# Constrained Clustering with Weak Label Prior

**Jing ZHANG, Ruidong FAN, Hong TAO, Jiacheng JIANG, Chenping HOU**

Frontiers of Computer Science, DOI: 10.1007/s11704-023-3355-7

# The propsed approach

- **The current problem:** In many real scenarios, two kinds of weak label prior information, e.g., pairwise constraints and cluster ratio, are easily obtained or already available. How to incorporate them to improve clustering performance is important but rarely studied.

- **Main idea:** A constrained Clustering with Weak Label Prior (CWLP) is proposed to consider compound weak label prior in an integrated framework.

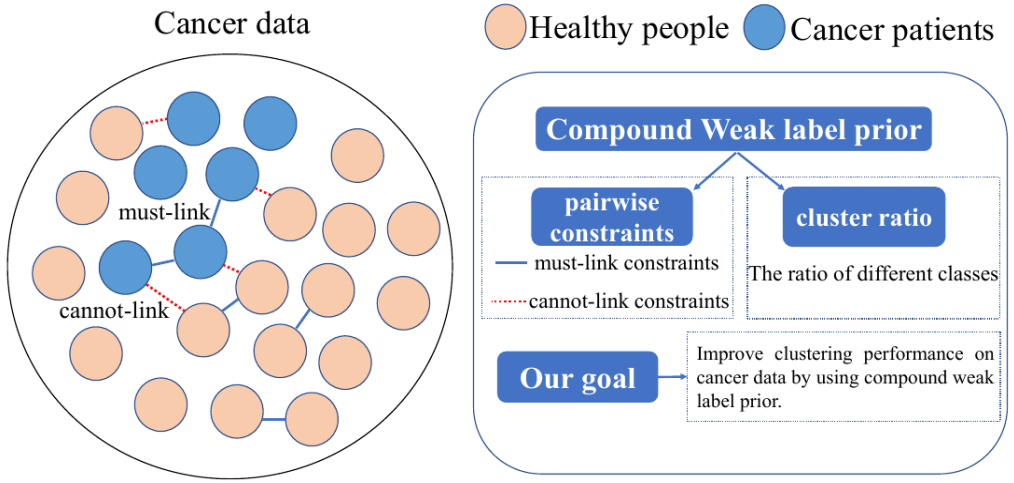



**Figure 2** Compound weak label prior plays an essential role in increasing clustering confidence.

Fig2 shows this example that compound weak label prior is used for clustering task. This example is just one of various real-world scenarios.

# Main experimental results

**Table 2** The average ACC (std) for all datasets. The best result on each dataset is highlighted in bold.

| Datasets | $K$-means | Ncut+SR | Rcut+SR | JSESR | PCOG | CPCG | SCCR | CWLP |
|---|---|---|---|---|---|---|---|---|
| Glass | 0.830(0.002) | 0.808(0.002) | 0.799(0.003) | 0.830(0.002) | **1.000(0.000)** | 0.836(0.000) | **1.000(0.000)** | **1.000(0.000)** |
| glass6 | 0.814(0.045) | 0.864(0.005) | 0.864(0.002) | 0.828(0.048) | 0.911(0.000) | 0.801(0.001) | 0.865(0.093) | **0.972(0.000)** |
| ecoli3 | 0.774(0.002) | 0.783(0.011) | 0.786(0.005) | 0.774(0.002) | 0.895(0.000) | 0.785(0.000) | 0.817(0.049) | **0.934(0.000)** |
| vowel0 | 0.523(0.007) | 0.552(0.001) | 0.552(0.002) | 0.617(0.104) | 0.461(0.000) | 0.595(0.079) | 0.843(0.019) | **0.849(0.020)** |
| balance-uni | 0.512(0.015) | 0.586(0.003) | 0.586(0.003) | 0.534(0.030) | 0.808(0.000) | 0.571(0.102) | 0.855(0.008) | **0.856(0.007)** |
| shuttlec0vsc4 | 0.950(0.023) | 0.978(0.003) | 0.983(0.022) | 0.966(0.030) | 0.950(0.000) | 0.951(0.024) | 0.963(0.013) | **0.992(0.018)** |
| dermatology6 | 0.503(0.005) | 0.637(0.001) | 0.637(0.002) | 0.645(0.103) | 0.935(0.095) | 0.615(0.216) | 0.897(0.025) | **0.954(0.030)** |
| shuttle2vs5 | 0.772(0.111) | 0.718(0.000) | 0.718(0.000) | 0.772(0.111) | 0.985(0.000) | 0.851(0.183) | 0.985(0.006) | **0.994(0.006)** |

