Supplement for the article Clustering of Distributions: A Case of Patent Citations

Vladimir Batagelj, Simona Korenjak-Černe, Nataša Kejžar
University of Ljubljana

January 25, 2011

1 Hierarchical clustering update formulas for different error functions δ

Let us denote

$$s_u = \sum_{X \in C_u} x, \quad S_u = \sum_{X \in C_u} x^2, \quad h_u = \sum_{X \in C_u} \frac{1}{x}, \quad H_u = \sum_{X \in C_u} \frac{1}{x^2}$$

and $C_z = C_u \cup C_v$, $C_u \cap C_v = \emptyset$. We have

$$s_z = s_u + s_v, \quad S_z = S_u + S_v, \quad h_z = h_u + h_v, \quad H_z = H_u + H_v$$

In the derivations of formulas we follow the same steps as in the case of δ_4.

1.1 δ_3

$$\delta_3(x,y) = (\frac{x-y}{y})^2$$

center equation: $us_u = S_u$
center z:

$$z(s_u + s_v) = S_u + S_v = us_u + vs_v$$
$$z = \frac{us_u + vs_v}{s_u + s_v}$$

difference:

$$\begin{aligned} \delta(x,z) - \delta(x,u) &= (\frac{x-z}{z})^2 - (\frac{x-u}{u})^2 = (\frac{x-z}{z} - \frac{x-u}{u})(\frac{x-z}{z} + \frac{x-u}{u}) \\ &= \frac{u-z}{z^2u^2}(x^2(u+z) - 2xuz) \end{aligned}$$

sum of differences:

$$\sum_{X \in C_u} \delta(x,z) - \delta(x,u) = \frac{u-z}{z^2u^2}(S_u(u+z) - 2s_u uz) = \frac{s_u}{z^2u}(u-z)^2 = \frac{s_u}{u}\delta(u,z)$$

dissimilarity:

$$D(C_u, C_v) = \frac{s_u}{u}\delta(u,z) + \frac{s_v}{v}\delta(v,z)$$

1.2 δ_5

$$\delta_5(x,y) = (\frac{x-y}{x})^2$$

center equation: $uH_u = h_u$
center z:

$$h_u + h_v = z(H_u + H_v) = z(\frac{h_u}{u} + \frac{h_v}{v})$$

$$z = \frac{h_u + h_v}{\frac{h_u}{u} + \frac{h_v}{v}}$$

difference:

$$\begin{aligned} \delta(x,z) - \delta(x,u) &= (\frac{x-z}{x})^2 - (\frac{x-u}{x})^2 \\ &= \frac{1}{x^2}((x-z)-(x-u))((x-z)+(x-u)) \\ &= \frac{u-z}{x^2}(2x-(u+z)) = (u-z)(\frac{2}{x} - \frac{u+z}{x^2}) \end{aligned}$$

sum of differences:

$$\sum_{X \in C_u} \delta(x,z) - \delta(x,u) = (u-z)(2h_u - (u+z)H_u) = \frac{h_u}{u}(u-z)^2 = uh_u\delta(u,z)$$

dissimilarity:

$$D(C_u, C_v) = uh_u\delta(u,z) + vh_v\delta(v,z)$$

1.3 δ_6

$$\delta_6(x,y) = \frac{(x-y)^2}{x}$$

center equation: $uh_u = |C_u|$
center z:

$$z(h_u + h_v) = |C_u| + |C_v|$$

$$z = \frac{|C_u| + |C_v|}{h_u + h_v}$$

difference:

$$\begin{aligned} \delta(x,z) - \delta(x,u) &= \frac{(x-z)^2}{x} - \frac{(x-u)^2}{x} \\ &= \frac{u-z}{x}(2x-(u+z)) = (u-z)(2-\frac{u+z}{x}) \end{aligned}$$

sum of differences:

$$\sum_{X \in C_u} \delta(x,z) - \delta(x,u) = (u-z)(2|C_u| - (u+z)h_u) = \frac{|C_u|}{u}(u-z)^2 = |C_u|\delta(u,z)$$

dissimilarity:

$$D(C_u, C_v) = |C_u|\delta(u,z) + |C_v|\delta(v,z)$$

1.4 δ_7

$$\delta_7(x,y) = \frac{(x-y)^2}{xy}$$

center equation: $u^2 h_u = s_u$
center z:

$$s_u + s_v = z^2(h_u + h_v) = z(\frac{s_u}{u^2} + \frac{s_v}{v^2})$$

$$z = \sqrt{\frac{s_u + s_v}{\frac{s_u}{u^2} + \frac{s_v}{v^2}}}$$

difference:

$$\begin{aligned} \delta(x,z) - \delta(x,u) &= \frac{(x-z)^2}{xz} - \frac{(x-u)^2}{xu} = \frac{1}{xuz}(ux^2 + uz^2 - zx^2 - zu^2) \\ &= \frac{u-z}{xuz}(x^2 - uz) = (u-z)(\frac{x}{uz} - \frac{1}{x}) \end{aligned}$$

sum of differences:

$$\sum_{X \in C_u} \delta(x,z) - \delta(x,u) = (u-z)(\frac{s_u}{uz} - h_u) = \frac{(u-z)s_u}{u}(\frac{1}{z} - \frac{1}{u}) = \frac{(u-z)^2 s_u}{u^2 z} = \frac{s_u}{u}\delta(u,z)$$

dissimilarity:

