

IB-type clustering for continuous finite data

Janne Sinkkonen

and Sami Kaski, and Janne Nikkilä, Jaakko Peltonen, and others

Laboratory of Computer and Information Science
Helsinki University of Technology, Finland

`Janne.Sinkkonen@hut.fi`

`http://www.cis.hut.fi/projects/mi/`

December 7, 2003

We propose algorithms for 1) finite and 2) continuous data

Data comes as counts $n(v, w)$.

Mutual information is defined for densities $p(v, w)$, not for counts.

Replace mutual information by a **Bayes factor**.

Use **continuous data** instead of discrete categories (words, documents).

Parameterization of clusters as Voronoi regions.

IB-like cost (mutual information, Bayes factor).

Examples

Hard IB finds margins for contingency tables

IB: compromise between dependency $\beta I(V; W)$ and complexity $I(V, X)$.

Soft clusters with an “ $\exp -\beta D_{KL}$ shape” follow.

$\beta \rightarrow \infty$:

Dependency $\beta I(V; W)$ remains, complexity $I(V, X)$ disappears

Clusters become hard.

For hard clusters, IB becomes a contingency table algorithm.

Contingency table with adjustable margins enables IB and more

No soft clusters!

Counts $n(v, w)$.

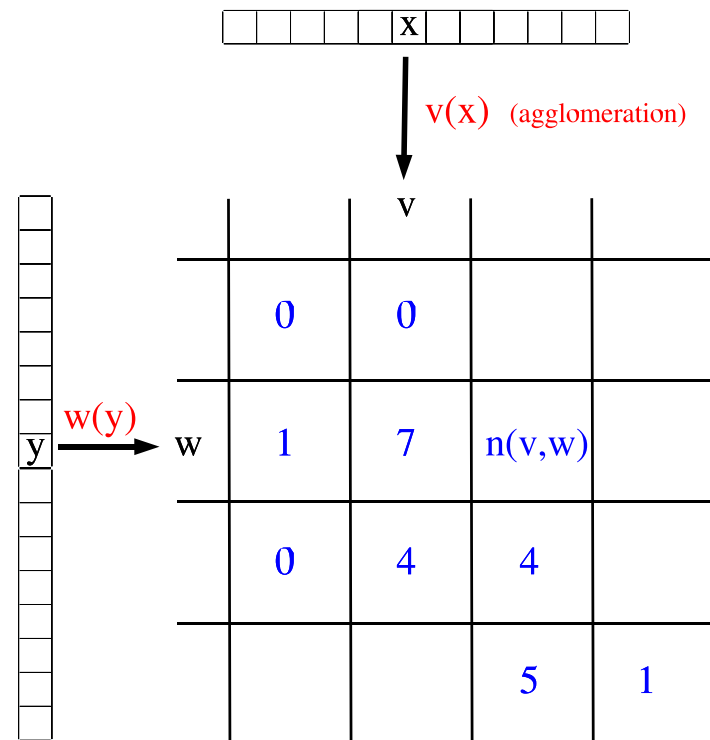
Representation for x : $v(x)$.

Representation for y : $w(y)$.

Dependency: $I(V, W)$ or something else.

Maximize dependency wrt.

$x \rightarrow v(x), y \rightarrow w(y)$.



Finding margins for contingency tables: **hard IB** vs. **extensions**

Optimize $I(V, W)$:

1. One-way, optimized by aggregation: **agglomerative IB**
2. Optimize by index shuffling: **sIB**
3. Two-way: **multivariate IB**

Change dependency measure: $I(X; V) \rightarrow$ **a Bayes factor (BF)**:

4. **finite-data** IB, sIB, multivariate IB

Bayes factor takes sampling uncertainty into account

Mutual information $I(V, W)$ ok, but actually defined for a known $p(v, w)$ only.

For $\hat{p}(v, w) = n(v, w) / \sum_{vw} n(v, w)$ MI becomes biased.

Are we measuring dependency or bias then?

Bayes factor is an alternative:

Defined for counts.

Bias less obvious (prior!).

Takes sampling uncertainty into account.

May improve results with small data sets.

Asymptotically still $I(V, W)$ for $n(v, w) \rightarrow \infty$.