**Table 3** The average F-score (std) for all datasets. The best result on each dataset is highlighted in bold.

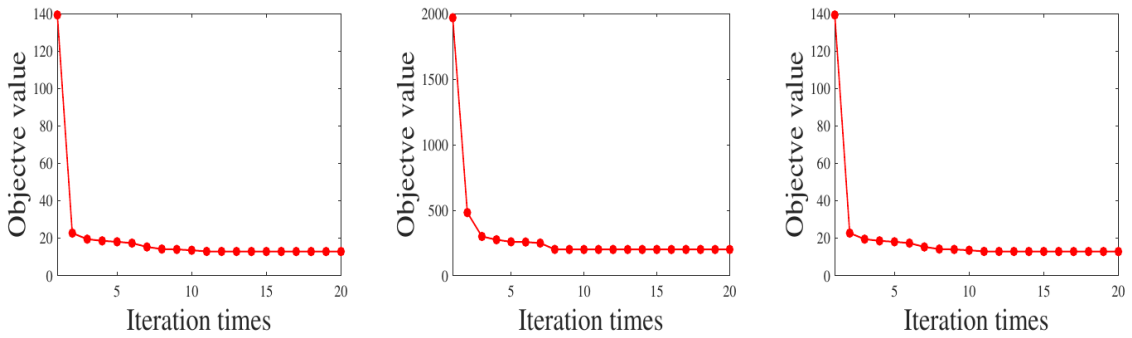
| Datasets | $K$-means | Ncut+SR | Rcut+SR | JSESR | PCOG | CPCG | SCCR | CWLP |
|---|---|---|---|---|---|---|---|---|
| Glass | 0.731(0.003) | 0.704(0.002) | 0.695(0.003) | 0.731(0.003) | **1.000(0.000)** | 0.739(0.000) | **1.000(0.000)** | **1.000(0.000)** |
| glass6 | 0.798(0.025) | 0.835(0.016) | 0.835(0.008) | 0.811(0.030) | 0.903(0.000) | 0.853(0.019) | 0.857(0.085) | **0.964(0.000)** |
| ecoli3 | 0.745(0.002) | 0.753(0.003) | 0.756(0.005) | 0.745(0.002) | 0.895(0.000) | 0.773(0.001) | 0.818(0.043) | **0.924(0.000)** |
| vowel0 | 0.626(0.001) | 0.631(0.000) | 0.631(0.000) | 0.670(0.075) | 0.495(0.000) | 0.656(0.041) | 0.842(0.015) | **0.847(0.017)** |
| balance-uni | 0.631(0.002) | 0.647(0.003) | 0.647(0.002) | 0.634(0.006) | 0.838(0.000) | 0.658(0.056) | 0.855(0.006) | **0.856(0.005)** |
| shuttlec0vsc4 | 0.949(0.026) | 0.976(0.002) | 0.981(0.001) | 0.963(0.030) | 0.961(0.000) | 0.951(0.023) | 0.960(0.014) | **0.991(0.020)** |
| dermatology6 | 0.641(0.003) | 0.684(0.007) | 0.684(0.008) | 0.695(0.061) | 0.935(0.000) | 0.733(0.087) | 0.897(0.024) | **0.952(0.029)** |
| shuttle2vs5 | 0.984(0.001) | 0.981(0.007) | 0.673(0.003) | 0.987(0.002) | 0.989(0.000) | 0.977(0.009) | 0.986(0.004) | **0.994(0.006)** |

**Table 4** The average Recall (std) for all datasets. The best result on each dataset is highlighted in bold.

| Datasets | $K$-means | Ncut+SR | Rcut+SR | JSESR | PCOG | CPCG | SCCR | CWLP |
|---|---|---|---|---|---|---|---|---|
| Glass | 0.775(0.003) | 0.746(0.005) | 0.735(0.010) | 0.775(0.003) | **1.000(0.000)** | 0.784(0.000) | **1.000(0.000)** | **1.000(0.000)** |
| glass6 | 0.834(0.077) | 0.900(0.006) | 0.900(0.008) | 0.841(0.077) | 0.830(0.000) | 0.865(0.060) | 0.857(0.085) | **0.964(0.000)** |
| ecoli3 | 0.912(0.001) | 0.916(0.004) | 0.917(0.004) | 0.912(0.001) | 0.820(0.000) | 0.913(0.001) | 0.818(0.043) | **0.924(0.000)** |
| vowel0 | 0.834(0.002) | 0.833(0.001) | 0.833(0.002) | 0.836(0.004) | 0.830(0.000) | 0.835(0.002) | 0.842(0.015) | **0.847(0.017)** |
| balance-uni | 0.855(0.002) | 0.857(0.001) | 0.857(0.001) | 0.855(0.000) | 0.855(0.000) | 0.856(0.003) | 0.855(0.006) | **0.857(0.005)** |
| shuttlec0vsc4 | 0.908(0.044) | 0.956(0.002) | 0.970(0.001) | 0.950(0.018) | **0.998(0.000)** | 0.908(0.044) | 0.960(0.014) | 0.991(0.020) |
| dermatology6 | 0.890(0.007) | 0.875(0.009) | 0.875(0.003) | 0.903(0.026) | 0.893(0.000) | 0.914(0.035) | 0.897(0.024) | **0.952(0.029)** |
| shuttle2vs5 | 0.973(0.001) | 0.965(0.000) | 0.965(0.000) | 0.973(0.001) | 0.971(0.000) | 0.976(0.007) | 0.985(0.006) | **0.994(0.006)** |

