1 Map Score

We have clustered same cells in two ways and got clusterings N and M, clustering N has $S(N)$ clusters $(N_1, \ldots, N_{S(N)})$, and clustering M has $S(M)$ clusters $(M_1, \ldots, M_{S(M)})$. The confusion matrix $C \in Mat_{S_N,S_M}$ is defined the following way: $C_{(i,j)} = |N_i \cap M_j|$, the cardinality of the intersection of clusters N_i and M_j. Notice that for each i, j, $C_{i,j} \in \mathbb{Z}^+$.

The goal is to construct a measure $L(N, M)$ between clusterings N and M that would fit the following criteria:

1. No dependence on cluster labels of N and M
2. If clusterings N and M are identical then $L(N, M) = 0$
3. If clusterings N and M are 'opposite' then $L(N, M) = 1$
4. The fewer elements of each cluster in N fall into different clusters in M the lower $L(N, M)$, and vice versa

Then the similarity measure $L(N, M)$ between clusterings will be defined in the following way:

$$V(N, M) = \frac{1}{2 \cdot S(N) \cdot (S(M) - 1)} \sum_{i=1}^{S(N)} \frac{\sum_{j=1}^{S(M)} C_{(i,j)} - \max_j C_{(i,j)}}{\max_j C_{(i,j)}}$$

Similarly,

$$V(M, N) = \frac{1}{2 \cdot (S(N) - 1) \cdot S(M)} \sum_{j=1}^{S(M)} \frac{\sum_{i=1}^{S(N)} C_{(i,j)} - \max_i C_{(i,j)}}{\max_i C_{(i,j)}}$$

Where $2 \cdot S(N) \cdot (S(M) - 1)$ and $2 \cdot (S(N) - 1) \cdot S(M)$ are normalisation terms. Finally,

$$L(N, M) = V(N, M) + V(M, N)$$