On Clustering Histograms with k-Means by Using Mixed α-Divergences

Entropy 16(6): 3273-3301 (2014)

Frank Nielsen[1,2] Richard Nock[3] Shun-ichi Amari[4]

[1] Sony Computer Science Laboratories, Japan
E-Mail: Frank.Nielsen@acm.org
[2] École Polytechnique, France
[3] NICTA/ANU, Australia
[4] RIKEN Brain Science Institute, Japan

2014

Clustering histograms

- Information Retrieval systems (IRs) based on **bag-of-words** paradigm (bag-of-textons, bag-of-features, bag-of-X)
- The rôle of distances:
  - Initially, create a dictionary of "words" by quantizing using k-means clustering (depends on the underlying distance)
  - At query time, find "closest" (histogram) document by querying with the histogram query
- Notation: Positive arrays h (counting histogram) versus frequency histograms $\tilde{h}$ (normalized counting) d bins

For IRs, prefer **symmetric distances** (not necessarily metrics) like the Jeffreys divergence or the Jensen-Shannon divergence (unified by a one parameterized family of divergences in [11])

Ali-Silvey-Csiszár f-divergences

An important class of divergences: f-divergences [10, 1, 7] defined for a convex generator f (with $f(1) = f'(1) = 0$ and $f''(1) = 1$):

$$\boxed{I_f(p : q) \doteq \sum_{i=1}^{d} q^i f\left(\frac{p^i}{q^i}\right)}$$

Those divergences preserve **information monotonicity** [3] under any arbitrary transition probability (Markov morphisms).
f-divergences can be extended to positive arrays [3].

Mixed divergences

Defined on three parameters:

$$\boxed{M_\lambda(p:q:r) \doteq \lambda D(p:q) + (1-\lambda) D(q:r)}$$

for $\lambda \in [0,1]$.
Mixed divergences include:

- the **sided divergences** for $\lambda \in \{0,1\}$,
- the **symmetrized** (arithmetic mean) divergence for $\lambda = \frac{1}{2}$.

Mixed divergence-based k-means clustering

k distinct seeds from the dataset with $l_i = r_i$.

Input: Weighted histogram set $\mathcal{H}$, divergence $D(\cdot,\cdot)$, integer $k > 0$, real $\lambda \in [0,1]$;
Initialize left-sided/right-sided seeds $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^k$;
repeat
  //Assignment
  for $i = 1, 2, ..., k$ **do**
    $\mathcal{C}_i \leftarrow \{h \in \mathcal{H} : i = \arg\min_j M_\lambda(l_j : h : r_j)\}$;
  // Dual-sided centroid relocation
  for $i = 1, 2, ..., k$ **do**
    $r_i \leftarrow \arg\min_x D(\mathcal{C}_i : x) = \sum_{h \in \mathcal{C}_i} w_j D(h : x)$;
    $l_i \leftarrow \arg\min_x D(x : \mathcal{C}_i) = \sum_{h \in \mathcal{C}_i} w_j D(x : h)$;
until *convergence*;
Output: Partition of $\mathcal{H}$ into k clusters following $\mathcal{C}$;

$\rightarrow$ different from the k-means clustering with respect to the symmetrized divergences

α-divergences

For $\alpha \in \mathbb{R} \neq \pm 1$, define α-divergences [6] on positive arrays [18] :

$$\boxed{D_\alpha(p:q) \doteq \sum_{i=1}^{d} \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}p^i + \frac{1+\alpha}{2}q^i - (p^i)^{\frac{1-\alpha}{2}}(q^i)^{\frac{1+\alpha}{2}}\right)}$$

with $D_\alpha(p:q) = D_{-\alpha}(q:p)$ and in the limit cases $D_{-1}(p:q) = \mathrm{KL}(p:q)$ and $D_1(p:q) = \mathrm{KL}(q:p)$, where KL is the extended Kullback–Leibler divergence:

$$\mathrm{KL}(p:q) \doteq \sum_{i=1}^{d} p^i \log \frac{p^i}{q^i} + q^i - p^i.$$

α-divergences belong to f-divergences

The α-divergences belong to the class of Csiszár f-divergences with the following generator:

$$f(t) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - t^{(1+\alpha)/2}\right), & \text{if } \alpha \neq \pm 1, \\ t \ln t, & \text{if } \alpha = 1, \\ -\ln t, & \text{if } \alpha = -1 \end{cases}$$

