Associative Clustering (AC): Technical Details

Janne Sinkkonen, Samuel Kaski, Janne Nikkilä, Leo Lahti

Abstract

This report contains derivations which did not fit into the paper [3]. Associative clustering (AC) is a method for separately clustering two data sets when one-to-one associations between the sets, implying statistical dependency, are available. AC finds Voronoi partitionings that maximize the visibility of the dependency on the cluster level. The main content of this paper are technical results related to the algorithm: A Bayes factor interpretation of AC, derivation of gradients for optimizing AC with a smoothing trick, and the connection of AC objective to mutual information.

I. Introduction

The abstract clustering task solved by associative clustering [3], [6] is the following: cluster two sets of data, with samples $\mathbf{x}$ and $\mathbf{y}$, each separately, such that (i) the clusterings would capture as much as possible of the dependencies within data pairs $(\mathbf{x}, \mathbf{y})$, and (ii) the clusters would contain (relatively) similar data points. The latter is roughly a definition of a cluster.

Figure 1 gives a brief overview of the method. For paired data $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ of real vectors $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, we search for partitionings $\{V_i^{(x)}\}$ for $\mathbf{x}$ and $\{V_j^{(y)}\}$ for $\mathbf{y}$. The partitions can be interpreted as clusters in the same way as in K-means; they are Voronoi regions parameterized by their prototype vectors $\mathbf{m}_i^{(x)}$: $\mathbf{x} \in V_i^{(x)}$ if $\|\mathbf{x} - \mathbf{m}_i^{(x)}\| \leq \|\mathbf{x} - \mathbf{m}_{i'}^{(x)}\|$ for all i', and correspondingly for $\mathbf{y}$.

II. Derivation of the Bayes factor for AC

Given a paired data set $\{(x_r, y_r)\}_r$, $x_r \in \mathcal{X}$, $y_r \in \mathcal{Y}$, and two parameterized families of clusterings, $f_x(x; \theta_x) : \mathcal{X} \to \{1, 2, \ldots, K\}$, and $f_y(y; \theta_y) : \mathcal{Y} \to \{1, 2, \ldots, L\}$, the goal of AC is to find parameters of cluster solutions, θ_x and θ_y, such that a measure of dependency between the cluster indices $\{f_x(x_r; \theta_x)\}_r$