$$D(C_u, C_v) = \frac{s_u}{u}\delta(u,z) + \frac{s_v}{v}\delta(v,z)$$

1.5 Summary

i	$\delta_i(x,y)$	z_i	$D_i(C_u, C_v)$
1	$(x-y)^2$	$\frac{\lvert C_u\rvert u+\lvert C_v\rvert v}{\lvert C_u\rvert+\lvert C_v\rvert}$	$\frac{\lvert C_u\rvert\cdot\lvert C_v\rvert}{\lvert C_u\rvert+\lvert C_v\rvert}\delta(u,v)$
3	$(\frac{x-y}{y})^2$	$\frac{us_u+vs_v}{s_u+s_v}$	$\frac{s_u}{u}\delta(u,z)+\frac{s_v}{v}\delta(v,z)$
4	$\frac{(x-y)^2}{y}$	$\sqrt{\frac{\lvert C_u\rvert u^2+\lvert C_v\rvert v^2}{\lvert C_u\rvert+\lvert C_v\rvert}}$	$\lvert C_u\rvert\delta(u,z)+\lvert C_v\rvert\delta(v,z)$
5	$(\frac{x-y}{x})^2$	$\frac{h_u+h_v}{\frac{h_u}{u}+\frac{h_v}{v}}$	$uh_u\delta(u,z)+vh_v\delta(v,z)$
6	$\frac{(x-y)^2}{x}$	$\frac{\lvert C_u\rvert+\lvert C_v\rvert}{h_u+h_v}$	$\lvert C_u\rvert\delta(u,z)+\lvert C_v\rvert\delta(v,z)$
7	$\frac{(x-y)^2}{xy}$	$\sqrt{\frac{s_u+s_v}{\frac{s_u}{u^2}+\frac{s_v}{v^2}}}$	$\frac{s_u}{u}\delta(u,z)+\frac{s_v}{v}\delta(v,z)$

2 Clustering results for US patent data: Additional variables

2.1 Patents granted in year 1980 with at least 20 citations

Table 1: p-values of statistical tests of possible factor explanatory variable assignee type and its relative frequencies for hierarchical clusters and the whole data set.

	assignee	probabilities for types (1–6)
cluster 1	0.001**	0.29, 0.17, 0.06, 0.16, 0.22, 0.11
cluster 2	0.000**	0.22, 0.25, 0.07, 0.20, 0.14, 0.11
cluster 3	0.476	0.25, 0.20, 0.14, 0.11, 0.13, 0.16
cluster 4	0.000**	0.29, 0.08, 0.24, 0.10, 0.13, 0.16
cluster 5	0.027*	0.22, 0.16, 0.15, 0.20, 0.11, 0.15
cluster 6	0.000**	0.16, 0.11, 0.20, 0.15, 0.18, 0.20
cluster 7	0.328	0.23, 0.16, 0.14, 0.12, 0.16, 0.18
overall		*0.24, 0.17, 0.14, 0.15, 0.15, 0.16*

2.2 Patents granted in years 1980–84, 15-year time period

Table 2: p-values of statistical tests of possible factor explanatory variable category (first) and assignee type (second) and their relative frequencies for hierarchical clusters and the whole data set.

	category	probabilities for categories (1–6)
cluster 1	0.000**	0.07, 0.53, 0.39, 0.01, 0.00, 0.01
cluster 2	0.000**	0.08, 0.57, 0.33, 0.00, 0.00, 0.01
cluster 3	0.000**	0.15, 0.59, 0.23, 0.01, 0.00, 0.01
cluster 4	0.000**	0.16, 0.56, 0.25, 0.01, 0.01, 0.01
cluster 5	0.000**	0.15, 0.57, 0.26, 0.01, 0.00, 0.02
cluster 6	0.005**	0.12, 0.57, 0.29, 0.01, 0.01, 0.01
cluster 7	0.054	0.11, 0.57, 0.29, 0.01, 0.00, 0.01
overall		*0.12, 0.57, 0.29, 0.01, 0.00, 0.01*
	assignee	probabilities for types (1–6)
cluster 1	0.000**	0.24, 0.15, 0.06, 0.20, 0.20, 0.14
cluster 2	0.000**	0.21, 0.22, 0.07, 0.21, 0.17, 0.13
cluster 3	0.000**	0.24, 0.18, 0.17, 0.13, 0.13, 0.15
cluster 4	0.000**	0.25, 0.13, 0.14, 0.13, 0.17, 0.19
cluster 5	0.000**	0.26, 0.15, 0.13, 0.14, 0.15, 0.17
cluster 6	0.004**	0.22, 0.16, 0.10, 0.21, 0.16, 0.16
cluster 7	0.000**	0.20, 0.20, 0.10, 0.19, 0.17, 0.15
overall		*0.23, 0.18, 0.11, 0.17, 0.16, 0.15*

Table 3: p-values of statistical tests of possible interval explanatory variables for hierarchical clusters (data from 1980–84).

	# citations	generality	originality	selfcitations
cluster 1	0.000**	0.000**	0.007**	0.000**
cluster 2	0.000**	0.000**	0.218	0.000**
cluster 3	0.000**	0.000**	0.012*	0.000**
cluster 4	0.000**	0.859	0.475	0.000**
cluster 5	0.000**	0.135	0.626	0.056
cluster 6	0.000**	0.789	0.106	0.002**
cluster 7	0.000**	0.206	0.985	0.163