BF compares the hypotheses of dependent vs. independent margins

Assume the counts $\{n_{vw}\}$ arise from a multinomial $\{\theta_{vw}\}$ over the table.

Dependent margins (\bar{H}): θ_{vw} are free, with a Dirichlet prior.

Independent margins (H): $\theta_{vw} = \theta_v \theta_w$, with Dirichlets for $\{\theta_v\}$, $\{\theta_w\}$.

Margins are functions of cluster assignments $v()$ and $w()$, hence BF is too:

$$BF(v(), w()) = \frac{P(\{n_{vw}\}|\bar{H})}{P(\{n_{vw}\}|H)} \text{ with } P(\{n_{vw}\}|\cdot) = \int_{\theta} p(\theta|\{n_{vw}\}, \cdot) p(\theta|\cdot) d\theta .$$

With the multinomial assumption and symmetric conjugate Dirichlet priors:

$$BF(v(), w()) = \frac{\prod_{vw} \Gamma(n_{vw} + n^0)}{\prod_v \Gamma(n_v + n^{0,x}) \prod_w \Gamma(n_w + n^{0,y})} .$$

Maximize BF with respect to cluster assignments $v()$, $w()$.

***BF* can replace $I(X; V)$ in the hard IB**

$BF \rightarrow I(V; W)$ when $n_{vw} \rightarrow \infty$.

sIB for finite data:

Replace the Jensen-Shannon assignment criterion by $p(doc|cluster)$.

Then the overall cost I becomes replaced by BF .

Overall, a nice analogy:

BF for large data: $\rightarrow I$

$p(doc|cluster)$ for large data: \rightarrow JS.

Difference between I 's: Jensen-Shannon

Difference between BF 's: $p(doc|cluster)$

Initial experiments: BF slightly better when only few documents per cluster.

Finding margins for contingency tables: IB vs. extensions

Optimize $I(V, W)$:

1. One-way, optimized by aggregation: **agglomerative IB**
2. Optimized by index shuffling: **sIB**
3. Two-way: **multivariate IB**

Change cost $I(X; V) \rightarrow$ a Bayes factor (BF):

4. finite-data IB, sIB, multivariate IB

$x \in \mathbb{R}^n$ and maybe $y \in \mathbb{R}^m$:

5. One-way: **discriminative clustering (DC)**
6. Two-way: **associative clustering (AC)**

Continuous x, y require parameterization

x and y nominal:

hard IB or something similar ok.

x and y from \mathbb{R}^n : parameterization needed.

Voronoi regions:

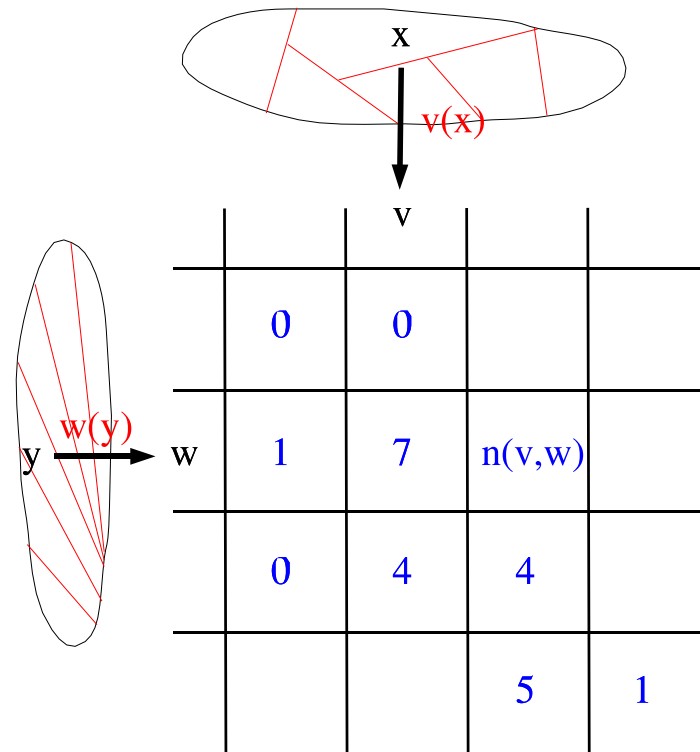
$v(x, \{m_v^x\}), w(y, \{m_w^y\})$

centroids $\{m_v^x\}$ and $\{m_w^y\}$

Optimize dependency

w.r.t. $\{m_v^x\}$ and $\{m_w^y\}$.