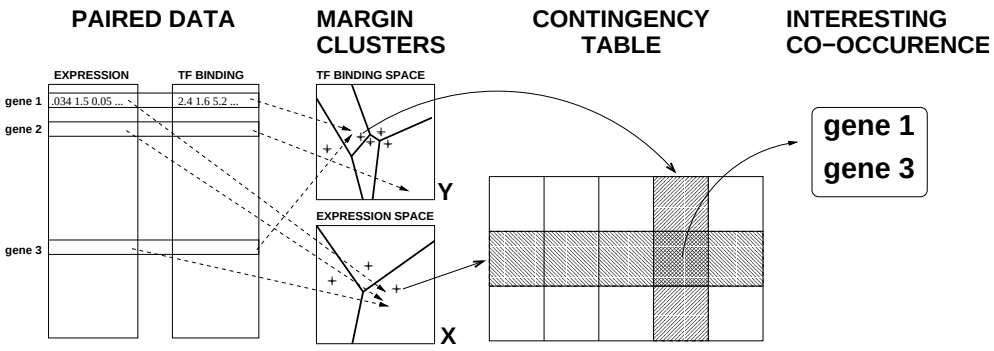

Fig. 1. Associative clustering in a nutshell. Two data sets are clustered into Voronoi regions. The Voronoi regions are defined in the standard way as sets of points closest to prototype vectors, but the prototypes are not chosen by minimizing a quantization error but by other means described in the text. In this example, the data sets are gene expression profiles and TF binding profiles. A one-to-one correspondence between the sets exist: Each gene has an expression profile and a TF binding profile. As each gene falls to a TF-Voronoi cluster and to an expression cluster, we get a contingency table by placing the two sets of clusters as rows and columns, and by counting genes falling to each combination of an expression and a TF cluster. Rows and columns, that is, the Voronoi regions defined within one data set, are consequently called *margin clusters*, while the combinations corresponding to the cells of the contingency table are called *cross clusters*. *Associative clustering* by definition finds Voronoi prototypes that maximize the dependency seen in the contingency table. Voronoi regions are representations for the data sets just as the linear combinations in canonical correlation analysis. In both cases, dependency between the two parametrized representations is maximized. Maximization of dependency in a contingency table results in a maximal amount of surprises, gene counts in cells not explainable by the margin distributions. The most surprising clusters with very high or low number of genes give rise to interesting interpretations. Reliability can be assessed by the bootstap.

and $\{f_y(y_r;\theta_y)\}_r$ is maximized. We will later assume f_x and f_y to be Voronoi partitionings parameterized by their the Voronoi prototypes, and $\mathcal{X}$ and $\mathcal{Y}$ to be real spaces. This section, however, derives a general form for the Bayes factor for which the parameterization of the clusters is irrelevant.

Because samples over r are assumed independent, a suffi cient statistics for cluster indices over the whole data set, $\{f_x(x_r;\theta_x), f_y(y_r;\theta_y)\}_r$, are just the counts of different combinations. Therefore, denote by n_{ij} the number of samples belonging to the cluster i in space $\mathcal{X}$ and to the cluster j in space $\mathcal{Y}$, and denote $n_{i\cdot} = \sum_j n_{ij}$, $n_{\cdot j} = \sum_i n_{ij}$.

The counts $\{n_{ij}\}_{ij}$ form a contingency table. We will measure dependency by the Bayes factor between likelihood for dependent margins (hypothesis M_D) and likelihood for independent margins (M_I):

$$BF = \frac{P(\{n_{ij}\}|M_D)}{P(\{n_{ij}\}|M_I)} \ . \tag{1}$$

The following result presented, e.g., by Good, 1976, is utilized [2]. The posterior probability of observing B-bin multinomial counts t_s with the sum $T = \sum_s t_s$, given a Dirichlet prior $p(\boldsymbol{\theta}) \propto \prod_s \theta_s^{t^0-1}$, is

$$P(\{t_s\}) = \int P(\{t_s\}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{\Gamma(Bt^0)\, T! \prod_s \Gamma(t_s + t^0)}{\Gamma(t^0)^B\, \Gamma(T + Bt^0) \prod_s t_s!} \ . \tag{2}$$

The hypothesis of dependent margins corresponds to the assumption of a Dirichlet prior (with n^d "prior data points" in each bin) over the whole contingency table. The numerator of (1) then comes directly from (2).

In the denominator, the hypothesis of independent margins, independent Dirichlet priors are assumed for the margins, with parameters $n^{(x)}$ and $n^{(y)}$. Under the hypothesis of independence, the prior over the cells of the contingency table is a product of the margin priors. For the denominator of (1), then,

$$\begin{aligned} P(\{n_{ij}\}|M_I) &= P(\{n_{ij}\}, \{n_{\cdot j}\}, \{n_{i\cdot}\}, M_I) \\ &= P(\{n_{ij}\}|\{n_{\cdot j}\}, \{n_{i\cdot}\}, M_I)\, P(\{n_{\cdot j}\}, \{n_{i\cdot}\}|M_I) \\ &= P(\{n_{ij}\}|\{n_{\cdot j}\}, \{n_{i\cdot}\}, M_I)\, P(\{n_{\cdot j}\}|M_I)\, P(\{n_{i\cdot}\}|M_I) \ , \end{aligned}$$

where the fi rst term comes from the hypergeometric distribution and the two last terms from (2).

