

## A Novel Idea of The Validation Criterion of Clustering

著者	AMANO Kou
内容記述	第14回情報メディア学会研究大会で発表したポスター（2015年6月27日 同志社大学 今出川キャンパス 良心館）
year	2015
その他のタイトル	クラスタリング指標における新しい評価基準のアイデア
URL	<a href="http://hdl.handle.net/2241/00131709">http://hdl.handle.net/2241/00131709</a>

# A Novel Idea of The Validation Criterion of Clustering

Kou AMANO<sup>†‡§</sup>  
 amano@brc.riken.jp

<sup>†</sup> RIKEN

<sup>‡</sup> University of Tsukuba

<sup>§</sup> National Institute of Agrobiological Sciences

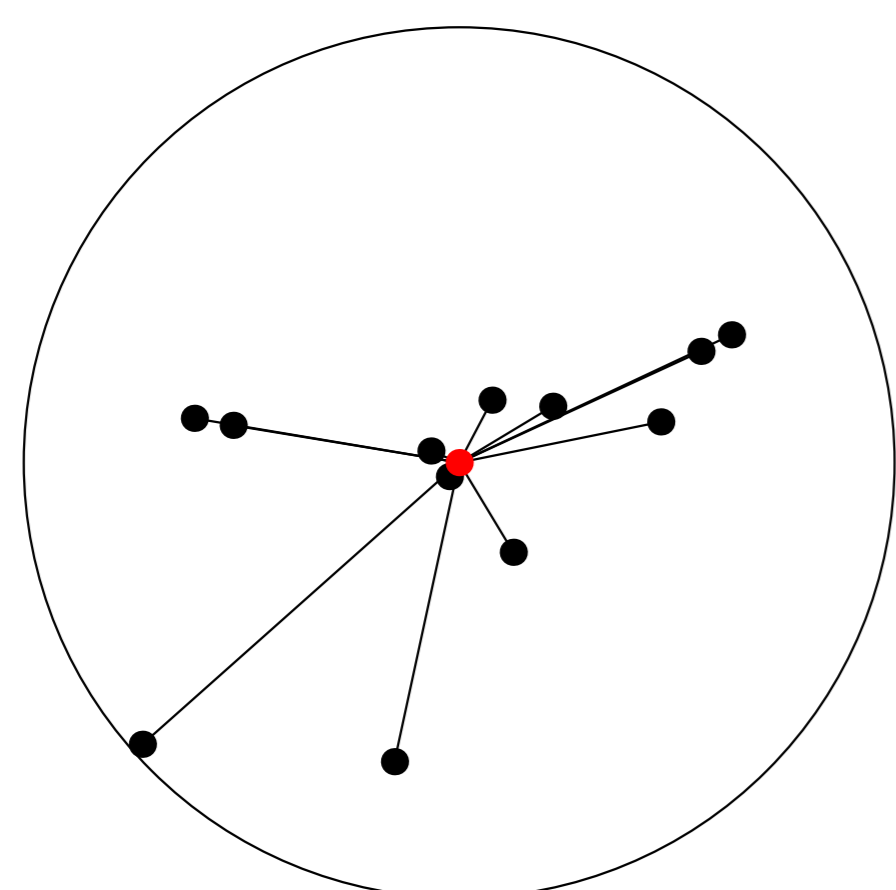
## Introduction

This paper introduce a new index of cluster validity only based on simplicity of cluster structure.

- Non-hierarchical clustering is the task of the k-partition problems with heuristics,
- and the solutions are neither global nor unique,
- hence it needs indices to assess the clustering results.