The Pearson and Neyman χ^2 distances are obtained for $\alpha = -3$ and $\alpha = 3$:

$$D_3(\tilde{p} : \tilde{q}) = \frac{1}{2} \sum_i \frac{(\tilde{q}^i - \tilde{p}^i)^2}{\tilde{p}^i},$$

$$D_{-3}(\tilde{p} : \tilde{q}) = \frac{1}{2} \sum_i \frac{(\tilde{q}^i - \tilde{p}^i)^2}{\tilde{q}^i}.$$

Squared Hellinger symmetric distance is a $\alpha = 0$-divergence

Divergence D_0 is the squared Hellinger symmetric distance (scaled by 4) extended to positive arrays:

$$D_0(p:q) = 2\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x = 4H^2(p,q),$$

with the Hellinger distance:

$$\boxed{H(p,q) = \sqrt{\frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x}}$$

Mixed α-divergences

- Mixed α-divergence between a histogram x to **two** histograms p and q:

$$
\begin{aligned}
M_{\lambda,\alpha}(p:x:q) &= \lambda D_\alpha(p:x) + (1-\lambda) D_\alpha(x:q), \\
&= \lambda D_{-\alpha}(x:p) + (1-\lambda) D_{-\alpha}(q:x), \\
&= M_{1-\lambda,-\alpha}(q:x:p),
\end{aligned}
$$

- α-Jeffreys symmetrized divergence is obtained for $\lambda = \frac{1}{2}$:

$$
\boxed{S_\alpha(p,q) = M_{\frac{1}{2},\alpha}(q:p:q) = M_{\frac{1}{2},\alpha}(p:q:p)}
$$

- skew symmetrized α-divergence is defined by:

$$
\boxed{S_{\lambda,\alpha}(p:q) = \lambda D_\alpha(p:q) + (1-\lambda) D_\alpha(q:p)}
$$

Coupled k-Means++ α-Seeding

Algorithm 1: Mixed α-seeding; $\mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$

Input: Weighted histogram set $\mathcal{H}$, integer $k \geq 1$, real $\lambda \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} \leftarrow h_j$ with uniform probability ;

for $i = 2, 3, \ldots, k$ **do**

Pick at random histogram $h \in \mathcal{H}$ with probability:

$$\pi_{\mathcal{H}}(h) \doteq \frac{w_h M_{\lambda,\alpha}(c_h : h : c_h)}{\sum_{y \in \mathcal{H}} w_y M_{\lambda,\alpha}(c_y : y : c_y)}, \quad (1)$$

//where $(c_h, c_h) \doteq \arg\min_{(z,z) \in \mathcal{C}} M_{\lambda,\alpha}(z : h : z)$;

$\mathcal{C} \leftarrow \mathcal{C} \cup \{(h, h)\}$;

Output: Set of initial cluster centers $\mathcal{C}$;

A guaranteed probabilistic initialization

Let $C_{\lambda,\alpha}$ denote for short the cost function related to the clustering type chosen (left-, right-, skew Jeffreys or mixed) in $\mathrm{MAS}$and $C_{\lambda,\alpha}^{opt}$ denote the optimal related clustering in k clusters, for $\lambda \in [0,1]$ and $\alpha \in (-1,1)$. Then, on average, with respect to distribution (1), the initial clustering of MAS satisfies:

$$E_\pi[C_{\lambda,\alpha}] \quad \leq \quad 4\left\{ \begin{array}{ll} f(\lambda)g(k)h^2(\alpha)C_{\lambda,\alpha}^{opt} & \text{if} \quad \lambda \in (0,1) \\ g(k)z(\alpha)h^4(\alpha)C_{\lambda,\alpha}^{opt} & \text{otherwise} \end{array} \right. .$$