With these priors, the whole Bayes factor becomes

$$\begin{aligned} BF &= P(\{n_{ij}\}|M_D) \times 1/P(\{n_{ij}\}|\{n_{\cdot j}\}, \{n_{i\cdot}\}, M_I) \times 1/P(\{n_{\cdot j}\}|M_I) \times 1/P(\{n_{i\cdot}\}|M_I) \\ &\propto \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_{ij} n_{ij}!} \times \frac{N! \prod_{ij} n_{ij}!}{\prod_j n_{\cdot j}! \prod_i n_{i\cdot}!} \times \frac{\prod_j n_{\cdot j}!}{\prod_j \Gamma(n_{\cdot j} + n^{(y)})} \frac{\prod_i n_{i\cdot}!}{\prod_i \Gamma(n_{i\cdot} + n^{(x)})} \\ &\propto \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_j \Gamma(n_{\cdot j} + n^{(y)}) \prod_i \Gamma(n_{i\cdot} + n^{(x)})} \ . \end{aligned}$$

Constants due to priors, fi xed bin number and N are omitted.

In summary, a suitable cost function for maximizing the dependence in the contingency table is

$$BF \propto \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_i \Gamma(n_{i\cdot} + n^{(x)}) \prod_j \Gamma(n_{\cdot j} + n^{(y)})} \ . \tag{3}$$

The parameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arise from Dirichlet priors. If all prior parameters are set to unity, *BF* becomes equivalent to the hypergeometric probability classically used as a dependency measure of contingency tables. In the limit of large data sets, (3) becomes mutual information of the margins (Section IV).

III. Derivation of the gradients of the log Bayes factor

Optimizing *BF* with respect to Voronoi region centroids determining the counts $n^{(\cdot)}$ is hard, for the *BF* is not a continuous function of the centroid vectors. In AC, a *smoothed version* of *BF* is optimized with respect to the parameters $\{\mathbf{m}^{(x)}\}$ and $\{\mathbf{m}^{(y)}\}$ by a conjugate-gradient algorithm (for a textbook account see [1]).

To start, from (3) one obtains the extended and log-transformed cost function

$$\log \textit{BF'} = \sum_{ij} \log\Gamma(n_{ij} + n^{(d)}) - \lambda^{(y)} \sum_{j} \log\Gamma(n_{\cdot j} + n^{(y)}) - \lambda^{(x)} \sum_{i} \log\Gamma(n_{i\cdot} + n^{(x)})$$

with $\log \textit{BF'} = \log \textit{BF} + \textit{const.}$ if $\lambda^{(\cdot)} = 1$. Values $\lambda^{(\cdot)} > 1$ are used for regularization. Rationale for regularization is discussed elsewhere [3] (but see also [4], [5] for more thorough discussion and testing in the one-margin case).

For shortness of notation, we will denote data samples that always appear inside sums by $\mathbf{x}, \mathbf{y}$ instead of $\mathbf{x}_r, \mathbf{y}_r$, and 'index' in the closing summation simply by $\mathbf{x}$ and $\mathbf{y}$. It is now assumed that $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}, \mathbf{y} \in \mathcal{Y} \subset \mathbb{R}$.

As an introduction to smoothing, we may think the counts $\{n_{ij}\}$ are produced as sums over *indicator functions* $g_i^{(x)}(\mathbf{x})$ and $g_j^{(y)}(\mathbf{y})$:

$$\begin{aligned} n_{ij} &= \sum_{(\mathbf{x},\mathbf{y})} g_i^{(x)}(\mathbf{x}) g_j^{(y)}(\mathbf{y}) \\ n_{\cdot j} &= \sum_{(\mathbf{x},\mathbf{y})} g_i^{(x)}(\mathbf{x}) \sum_{j} g_j^{(y)}(\mathbf{y}) = \sum_{\mathbf{x}} g_i^{(x)}(\mathbf{x}) \\ n_{i\cdot} &= \sum_{\mathbf{y}} g_j^{(y)}(\mathbf{y}) \end{aligned}$$

Here $g_i^{(x)}(\mathbf{x}) = 1$ if the sample $\mathbf{x}$ falls into cluster j, and zero otherwise. $g_j^{(y)}(\mathbf{y})$ is defined analogously for $\mathbf{y}$. Note that the clusters in $\mathcal{X}$ are indexed by i or i', and the clusters in $\mathcal{Y}$ by j.

