as serious as one might have expected. Using an adaptation of the "direct" estimation method (4), we have found that one can obtain reliable estimates of mutual (and sometimes multi–) information values for a variety of data types, including gene expression data (5).

Summary. We have suggested a purely information theoretic approach to clustering and categorization: relatedness among elements is defined by the mutual (or multi–) information, and optimal clustering is defined as the best tradeoff between maximizing this average relatedness within clusters and minimizing the number of bits required to describe the data. The result is a formulation of clustering that trades bits of similarity against bits of descriptive power, with no further assumptions.

Most clustering algorithms embody different models of the underlying statistical structure. In principle, more accurate models should lead to more meaningful clusters. However, the question of how to construct an accurate model obviously is quite involved, raising further issues that often are addressed arbitrarily before the cluster analysis begins. Moreover, an algorithm or model which is successful in one data type might fail completely in a different domain. Our use of information theory allows us to capture the relatedness of different patterns independent of assumptions about the nature of this relatedness. Correspondingly, we have a single approach which achieves high performance across different domains (3).

Our approach can further succeed where other methods would fail qualitatively. Conventional algorithms search for linear or approximately linear relations among the different variables, while our information theoretic approach is responsive to any type of dependencies, including strongly nonlinear structures. In addition, while much of the cluster analysis literature has focused thus far on pairwise relations and similarity measures, our approach sets a sound theoretical framework for analyzing complex data based on higher order relations. Indeed, it was recently demonstrated that in some situations the data structure is obscured at the pairwise level, but clearly manifests itself only at higher levels (6,7). The question of how common such data are, as well as the associated computational difficulties in analyzing such higher order relations, is yet to be explored.

1. Shannon, C. E. (1948) *The Bell Sys Tech J* **27**, 379–423 & 623–656.

2. Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory* (John Wiley and Sons, New York).

3. Slonim, N., Atwal, G. S., Tkačik, G. & Bialek W. (2005). Information based clustering. *Proc Nat Acad Sci (USA)* in press. See also *http://www.genomics.princeton.edu/biophysics-theory/Clustering/web-content/index.html*.

4. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. (1998) *Phys Rev Lett* **80**, 197–200.

5. Slonim, N., Atwal, G. S., Tkačik, G. & Bialek, W. (2005) *http://arxiv.org/abs/cs.IT/0502017*.

6. Schneidman, E., Still, S., Berry II, M. J. & Bialek, W. (2003) *Phys Rev Lett* **91**, 238701(4).

7. Bowers, P. M., Cokus, S. J., Eisenberg, D. & Yeates, T. O. (2004) *Science* **306**, 2246–2249.