Here, $f(\lambda) = \max\left\{\frac{1-\lambda}{\lambda}, \frac{\lambda}{1-\lambda}\right\}$, $g(k) = 2(2+\log k)$, $z(\alpha) = \left(\frac{1+|\alpha|}{1-|\alpha|}\right)^{\frac{8|\alpha|^2}{(1-|\alpha|)^2}}$, $h(\alpha) = \max_i p_i^{|\alpha|} / \min_i p_i^{|\alpha|}$; the min is defined on strictly positive coordinates, and π denotes the picking distribution.

Mixed α-hard clustering: $\mathrm{MAhC}(\mathcal{H}, k, \lambda, \alpha)$

Input: Weighted histogram set $\mathcal{H}$, integer $k > 0$, real $\lambda \in [0, 1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^{k} \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

    //Assignment

    for $i = 1, 2, ..., k$ **do**

        $\mathcal{A}_i \leftarrow \{h \in \mathcal{H} : i = \arg\min_j M_{\lambda,\alpha}(l_j : h : r_j)\}$;

    // Centroid relocation

    for $i = 1, 2, ..., k$ **do**

        $r_i \leftarrow \left(\sum_{h \in \mathcal{A}_i} w_i h^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}}$;

        $l_i \leftarrow \left(\sum_{h \in \mathcal{A}_i} w_i h^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}}$;

until *convergence*;

Output: Partition of $\mathcal{H}$ in k clusters following $\mathcal{C}$;

Sided Positive α-Centroids [14]

The left-sided l_α and right-sided r_α positive weighted α-centroid coordinates of a set of n positive histograms $h_1, \ldots, h_n$ are weighted α-**means**:

$$r_\alpha^i = f_\alpha^{-1}\left(\sum_{j=1}^n w_j f_\alpha(h_j^i)\right), l_\alpha^i = r_{-\alpha}^i$$

with $f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & \alpha \neq \pm 1, \\ \log x & \alpha = 1. \end{cases}$

Sided Frequency α-Centroids [2]

Theorem (Amari, 2007)

The coordinates of the sided frequency α-centroids of a set of n weighted frequency histograms are the normalised weighted α-means.

Positive and Frequency α-centroids

Summary:

- $r_\alpha^i = \begin{cases} \left(\sum_{j=1}^n w_j (h_j^i)^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & \alpha \neq 1 \\ r_1^i = \prod_{j=1}^n (h_j^i)^{w_j} & \alpha = 1 \end{cases}$
- $l_\alpha^i = r_{-\alpha}^i = \begin{cases} \left(\sum_{j=1}^n w_j (h_j^i)^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}} & \alpha \neq -1 \\ l_{-1}^i = \prod_{j=1}^n (h_j^i)^{w_j} & \alpha = -1 \end{cases}$
- $\tilde{r}_\alpha^i = \frac{r_\alpha^i}{w(\tilde{r}_\alpha)}$
- $\tilde{l}_\alpha^i = \tilde{r}_{-\alpha}^i = \frac{r_{-\alpha}^i}{w(\tilde{r}_{-\alpha})}$

Mixed α-Centroids

Two centroids minimizer of:

$$\sum_j w_j M_{\lambda,\alpha}(l : h_j : r)$$

Generalizing mixed Bregman divergences [16]:

Theorem

The two mixed α-centroids are the left-sided and right-sided α-centroids.