The indicator functions are (here implicitly) parametrized by the Voronoi prototypes: $g_i^{(x)}(\mathbf{x}) = 1$ if the prototype with index i is the closest of all prototypes in $\mathcal{X}$ to the sample $\mathbf{x}$. It is clear that the values of the indicator functions are mostly constant but change noncontinuously as a functions of locations of prototypes: when a sample $\mathbf{x}$ crosses the Voronoi region border from region i to region k, $g_i^{(x)}(\mathbf{x})$ changes abruptly from one to zero, and $g_{i'}^{(x)}(\mathbf{x})$ does the opposite. The gradient of $\{g_i^{(x)}(\mathbf{x})\}$ or $\{g_j^{(y)}(\mathbf{y})\}$ with respect to the Voronoi prototype is therefore almost always zero, and sometimes does not exist. The same then holds for the gradient of the Bayes factor $\log \textit{BF'}$, which renders conventional nonlinear optimization methods unusable.

To get good gradients, we extend the concept of indicator functions such that $g_i^{(x)}(\mathbf{x}) \in [0, 1]$ (analogously for y). At the limit $\sigma_{(x)} \to 0$ of the smoothing parameter $\sigma_{(x)}$, the original indicator functions and therefore original, non-smoothed Voronoi regions are obtained. Specifically, we set

$$\begin{aligned} g_i^{(x)}(\mathbf{x}) &= Z^{(x)}(\mathbf{x})^{-1} e^{-\|\mathbf{x}-\mathbf{m}_i^{(x)}\|^2/\sigma_{(x)}^2} , \\ g_j^{(y)}(\mathbf{y}) &= Z^{(y)}(\mathbf{y})^{-1} e^{-\|\mathbf{y}-\mathbf{m}_j^{(y)}\|^2/\sigma_{(y)}^2} . \end{aligned}$$

with $Z^{(x)}$ and $Z^{(y)}$ such that $\sum_j g_j^{(y)}(\mathbf{x}) = \sum_i g_i^{(x)}(\mathbf{y}) = 1$.

Denote for brevity $t_{ij} = n_{ij} + n^{(d)}$. The gradient of the cost function with respect to the Voronoi centers

$\mathbf{m}_i^{(x)}$ of space $\mathcal{X}$ is

$$\begin{aligned}\nabla_{\mathbf{m}_i^{(x)}} \log BF' &= \sum_{i'j} \Psi(t_{i'j}) \sum_{(\mathbf{x},\mathbf{y})} g_j^{(y)}(\mathbf{y}) \nabla_{\mathbf{m}_i^{(x)}} g_{i'}^{(x)}(\mathbf{x}) - \lambda^{(x)} \sum_{\mathbf{x},i'} \Psi(t_{k\cdot}) \nabla_{\mathbf{m}_i^{(x)}} g_{i'}^{(x)}(\mathbf{x}) \\ &= \sum_{(\mathbf{x},\mathbf{y}),i'} \left[\sum_j \Psi(t_{i'j}) g_j^{(y)}(\mathbf{y}) - \lambda^{(x)} \Psi(t_{i'\cdot}) \right] \nabla_{\mathbf{m}_i^{(x)}} g_{i'}^{(x)}(\mathbf{x}) \ ,\end{aligned}$$

where $\Psi(\cdot)$ is the psi or digamma function, the derivative of $\log\Gamma(\cdot)$. The gradient of the smoothed indicator functions is

$$\nabla_{\mathbf{m}_i^{(x)}} g_{i'}^{(x)}(\mathbf{x}) = \frac{1}{\sigma^2}(\mathbf{x} - \mathbf{m}_i^{(x)})(\delta_{i'i} - g_{i'}^{(x)}(\mathbf{x})) g_i^{(x)}(\mathbf{x}) \ .$$