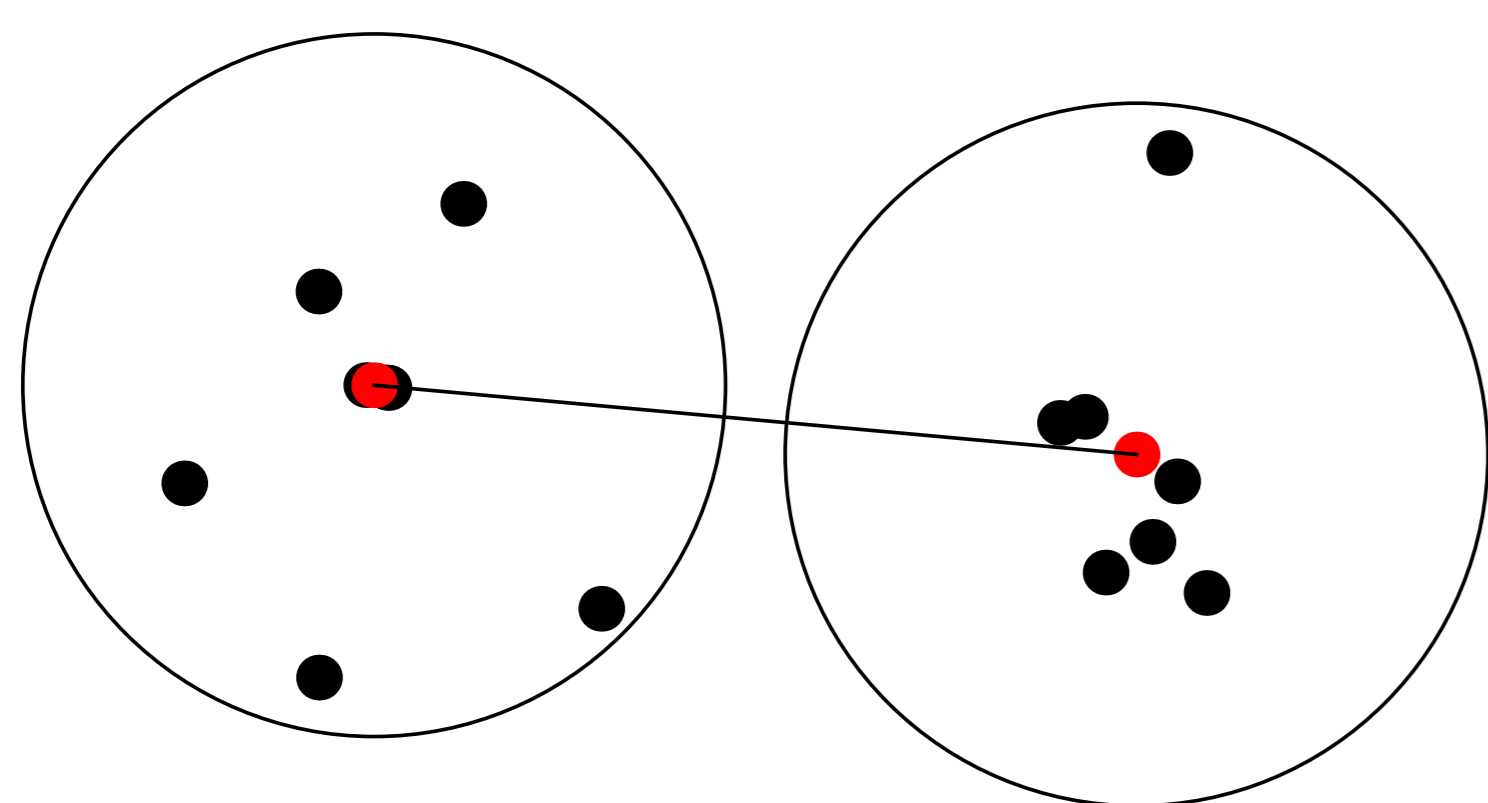
## Existing indices

- Criteria
  - Compactness



Within-cluster distance

- Separability

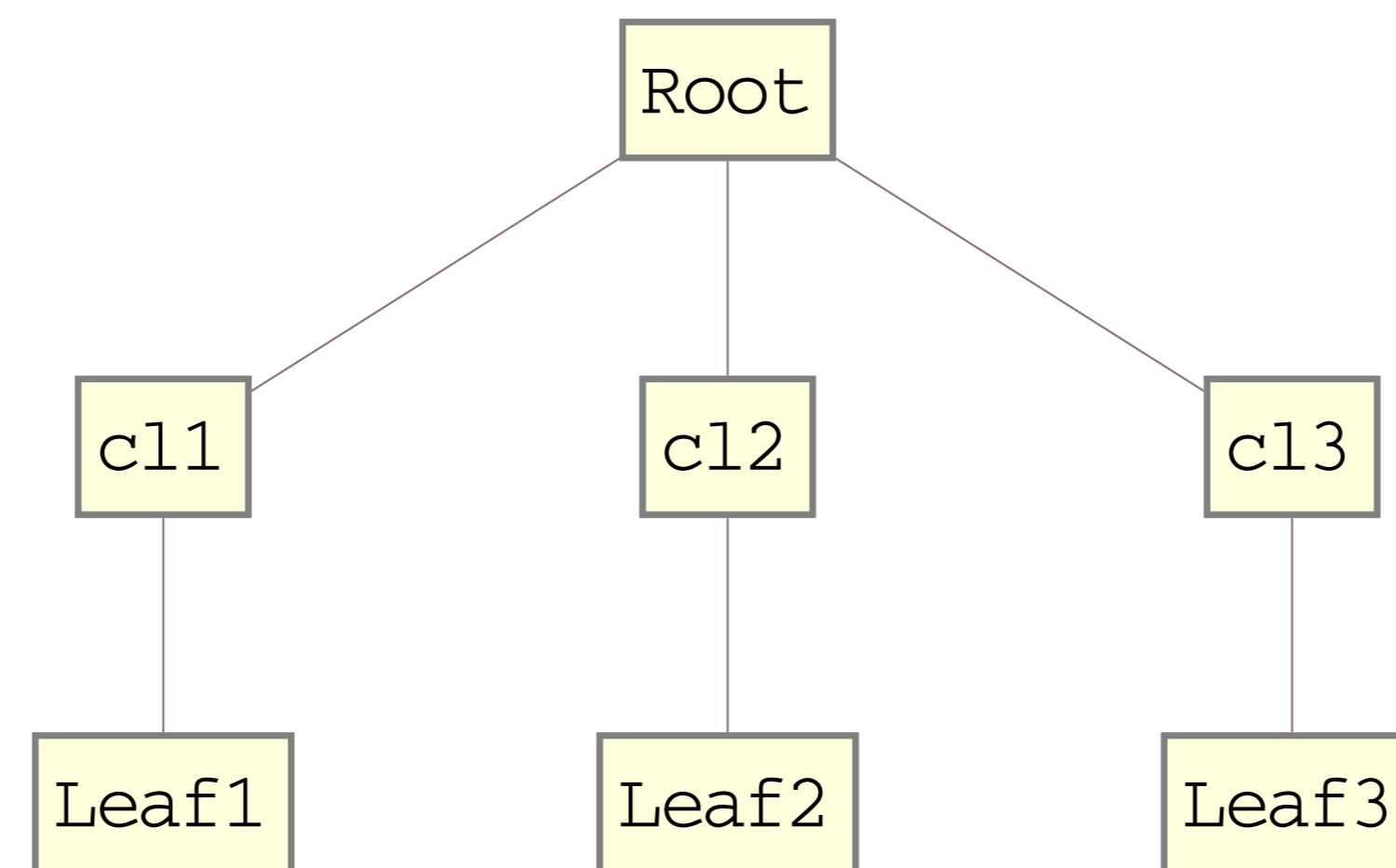


Between-cluster distance

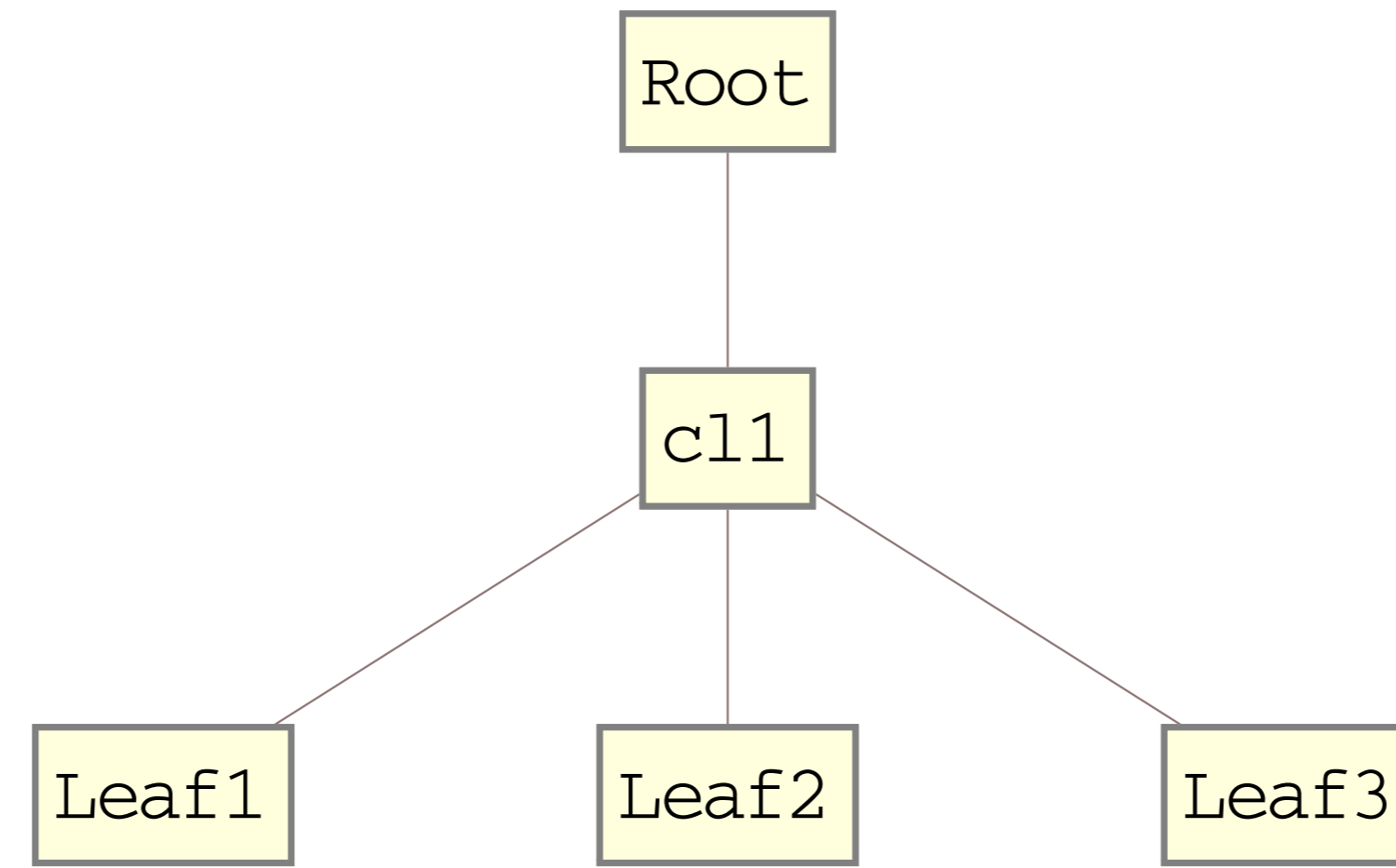
- Elements
  - Number of clusters
  - Number of members of each cluster
  - representative (or each) distance within cluster
  - representative (or each) distance between clusters
- Well-known indices
  - Dunn index[1]
  - Calinski-Harabazs index[2]
  - Davies-Bouldin index[3]

## Desirable properties of indices

- The index would indicate identical values to cluster structures with the same topology:



$$k = N ; \{ \{Leaf1\}, \{Leaf2\}, \{Leaf3\} \} \xrightarrow{\text{simplify}} \{Leaf1, Leaf2, Leaf3\}$$

