k-averages : criterions, convergence

Mathias Rossignol

23 janvier 2014

1 Conventions, possible objective functions

Partition of n objects $\mathcal{O} = \{o_1, \ldots, o_n\}$ into k classes $\mathcal{C} = \{c_1, \ldots, c_k\}$. We'll write n_{c_i} the cardinal of class c_i, and o_{ij} its elements : $c_i = \left\{o_{i1}, \ldots, o_{in_{c_i}}\right\}$.

To simplify below, when we're simply considering one class, no matter which, we shall omit the first index and write $c = \{o_1, \ldots, o_{n_c}\}$.

The similarity between objects shall be written $s(o_i, o_j)$.

What we're interested in : maximising the average intra-class object-to-object similarity. That's pretty straightforward, as far as I can tell there's only one possible variant : whether we have a class-by-class normalization or only a global one.

In the first case, objects appearing in small classes will have more « weight » in the function. In the second, big classes will tend to dominate the objective function. So : neither is 100% satisfying. Better try to implement both and test.

Note that since the number of obejcts does not vary, global normalization (divide by number of objects) is equivalent to no normalization at all.

Other possible variants : using other averaging methods (geometric mean ? Does that make sense ?). Maybe later.

2 Impact of object reallocation on class quality

Let us extend the notation s to the *similarity of an object to a class*, which we define as the average similarity of that object with all objects of the class. $s(o, c)$ accepts two definitions, depending on whether or not o is in c :

If $o \notin c$,

$$s(o, c) = \frac{1}{n_c} \sum_{i=1}^{n_c} s(o, o_i) \tag{1}$$

If $o \in c$, then necessarily $\exists i \mid o = o_i$

$$s(o, c) = s(o_i, c) = \frac{1}{n_c - 1} \sum_{j=1 \ldots n_c, j \neq i} s(o_i, o_j) \tag{2}$$

Let's call « quality » of a class the average intra-class object-to-object similarity, and write it $\mathcal{Q}$:

$$\mathcal{Q}(c) = \frac{1}{n_c} \sum_{i=1}^{n_c} s(o_i, c) \tag{3}$$

Since all objects are in c, we use the formula in (2) to get :

$$\begin{aligned}\mathcal{Q}(c) &= \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{1}{n_c-1}\sum_{\substack{j=1\dots n_c\\ j\neq i}} s(o_i,o_j)\\ &= \frac{1}{n_c(n_c-1)}\sum_{i=1}^{n_c}\sum_{\substack{j=1\dots n_c\\ j\neq i}} s(o_i,o_j)\end{aligned} \tag{4}$$

Using the assumption that the similarity matrix is symmetrical, we can reach (this is an indispensable transformation for future calculations) :

$$\mathcal{Q}(c) = \frac{2}{n_c(n_c-1)}\sum_{i=2}^{n_c}\sum_{j=1}^{i-1} s(o_i,o_j) \tag{5}$$

For future use, let us define the notation :

$$\Sigma(c) = \sum_{i=2}^{n_c}\sum_{j=1}^{i-1} s(o_i,o_j) \tag{6}$$

Thus :

$$\mathcal{Q}(c) = \frac{2}{n_c(n_c-1)}\Sigma(c) \qquad \text{and} \qquad \Sigma(c) = \frac{n_c(n_c-1)\mathcal{Q}(c)}{2} \tag{7}$$

2.1 Removing an object from a class

Assuming that $o \in c$, necessarily $\exists i \mid o = o_i$. Since the numbering of objects is arbitrary, we can assume that $o = o_{n_c}$ then generalize from the result thus obtained.

$$\begin{aligned}\mathcal{Q}(c \setminus o_{n_c}) &= \frac{2}{(n_c-1)(n_c-2)}\sum_{i=2}^{n_c-1}\sum_{j=1}^{i-1} s(o_i,o_j)\\ &= \frac{2}{(n_c-1)(n_c-2)}\left[\Sigma(c) - \sum_{j=1}^{n_c-1} s(o_{n_c},o_j)\right]\\ &= \frac{2}{(n_c-1)(n_c-2)}\left[\Sigma(c) - (n_c-1)s(o_{n_c},c)\right]\\ &= \frac{2n_c(n_c-1)\mathcal{Q}(c)}{2(n_c-1)(n_c-2)} - \frac{2(n_c-1)s(o_{n_c},c)}{(n_c-1)(n_c-2)}\\ &= \frac{n_c\mathcal{Q}(c) - 2s(o_{n_c},c)}{n_c-2}\end{aligned} \tag{8}$$