We will denote the original cost function $\log BF'$ extended by the smoothed indicators by $\log BF''$. (The original cost is obtained by setting $\sigma_{(x)} \to 0, \sigma_{(y)} \to 0$.) Substituting the gradients of the smooth indicators and applying the algebraic identity

$$\sum_{i'} (\delta_{i'i} - y_{i'}) y_i L_{i'} = \sum_{i'} y_{i'} y_i (L_i - L_{i'})$$

gives

$$\begin{aligned}\sigma_{(x)}^2 \nabla_{\mathbf{m}_i^{(x)}} \log BF'' &= \sum_{(\mathbf{x},\mathbf{y}),i'} (\mathbf{x} - \mathbf{m}_i^{(x)})(\delta_{i'i} - g_{i'}^{(x)}(\mathbf{x})) g_i^{(x)}(\mathbf{x}) \left[\sum_j \Psi(t_{i'j}) g_j^{(y)}(\mathbf{y}) - \lambda^{(x)} \Psi(t_{i'\cdot}) \right] \\ &= \sum_{(\mathbf{x},\mathbf{y}),i'} (\mathbf{x} - \mathbf{m}_i^{(x)}) g_{i'}^{(x)}(\mathbf{x}) g_i^{(x)}(\mathbf{x}) (L_i^{(x)}(\mathbf{y}) - L_{i'}^{(x)}(\mathbf{y})) \ ,\end{aligned}$$

where

$$L_i^{(x)}(\mathbf{y}) = \sum_j \Psi(n_{ij} + n^{(d)}) g_j^{(y)}(\mathbf{y}) - \lambda^{(x)} \Psi(n_{i\cdot} + n^{(x)}) \ .$$

Since the cost function is symmetric with respect to the two spaces $\mathcal{X}$ and $\mathcal{Y}$, the gradient with respect to a cluster of $\mathcal{Y}$-space is obtained analogously.

IV. Connection to maximization of mutual information

Applying Stirling's approximation $\log\Gamma(x+1) = x\log(x) - x + O(\log(x))$ to the logarithmic Bayes factor

$$\frac{1}{N} \log BF = \frac{1}{N} \sum_{i,j} \log\Gamma(n_{ij} + n^{(d)}) - \frac{1}{N} \sum_i \log\Gamma(n_{i\cdot} + n^{(y)}) - \frac{1}{N} \sum_j \log\Gamma(n_{\cdot j} + n^{(x)}) \quad (4)$$

yields

$$\begin{aligned}\frac{1}{N} \log BF = &\frac{1}{N} \sum_{i,j} [(n_{ij} + a)\log(n_{ij} + a) - (n_{ij} + a) + O(\log(n_{ij} + a))] \\ &- \frac{1}{N} \sum_i [(n_{i\cdot} + b)\log(n_{i\cdot} + b) - (n_{i\cdot} + b) + O(\log(n_{i\cdot} + b))] \\ &- \frac{1}{N} \sum_j [(n_{\cdot j} + c)\log(n_{\cdot j} + c) - (n_{\cdot j} + c) + O(\log(n_{\cdot j} + c))] \quad (5)\end{aligned}$$

where $a = n^{(d)} - 1$, $b = n^{(x)} - 1$, and $c = n^{(y)} - 1$. Applying the algebraic identity $(x+y)\log(x+y) = x\log(x) + x\log(1+y/x) + y\log(x+y)$ further yields

$$\begin{aligned}\frac{1}{N}\log \mathit{BF} = &\frac{1}{N}\sum_{i,j}\left[n_{ij}\log n_{ij} + n_{ij}\log(1+\frac{a}{n_{ij}}) + a\log(n_{ij}+a) - (n_{ij}+a) + O(\log(n_{ij}+a))\right] \\ &- \frac{1}{N}\sum_{i}\left[n_{i\cdot}\log n_{i\cdot} + n_{i\cdot}\log(1+\frac{b}{n_{i\cdot}}) + b\log(n_{i\cdot}+b) - (n_{i\cdot}+b) + O(\log(n_{i\cdot}+b))\right] \\ &- \frac{1}{N}\sum_{j}\left[n_{\cdot j}\log n_{\cdot j} + n_{\cdot j}\log(1+\frac{c}{n_{\cdot j}}) + c\log(n_{\cdot j}+c) - (n_{\cdot j}+c) + O(\log(n_{\cdot j}+c))\right] \quad . \quad (6)\end{aligned}$$