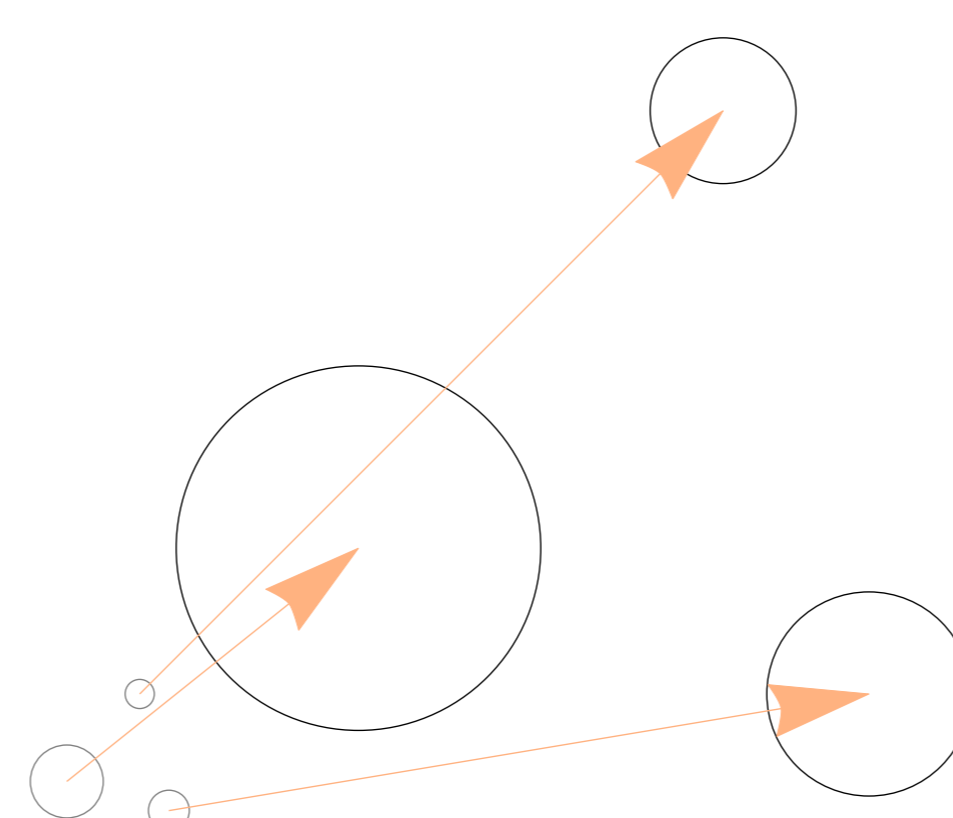


$$k = 1 ; \{ \{Leaf1, Leaf2, Leaf3\} \} \xrightarrow{\text{simplify}} \{Leaf1, Leaf2, Leaf3\}$$

$$\downarrow$$

$$Id_{k=N}(S) = Id_{k=1}(S)$$

- The index would indicate identical values to sample sets with a similar distribution:



$$\text{Scale ; } Id(a \times S) = Id(S)$$



$$\text{Shift ; } Id(b + S) = Id(S)$$

## Simplicity Index(SI): a novel index

- New criterion
  - Simplicity
- Elements
  - $k$  : number of clusters
  - $c$  : number of members of each cluster
  - $v$  : space capacity
- Definition

$$SI = k \prod_{n=1}^k c_n^{\frac{r_n}{R}}$$

where  $r_n$  : radius of cluster  $n$ , and  $R$  : radius of complete samples.  $\frac{r_n}{R}$  is used for  $v$ .

## Simplicity Ratio(SR): a derivative of SI

- Definition
- $$SR = SI/N$$
- where  $N$  : total number of samples.
- Failure detection

$$\begin{cases} SR < 1 & \text{Success} \\ SR \geq 1 & \text{Failure} \end{cases}$$

## Conclusion

The index  $SI$  and the derivative  $SR$  have been introduced. These are only based on simplicity of the cluster structure.

## References and Acknowledgement

- [1] J. C. Dunn. Journal of Cybermetrics. 3,32, (1973).
  - [2] T. Calinski; T. Harabasz. Communication in Statistics. 3,1, (1974).
  - [3] D. L. Davies; D. W. Bouldin. Pattern Analysis and Machine Intelligence, IEEE Transactions on. PAMI-1,224, (1979).
- [Ack] Dr. Kaoru Fukami, RIKEN BioResource Center  
 [Ack] Mr. Masamichi Wada, JST  
 [Ack] Mathematica Kenkyu-Kai

2015-06-27, 情報メディア学会 研究大会  
 同志社大学 今出川キャンパス 良心館

