Photo: Vladimir Batagelj, ***UNI-LJ***

Metoda voditeljev

Vladimir Batagelj

FMF, Univerza Ljubljani

1143. Sredin seminar, Ljubljana, 21. marec 2007

Outline

Leaders method

In order to support our intuition in further development we shall briefly describe a simple version of dynamic clusters method – the *leaders* or k-means method, which is the basis of the ISODATA program (one among the most popular clustering programs) and several recent 'data-mining' methods. In the leaders method the criterion function has the form SR.

The basic scheme of leaders method is simple:

determine $\mathbf{C}_0$; $\mathbf{C} := \mathbf{C}_0$;
repeat
    determine for each $C \in \mathbf{C}$ its leader $\overline{C}$;
    the new clustering $\mathbf{C}$ is obtained by assigning each unit
        to its nearest leader
until leaders stabilize

To obtain a 'good' solution and an impression of its quality we can repeat this procedure with different (random) $\mathbf{C}_0$.

The dynamic clusters method

The dynamic clusters method is a generalization of the above scheme. Let us denote:

Λ	– set of *representatives*
$\mathrm{L} \subseteq \Lambda$	– *representation*
Ψ	– set of *feasible representations*
$W : \Phi \times \Psi \rightarrow \mathbb{R}_0^+$	– *extended criterion function*
$G : \Phi \times \Psi \rightarrow \Psi$	– *representation function*
$F : \Phi \times \Psi \rightarrow \Phi$	– *clustering function*

and

Basic scheme of the dynamic clusters method

the following conditions have to be satisfied:

W0. $P(\mathbf{C}) = \min_{\mathrm{L}\in\Psi} W(\mathbf{C}, \mathrm{L})$

the functions G and F tend to improve (diminish) the value of the extended criterion function W:

W1. $W(\mathbf{C}, G(\mathbf{C}, \mathrm{L})) \leq W(\mathbf{C}, \mathrm{L})$
W2. $W(F(\mathbf{C}, \mathrm{L}), \mathrm{L}) \leq W(\mathbf{C}, \mathrm{L})$

then the *dynamic clusters method* can be described by the scheme:

$\mathbf{C} := \mathbf{C}_0; \mathrm{L} := \mathrm{L}_0;$
repeat
$\mathrm{L} := G(\mathbf{C}, \mathrm{L});$
$\mathbf{C} := F(\mathbf{C}, \mathrm{L})$
until the clustering stabilizes

Properties of DCM

To this scheme corresponds the sequence $v_n = (\mathbf{C}_n, \mathrm{L}_n), n \in \mathbb{N}$ determined by relations

$$\mathrm{L}_{n+1} = G(\mathbf{C}_n, \mathrm{L}_n) \quad \text{and} \quad \mathbf{C}_{n+1} = F(\mathbf{C}_n, \mathrm{L}_{n+1})$$

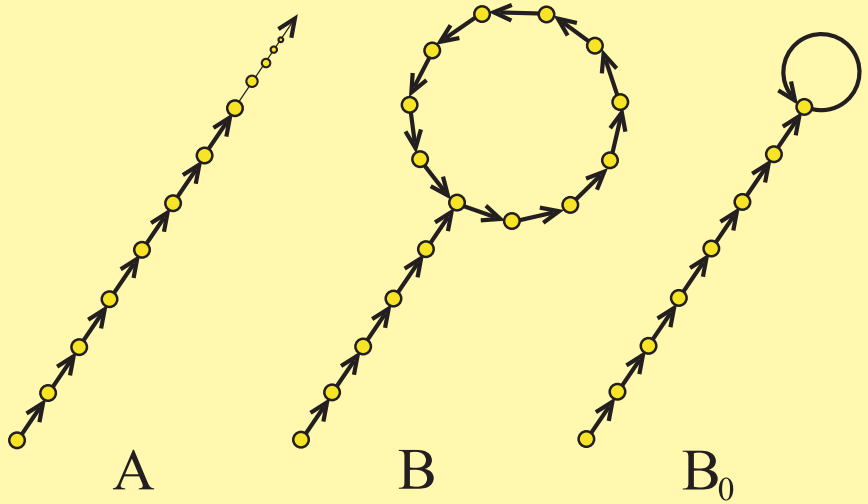
and the sequence of values of the extended criterion function $u_n = W(\mathbf{C}_n, \mathrm{L}_n)$. Let us also denote $u^* = P(\mathbf{C}^*)$. Then it holds:

Theorem 1.1 *For every* $n \in \mathbb{N}$, $u_{n+1} \leq u_n$, $u^* \leq u_n$, *and if for* $k > m$, $v_k = v_m$ *then* $\forall n \geq m : u_n = u_m$.