Dependency may be I or BF .



Optimization is easier with soft clusters

Problem for real data sets: no gradients w.r.t. $\{m_v^x\}$ and $\{m_w^y\}$.

Convert cluster assignments into a continuum (soft clusters):

$$y_v(x) = \frac{1}{Z(x)} \exp -\frac{1}{\sigma^2} \|x - m\|^2 .$$

Hard clusters are obtained by $\sigma \rightarrow 0$.

Looks like the cluster shape obtained by IB, but is not!

Euclidean instead of KL.

The IB shape is not obtainable.

IB shape would be a function of $D_{KL}(p(w|x)||p(w|v))$,
with $p(w|x)$ unknown ($x \in \mathbb{R}^n$).

With density estimators $\hat{p}(w|x)$: yes, but not very elegant.

$I(V, W)$ as the dependency leads to a simple algorithm

Maximize $I(V; W)$ w.r.t. $v()$ and $w()$ or more concretely $\{m_v^x\}, \{m_w^y\}$.

Apply the **cluster smoothing trick**.

A simple on-line algorithm follows:

Draw (x, y) . Assign $y \rightarrow w$ by $w(y)$.

Draw neighbourhood partitions (v, v') on the basis of $\{y_v(x)\}$.

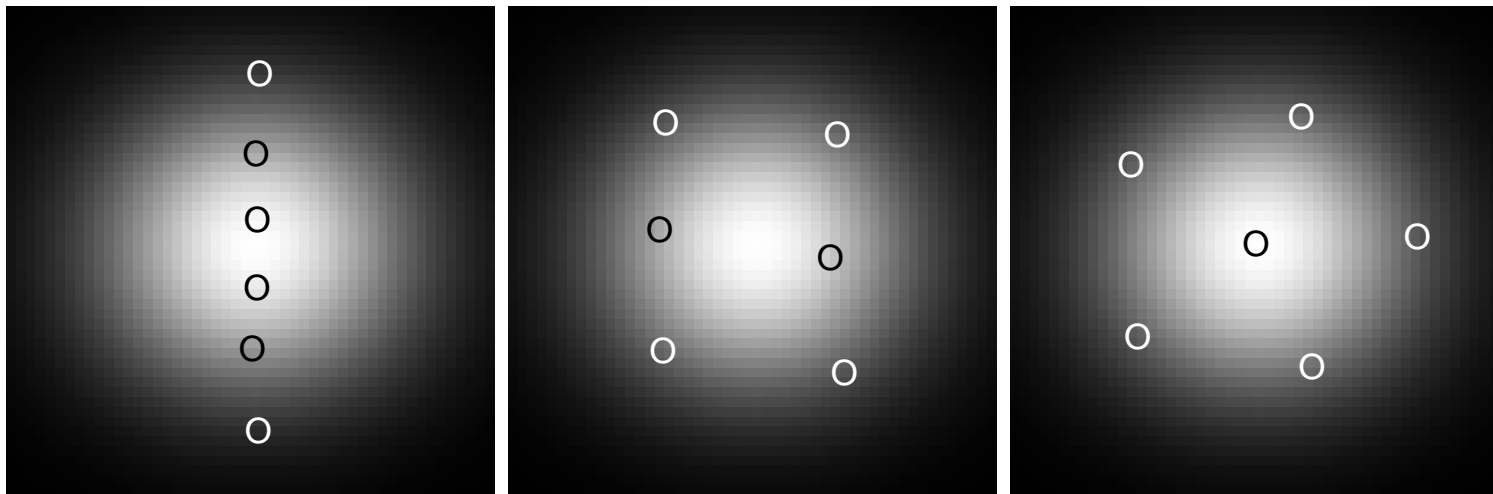
Update the Voronoi centroid:

$$\Delta m_v \propto (m_v - x) \log \frac{p(w|v)}{p(w|v')}$$

Looks much like **VQ, LVQ**.

For one fixed margin: works, but sensitive to **smoothing** (σ).

Example: toy data, one-margin case



Simple yes, but *BF* as dependency provides better results

Maximize $BF(v(), w())$ w.r.t. $v()$ and $w()$ or actually $\{m_v^x\}, \{m_w^y\}$.