Symmetrized Jeffreys-Type α-Centroids

$$
\begin{aligned}
S_\alpha(p,q) &= \frac{1}{2}\left(D_\alpha(p:q) + D_\alpha(q:p)\right) = S_{-\alpha}(p,q), \\
&= M_{\frac{1}{2}}(p:q:p),
\end{aligned}
$$

For $\alpha = \pm 1$, we get half of Jeffreys divergence:

$$
\boxed{S_{\pm 1}(p,q) = \frac{1}{2}\sum_{i=1}^{d}(p^i - q^i)\log\frac{p^i}{q^i}}
$$

Jeffreys α-divergence and Heinz means

When p and q are frequency histograms, we have for $\alpha \neq \pm 1$:

$$J_\alpha(\tilde{p} : \tilde{q}) = \frac{8}{1-\alpha^2}\left(1 + \sum_{i=1}^{d} H_{\frac{1-\alpha}{2}}(\tilde{p}^i, \tilde{q}^i)\right)$$

where $H_{\frac{1-\alpha}{2}}(a, b)$ a symmetric Heinz mean [8, 5]:

$$\boxed{H_\beta(a, b) = \frac{a^\beta b^{1-\beta} + a^{1-\beta} b^\beta}{2}}$$

Heinz means interpolate the arithmetic and geometric means and satisfies the inequality:

$$\sqrt{ab} = H_{\frac{1}{2}}(a, b) \leq H_\alpha(a, b) \leq H_0(a, b) = \frac{a+b}{2}.$$

Jeffreys divergence in the limit case

For $\alpha = \pm 1$, $S_\alpha(p, q)$ tends to the Jeffreys divergence:

$$\boxed{J(p, q) = \mathrm{KL}(p, q) + \mathrm{KL}(q, p) = \sum_{i=1}^{d} (p^i - q^i)(\log p^i - \log q^i)}$$

The Jeffreys divergence writes mathematically the same for frequency histograms:

$$\boxed{J(\tilde{p}, \tilde{q}) = \mathrm{KL}(\tilde{p}, \tilde{q}) + \mathrm{KL}(\tilde{q}, \tilde{p}) = \sum_{i=1}^{d} (\tilde{p}^i - \tilde{q}^i)(\log \tilde{p}^i - \log \tilde{q}^i)}$$

Analytic formula for the positive Jeffreys centroid [12]

Theorem (Jeffreys positive centroid [12])

The Jeffreys positive centroid $c = (c^1, ..., c^d)$ *of a set* $\{h_1, ..., h_n\}$ *of* n *weighted positive histograms with* d *bins can be calculated component-wise exactly using the Lambert* W *analytic function:*

$$\boxed{c^i = \frac{a^i}{W(\frac{a^i}{g^i} e)}}$$

where $a^i = \sum_{j=1}^n \pi_j h_j^i$ *denotes the coordinate-wise arithmetic weighted means and* $g^i = \prod_{j=1}^n (h_j^i)^{\pi_j}$ *the coordinate-wise geometric weighted means.*

The Lambert analytic function W [4] (positive branch) is defined by $W(x)e^{W(x)} = x$ for $x \geq 0$.

Jeffreys frequency centroid [12]

Theorem (Jeffreys frequency centroid [12])

Let $\tilde{c}$ denote the Jeffreys frequency centroid and $\tilde{c}' = \frac{c}{w_c}$ the normalised Jeffreys positive centroid. Then, the approximation factor $\alpha_{\tilde{c}'} = \frac{S_1(\tilde{c}', \tilde{\mathcal{H}})}{S_1(\tilde{c}, \tilde{\mathcal{H}})}$ is such that $1 \leq \alpha_{\tilde{c}'} \leq \frac{1}{w_c}$ (with $w_c \leq 1$).

better upper bounds in [12].