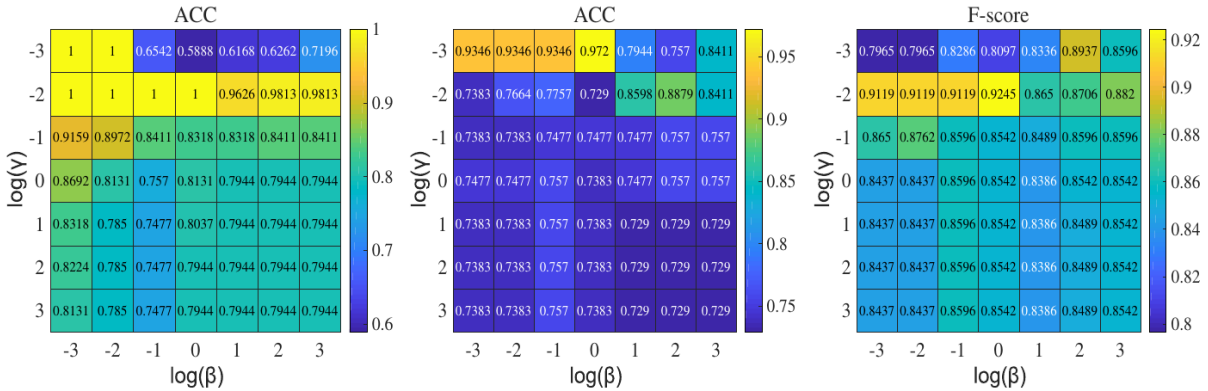
**Table 5** The average Purity (std) for all datasets. The best result on each dataset is highlighted in bold.

| Datasets | $K$-means | Ncut+SR | Rcut+SR | JSESR | PCOG | CPCG | SCCR | CWLP |
|---|---|---|---|---|---|---|---|---|
| Glass | 0.830(0.002) | 0.808(0.002) | 0.799(0.005) | 0.829(0.002) | **1.000(0.000)** | 0.834(0.000) | **1.000(0.000)** | **1.000(0.000)** |
| glass6 | 0.868(0.025) | 0.864(0.004) | 0.864(0.002) | 0.878(0.018) | 0.953(0.000) | 0.902(0.031) | 0.908(0.038) | **0.972(0.000)** |
| ecoli3 | 0.774(0.002) | 0.783(0.005) | 0.786(0.006) | 0.774(0.002) | **0.988(0.005)** | 0.801(0.001) | 0.899(0.007) | 0.934(0.000) |
| vowel0 | 0.529(0.007) | 0.570(0.000) | 0.570(0.000) | 0.638(0.126) | 0.485(0.000) | 0.614(0.096) | **0.909(0.000)** | **0.909(0.000)** |
| balance-uni | 0.514(0.016) | 0.600(0.003) | 0.600(0.007) | 0.588(0.035) | 0.867(0.040) | 0.588(0.115) | **0.922(0.003)** | **0.922(0.002)** |
| shuttlec0vsc4 | 0.989(0.013) | 0.978(0.004) | 0.983(0.002) | 0.969(0.021) | 0.950(0.000) | 0.991(0.011) | 0.963(0.013) | **0.992(0.018)** |
| dermatology6 | 0.553(0.012) | 0.693(0.002) | 0.693(0.012) | 0.672(0.106) | **0.991(0.000)** | 0.731(0.120) | 0.946(0.007) | 0.962(0.016) |
| shuttle2vs5 | 0.777(0.116) | 0.733(0.000) | 0.733(0.000) | 0.777(0.116) | 1.000(0.000) | 0.857(0.188) | 0.987(0.002) | **0.994(0.006)** |




(a) ecoli3 (b) vowel0 (c) dermatology6

**Figure 3** Convergence curve of the proposed CWLP on three datasets. (a) ecoli3. (b) vowel0. (c) dermatology6. These figures show the value of the objective function decreases rapidly and achieves convergence after a few iterations.




(a) Glass (b) glass6 (c) ecoil3

**Figure 4** The ACC (F-score) value as a function of $\beta$ and $\gamma$ for three datasets. (a) Glass. (b) glass6. (c)