No on-line algorithm follows.

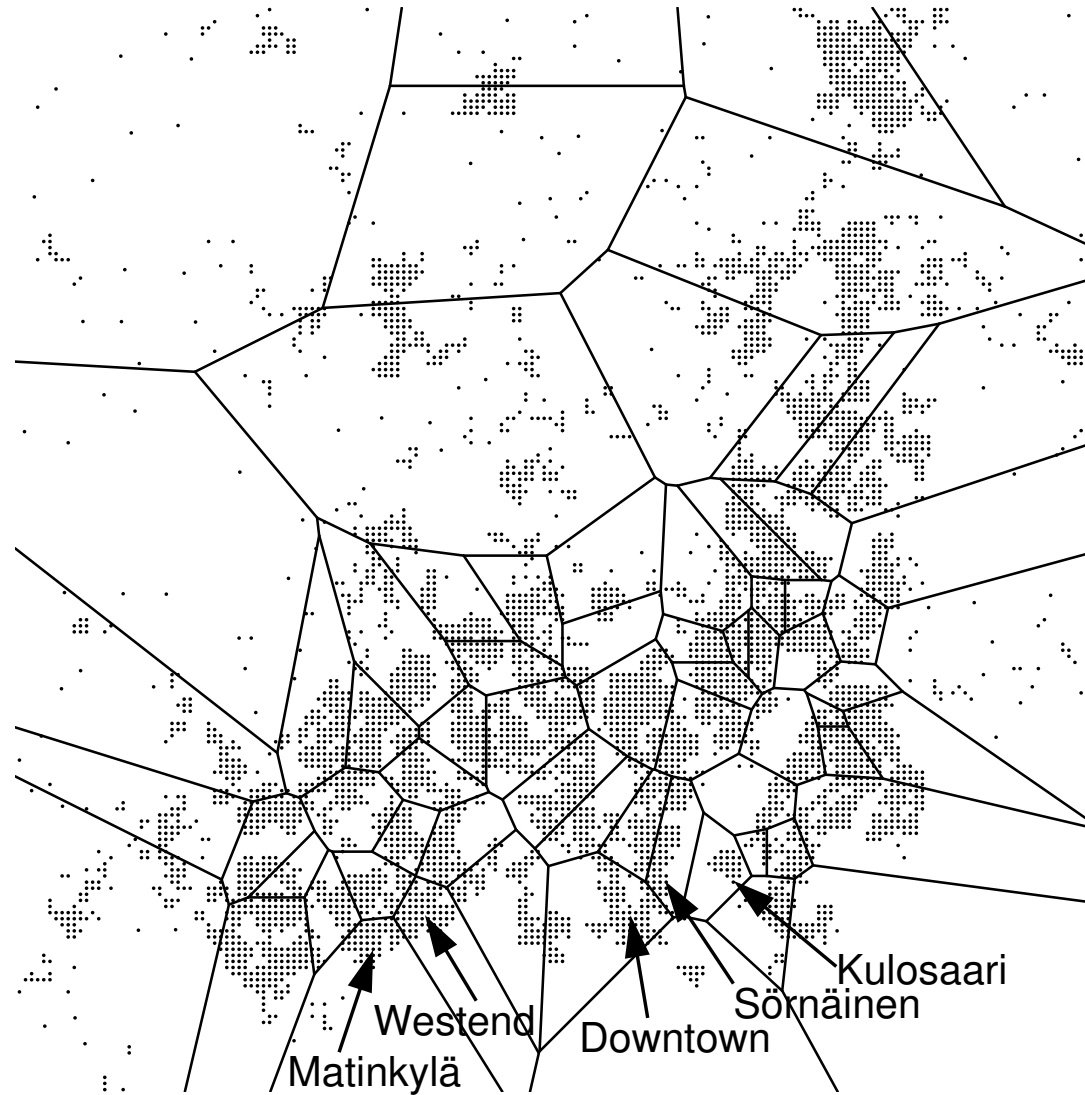
No EM-like fast, alternating optimization method found yet.

Conjugate gradients: better results than with $I(V; W)$.