Reducing a n-size problem to a 2-size problem

Generalize [17] (symmetrized Kullback–Leibler divergence) and [15] (symmetrized Bregman divergence)

Lemma (Reduction property)

The symmetrized J_α-centroid of a set of n weighted histograms amount to computing the symmetrized α-centroid for the weighted α-mean and $-\alpha$-mean:

$$\min J_\alpha(x, \mathcal{H}) = \min_x \left(D_\alpha(x : r_\alpha) + D_\alpha(l_\alpha : x) \right).$$

Frequency symmetrized α-centroid

Minimizer of $\min_{\tilde{x}\in\Delta_d} \sum_j w_j S_\alpha(\tilde{x}, \tilde{h}_i)$
Instead of seeking for $\tilde{x}$ in the probability simplex, we can optimize on the unconstrained domain $\mathbb{R}^{d-1}$ by using the natural parameter reparameterization [13] of multinomials.

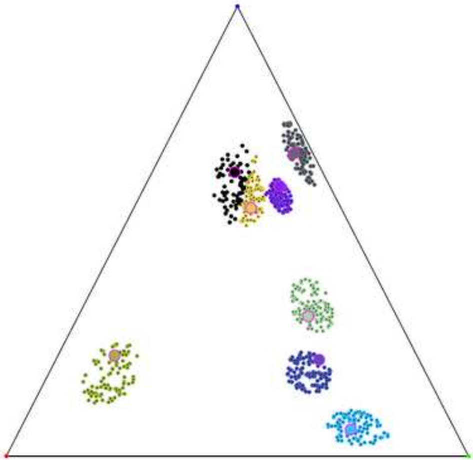
Lemma

The α-divergence for distributions belonging to the same exponential families amounts to computing a divergence on the corresponding natural parameters:

$$A_\alpha(p:q) = \frac{4}{1-\alpha^2}\left(1 - e^{-J_F^{\left(\frac{1-\alpha}{2}\right)}(\theta_p:\theta_q)}\right),$$

where $J_F^\beta(\theta_1:\theta_2) = \beta F(\theta_1) + (1-\beta)F(\theta_2) - F(\beta\theta_1 + (1-\beta)\theta_2)$ *is a skewed Jensen divergence defined for the log-normaliser F of the family.*

Implementation (in `processing.org`)



Snapshot of the α-clustering software. Here, $n = 800$ frequency histograms of three bins with $k = 8$, and $\alpha = 0.7$ and $\lambda = \frac{1}{2}$.

Soft Mixed α-Clustering

Learn both α and λ (α-EM [9])

Input: Histogram set $\mathcal{H}$ with $|\mathcal{H}| = m$, integer $k > 0$, real $\lambda \leftarrow \lambda_{\mathrm{init}} \in [0,1]$, real $\alpha \in \mathbb{R}$;

Let $\mathcal{C} = \{(l_i, r_i)\}_{i=1}^{k} \leftarrow \mathrm{MAS}(\mathcal{H}, k, \lambda, \alpha)$;

repeat

- //Expectation
- **for** $i = 1, 2, ..., m$ **do**
  - **for** $j = 1, 2, ..., k$ **do**
    - $p(j|h_i) = \frac{\pi_j \exp(-M_{\lambda,\alpha}(l_j : h_i : r_j))}{\sum_{j'} \pi_{j'} \exp(-M_{\lambda,\alpha}(l_{j'} : h_i : r_{j'}))}$;
- //Maximization
- **for** $j = 1, 2, ..., k$ **do**
  - $\pi_j \leftarrow \frac{1}{m} \sum_i p(j|h_i)$;
  - $l_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1+\alpha}{2}} \right)^{\frac{2}{1+\alpha}}$;
  - $r_i \leftarrow \left(\frac{1}{\sum_i p(j|h_i)} \sum_i p(j|h_i) h_i^{\frac{1-\alpha}{2}} \right)^{\frac{2}{1-\alpha}}$;
- //Alpha - Lambda
- $\alpha \leftarrow \alpha - \eta_1 \sum_{j=1}^{k} \sum_{i=1}^{m} p(j|h_i) \frac{\partial}{\partial \alpha} M_{\lambda,\alpha}(l_j : h_i : r_j)$;
- **if** $\lambda_{\mathrm{init}} \neq 0, 1$ **then**
  - $\lambda \leftarrow \lambda - \eta_2 \left(\sum_{j=1}^{k} \sum_{i=1}^{m} p(j|h_i) D_\alpha(l_j : h_i) - \sum_{j=1}^{k} \sum_{i=1}^{m} p(j|h_i) D_\alpha(h_i : r_j) \right)$;
  - //for some small η_1, η_2; ensure that $\lambda \in [0,1]$.