The Theorem 2.6 states that the sequence u_n is monotonically decreasing and bounded, therefore it is convergent. Note that the limit of u_n is not necessarily u^* – the dynamic clusters method is a local optimization method.

... types of of DCM sequences

Type A: $\neg\exists k, m \in \mathbb{N}, k > m : v_k = v_m$
Type B: $\exists k, m \in \mathbb{N}, k > m : v_k = v_m$
Type B_0: Type B with $k = m + 1$



The DCM sequence (v_n) is of type B if

- sets Φ and Ψ are both finite.
  For example, when we select a representative of C among its members.
- $\exists\delta > 0 : \forall n \in \mathbb{N} : (v_{n+1} \neq v_n \Rightarrow u_n - u_{n+1} > \delta)$

Because the sets $\mathbf{U}$ and consequently Φ are finite we expect from a good dynamic clusters procedure to stabilize in finite number of steps – is of type B.

Additional requirement

The conditions W0, W1 and W2 are not strong enough to ensure this. We shall try to compensate the possibility that the set of representations Ψ is infinite by the additional requirement:

W3. $W(\mathbf{C}, G(\mathbf{C}, \mathrm{L})) = W(\mathbf{C}, \mathrm{L}) \Rightarrow \mathrm{L} = G(\mathbf{C}, \mathrm{L})$

With this requirement the 'symmetry' between Φ and Ψ is distroyed. We could reestablish it by the requirement:

W4. $W(F(\mathbf{C}, \mathrm{L}, \mathrm{L})) = W(\mathbf{C}, \mathrm{L}) \Rightarrow \mathbf{C} = F(\mathbf{C}, \mathrm{L})$

but it turns out that W4 often fails. For this reason we shall avoid it.

Theorem 1.2 *If W3 holds and if there exists* $m \in \mathbb{N}$ *such that* $u_{m+1} = u_m$, *then also* $\mathrm{L}_{m+1} = \mathrm{L}_m$.

Simple clustering and representation functions

Usually, in the applications of the DCM, the clustering function takes the form $F : \Psi \rightarrow \Phi$. In this case the condition W2 simplifies to: $W(F(\mathrm{L}), \mathrm{L}) \leq W(\mathbf{C}, \mathrm{L})$ which can be expressed also as $F(\mathrm{L}) \in \mathrm{Min}_{\mathbf{C} \in \Phi} W(\mathbf{C}, \mathrm{L})$. For such, *simple* clustering functions it holds:

Theorem 1.3 *If the clustering function F is simple and if there exists $m \in \mathbb{N}$ such that $\mathrm{L}_{m+1} = \mathrm{L}_m$, then for every $n \geq m : v_n = v_m$.*

What can be said about the case when G is *simple* – has the form $G : \Phi \rightarrow \Psi$?

Theorem 1.4 *If W3 holds and the representation function G is simple then:*

a. $G(\mathbf{C}) = \arg\min_{\mathrm{L} \in \Psi} W(\mathbf{C}, \mathrm{L})$

b. $\exists k, m \in \mathbb{N}, k > m \forall i \in \mathbb{N} : v_{k+i} = v_{m+i}$

c. $\exists m \in \mathbb{N} \forall n \geq m : u_n = u_m$

d. *if also F is simple then* $\exists m \in \mathbb{N} \forall n \geq m : v_n = v_m$

Original DCM

In the original dynamic clusters method (Diday, 1979) both functions F and G are simple – $F : \Psi \rightarrow \Phi$ and $G : \Phi \rightarrow \Psi$.

We proved, if also W3 holds and the functions F and G are simple, then:

G0. $G(\mathbf{C}) = \underset{\mathrm{L}\in\Psi}{\operatorname{argmin}} W(\mathbf{C}, \mathrm{L})$

and

F0. $F(\mathrm{L}) \in \operatorname{Min}_{\mathbf{C}\in\Phi} W(\mathbf{C}, \mathrm{L})$

In other words, given an extended criterion function W, the relations G0 and F0 define an appropriate pair of functions G and F such that the DCM stabilizes in finite number of steps.