Mirkin’s Rules for Cluster Interpretation

(Supplement to Lecture 2020_2))

1. Rules 1 to 8 2--7
2. Example: Interpretation
   of Iris taxa 8—14
3. Conclusion 15

Rules =
Points, to be taken seriously 1:

1. Each cluster is to be interpreted separately.
2. A set F of features to be interpreted is selected by the user; any external (not used at clustering) feature(s) may be included too.

Relative difference 2:

3. Given a cluster k and a quantitative feature v F, the relative difference is computed:

$$\mathbf{d_{kv}=100[c_{kv}/c_v - 1]} \text{ (per cent)}$$

Here $\mathbf{c_{kv}}$ is within-cluster mean of v, and $\mathbf{c_v}$ is grand mean (mean over the dataset) of v

Relative difference for dummy 1/0 feature

4. Given a cluster k and a category v F, the Quetelet index is computed:

$$\mathbf{q_{kv} = 100[p_{kv} / (p_k\, p_v) - 1]} \text{ (per cent)}$$

Here $\mathbf{p_{kv}}$ is the proportion of entities falling in both cluster k and category v, $\mathbf{p_v}$ is proportion of category v in the dataset,

$\mathbf{p_k}$ is proportion of cluster k in the dataset.

In fact, $\mathbf{p_{kv}} = \mathbf{d_{kv}}$ if category v is represented by a 1/0 dummy.

Interpretative features, V^+ and V^-

5. Given a cluster k, pick up those features and categories v F for which values of $\mathbf{d_{kv}}$ or $\mathbf{q_{kv}}$ are far from 0, say, **greater than 35%**, forming set V_k^+, or **smaller than }35%,** forming set $V_k^{\}}$.

6. **Describe** cluster k as that characterized by features from V_k^+as those “much greater than the average” and features from $V_k^{\}}$as those “much smaller than the average”. (For larger deviations, you may use “very much” modifier.)

Conceptualization 5:

7. After you have described cluster k by sets V^+ and V^-, try to conceptualize the description on a deeper level, in more general terms. If you can, put your conceptualization down in writing. If you cannot, do not get frustrated: you may get more lucky next time.

Super-Conceptuaization 6:

8. After you have conceptualized all the clusters, take a look at the conceptual descriptions and try conceptualize the entire partition.

If you can, put your conceptualization down in writing. If you cannot, do not get frustrated: you may get more lucky next time.

Example: Interpreting Iris taxa

1. Take first taxon T1 (the first 50 specimens) to interpret.
2. Take all four Iris dataset features (Sepal length, Sepal width, Petal length, Petal width) as F set of features.

Example: Interpreting Iris taxon T1:
3. Compute relative differences

	SLength	SWidth	PLength	PWidth
Taxon center $c_k=(c_{kv})$	**5.006**	**3.428**	**1.462**	**0.246**
Grand mean $c=(c_v)$	5.843	3.057	3.758	1.199
Difference	}0.837	0.371	}2.296	}0.953
Relative difference, d_{kv}%	}14.3	+12.1	**}61.1**	**}79.5**

$d_{kv}= (c_{kv} - c_v)/c_v$, per cent!

Example: Interpreting taxon T1:

4. Set of interpreting categories is empty, since we have no nominal categories in F

Interpretation of taxon T1 in Iris dataset, 1

5. V_{T1}^{+} is empty; $V_{T1}^{\}}$= {Petal length, Petal width}

6. Conceptualize taxon T1 as that characterized by this statement:

T1 = Those specimens at which the Petal is much smaller than the average (on both length and width).

Interpretation of taxon T1 in Iris dataset

7. A more parsimonious concept: “Small petals”.

8. Conceptual interpretation of the partition of Iris in three taxa, 1:

- Relative Difference: 100*(CMean –GMean)/GMean

	SL	SW	PL	PW
T1	-14.3297	12.1239	**-61.0963**	**-79.4886**
T2	1.5859	-9.3982	13.3582	10.5614
T3	12.7439	-2.7257	**47.7382**	**68.9272**

Taxa conceptual descriptions:

- T1 is “small petals”, T3 is “large petals”,

T2 is “just about the average”

9. Super-Conceptual Description of the partition of Iris in three taxa, 2:

Taxa conceptual descriptions:

- T1 is “small petals”, T3 is “large petals”, T2 is “just about the average”

- A deeper level yet:

“Sepal is not used in the description”

Why is that? I am not a botanist, cannot explain. Should undertake a research inspired by the data analysis.

Conclusion

- [«Бди!» Козьма Прутков], that is:
- “Be on Alert!” Koz’ma Prutkov, a famous Russian 19-century poet