until *convergence*;

Output: Soft clustering of $\mathcal{H}$ according to k densities $p(j|.)$ following $\mathcal{C}$;

Summary

1. Mixed divergences,mixed divergence k-means++ seeding, coupled k-means seeding
2. Sided left or right α-centroid k-means
3. Coupled k-means with respect to mixed α-divergences relying on dual α-centroids
4. Symmetrized Jeffreys-type α-centroid (variational) k-means,

All technical proofs and details in:
Entropy 16(6): 3273-3301 (2014)

Bibliographic references I

Syed Mumtaz Ali and Samuel David Silvey.
A general class of coefficients of divergence of one distribution from another.
Journal of the Royal Statistical Society, Series B, 28:131–142, 1966.

Shun-ichi Amari.
Integration of stochastic models by minimizing α-divergence.
Neural Computation, 19(10):2780–2796, 2007.

Shun-ichi Amari.
alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes.
IEEE Transactions on Information Theory, 55(11):4925–4931, 2009.

D. A. Barry, P. J. Culligan-Hensley, and S. J. Barry.
Real values of the W-function.
ACM Trans. Math. Softw., 21(2):161–171, June 1995.

Ádám Besenyei.
On the invariance equation for Heinz means.
Mathematical Inequalities & Applications, 15(4):973–979, 2012.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari.
Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization.
Entropy, 13(1):134–170, 2011.

Imre Csiszár.
Information-type measures of difference of probability distributions and indirect observation.
Studia Scientiarum Mathematicarum Hungarica, 2:229–318, 1967.

Bibliographic references II

Erhard Heinz.
Beiträge zur störungstheorie der spektralzerlegung.
Mathematische Annalen, 123:415–438, 1951.

Yasuo Matsuyama.
The alpha-EM algorithm: surrogate likelihood maximization using alpha-logarithmic information measures.
IEEE Transactions on Information Theory, 49(3):692–706, 2003.

Tetsuzo Morimoto.
Markov processes and the *h*-theorem.
Journal of the Physical Society of Japan, 18(3), March 1963.

Frank Nielsen.
A family of statistical symmetric divergences based on Jensen's inequality.
CoRR, abs/1009.4004, 2010.

Frank Nielsen.
Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms.
IEEE Signal Processing Letters (SPL), 20(7), July 2013.

Frank Nielsen and Vincent Garcia.
Statistical exponential families: A digest with flash cards, 2009.
arXiv.org:0911.4863.

Frank Nielsen and Richard Nock.
The dual Voronoi diagrams with respect to representational Bregman divergences.
In *International Symposium on Voronoi Diagrams (ISVD)*, pages 71–78, DTU Lyngby, Denmark, June 2009. IEEE.

Bibliographic references III

Frank Nielsen and Richard Nock.
Sided and symmetrized Bregman centroids.
IEEE Transactions on Information Theory, 55(6):2882–2904, 2009.

Richard Nock, Panu Luosto, and Jyrki Kivinen.
Mixed Bregman clustering with approximation guarantees.
In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, pages 154–169, Berlin, Heidelberg, 2008. Springer-Verlag.

Raymond N. J. Veldhuis.
The centroid of the symmetrical Kullback-Leibler distance.
IEEE signal processing letters, 9(3):96–99, March 2002.

Huaiyu Zhu and Richard Rohwer.
Measurements of generalisation based on information geometry.
In StephenW. Ellacott, JohnC. Mason, and IainJ. Anderson, editors, *Mathematics of Neural Networks*, volume 8 of *Operations Research/Computer Science Interfaces Series*, pages 394–398. Springer US, 1997.