The quality of a class after removal of an object is thus :

$$\mathcal{Q}(c \setminus o) = \frac{n_c\mathcal{Q}(c) - 2s(o,c)}{n_c-2} \tag{9}$$

And the change in quality from its previous value :

$$
\begin{aligned}
\mathcal{Q}\left(c \setminus o\right) - \mathcal{Q}\left(c\right) &= \frac{n_c \mathcal{Q}(c) - (n_c - 2)\mathcal{Q}(c) - 2s\left(o, c\right)}{n_c - 2} \\
&= \frac{2\left(\mathcal{Q}(c) - s\left(o, c\right)\right)}{n_c - 2}
\end{aligned} \tag{10}
$$

2.2 Adding an object to a class

Assuming that $o \notin c$, we'll consider for the sake of simplicity that o becomes o_{n_c+1} in the modified class c. Following a path similar to above, we get :

$$
\begin{aligned}
\mathcal{Q}(c \cup o_{n_c+1}) &= \frac{2}{n_c(n_c+1)} \sum_{i=2}^{n_c+1} \sum_{j=1}^{i-1} s\left(o_i, o_j\right) \\
&= \frac{2}{n_c(n_c+1)} \left[\Sigma(c) + n_c s\left(o_{n_c+1}, c\right)\right] \\
&= \frac{(n_c - 1)\mathcal{Q}(c) + 2s\left(o_{n_c+1}, c\right)}{n_c + 1}
\end{aligned} \tag{11}
$$

The quality of a class c after adding an object o is thus :

$$
\mathcal{Q}\left(c \cup o\right) = \frac{(n_c - 1)\mathcal{Q}(c) + 2s\left(o, c\right)}{n_c + 1} \tag{12}
$$

And the change in quality from its previous value :

$$
\begin{aligned}
\mathcal{Q}\left(c \cup o\right) - \mathcal{Q}\left(c\right) &= \frac{(n_c - 1)\mathcal{Q}(c) - (n_c + 1)\mathcal{Q}(c) + 2s\left(o, c\right)}{n_c + 1} \\
&= \frac{2\left(s\left(o, c\right) - \mathcal{Q}(c)\right)}{n_c + 1}
\end{aligned} \tag{13}
$$

3 Impact of object reallocation on global quality (objective function)

3.1 Class-normalized objective function

In that case, the calculation is direct : from (10) and (13), we can see that the impact on the objective function of moving an object c from class c_s ("source"), to whom it belongs, to a distinct class c_t ("target") is :

$$
\delta = \frac{2\left(s\left(o, c_t\right) - \mathcal{Q}(c_t)\right)}{n_{c_t} + 1} + \frac{2\left(\mathcal{Q}(c_s) - s\left(o, c_s\right)\right)}{n_{c_s} - 2} \tag{14}
$$

Using this value as the basis to decide object reallocation, and only performing reallocations while it is strictly positive, ensures that the objective function is strictly increasing, and therefore that the process converges to one of its local maxima.

3.2 Object-normalized objective function

This complicates the calculation a bit, but not much : when moving an object c from class c_s ("source"), to whom it belongs, to a distinct class c_t ("target"), $(n_{c_s} - 1)$ objects are affected by the variation in (10), and n_{c_t}) are affected by that in (13), in addition to the variation in simiarity of o to the class it belongs to :

$$\delta = \frac{2n_{c_t}\left(s\left(o, c_t\right) - \mathcal{Q}(c_t)\right)}{n_{c_t} + 1} + \frac{2(n_{c_s} - 1)\left(\mathcal{Q}(c_s) - s\left(o, c_s\right)\right)}{n_{c_s} - 2} + s(o, c_t) - s(o, c_s) \tag{15}$$

Once again, using this value as the basis to decide object reallocation, and only performing reallocations while it is strictly positive, ensures that the objective function is strictly increasing, and therefore that the process converges to one of its local maxima.