With $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$, we may rewrite

$$\begin{aligned}\frac{1}{N}\log \mathit{BF} = &\sum_{i,j}\frac{n_{ij}}{N}\log\frac{n_{ij}}{n_{i\cdot}n_{\cdot j}} + 1 + \sum_{i,j}\frac{n_{ij}}{N}\log\frac{1+\frac{a}{n_{ij}}}{(1+\frac{b}{n_{i\cdot}})(1+\frac{c}{n_{\cdot j}})} \\ &+ \sum_{i,j}\left[\frac{a}{N}\log(n_{ij}+a) - \frac{a}{N} + O(\frac{1}{N}\log(n_{ij}+a))\right] \\ &- \sum_{i\cdot}\left[\frac{c}{N}\log(n_{i\cdot}+b) - \frac{b}{N} + O(\frac{1}{N}\log(n_{i\cdot}+b))\right] \\ &- \sum_{\cdot j}\left[\frac{c}{N}\log(n_{\cdot j}+c) - \frac{c}{N} + O(\frac{1}{N}\log(n_{\cdot j}+c))\right] \quad . \quad (7)\end{aligned}$$

On the basis of Taylor approximations of the type

$$\log\left(1+\frac{a}{n_{ij}}\right) = \frac{a}{n_{ij}} + O\left(\frac{a^2}{n_{ij}^2}\right)$$

the term

$$\sum_{i,j}\frac{n_{ij}}{N}\log\frac{1+\frac{a}{n_{ij}}}{(1+\frac{b}{n_{i\cdot}})(1+\frac{c}{n_{\cdot j}})}$$

of (7) is bounded by $O(N^{-1})$ and therefore also by $O(N^{-1}\log N)$. Likewise, because a, b, and c are constants and because $n_{ij} \leq N$, $n_{\cdot j} \leq N$, and $n_{i\cdot} < N$, all the rest of the terms are also bounded by $O(N^{-1}\log N)$. Therefore we have

$$\frac{1}{N}\log \mathit{BF} = \sum_{i,j} p_{ij}\log\frac{p_{ij}}{\hat{p}_i p_j} - \log N + 1 + O\left(\frac{1}{N}\log N\right) = \hat{I}(I,J) - \log N + 1 + O\left(\frac{1}{N}\log N\right) \quad , \quad (8)$$

that is, the logarithmic Bayes factor approaches mutual information of the distribution $p_{ij} = n_{ij}/N$ with the margins $p_i = n_{i\cdot}/N$ and $p_j = n_{\cdot j}/N$, plus a constant term.

ACKNOWLEDGMENT

This work has been supported by the Academy of Finland, decisions #79017 and #207467.

REFERENCES

[1] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley, New York, 1993.

[2] I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4:1159–1189, 1976.

[3] S. Kaski, J. Nikkilä, J. Sinkkonen, L. Lahti, J. Knuuttila, and C. Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE TCBB*, 2005. In press.

[4] S. Kaski, J. Sinkkonen, and A. Klami. Discriminative clustering. *Neurocomputing*, 2005. To appear.

[5] Samuel Kaski, Janne Sinkkonen, and Arto Klami. Regularized discriminative clustering. In Christophe Molina, Tülay Adali, Jan Larsen, Marc Van Hulle, Scott Douglas, and Jean Rouat, editors, *Neural Networks for Signal Processing XIII*, pages 289–298. IEEE, New York, NY, 2003.

[6] Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, pages 396–406. Springer, Berlin, 2004.