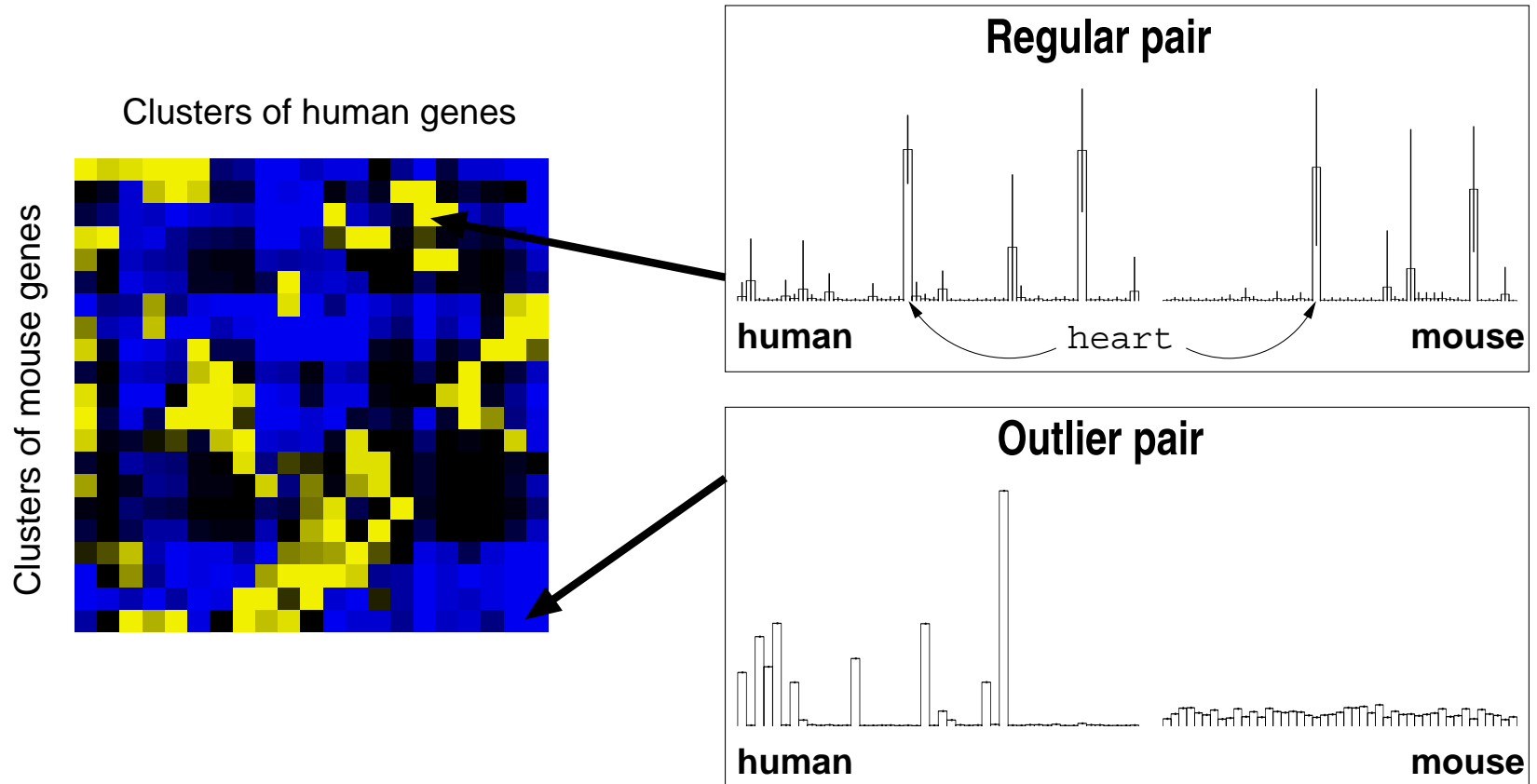
Less bias problems?

Still, there are the priors to decide.

Example: demographics, two-margin case



Example: gene expression, two margins optimized



Discussion/summary

Hard IB finds margins for contingency tables.

BF is an alternative to *I* as a dependency measure:

- Semi-generative.

- Asymptotically still *I*.

- Better for small data sets.

- Priors pose a kind of problem.

Hard IB is generalizable to continuous data.

- Voronoi partitions are practical.