"Cluster Validation by Prediction Strength" review

Zhongkai Wang

University of Florida
Department of Biostatistics

zkwang@ufl.edu

June 3, 2015

Overview

1. Theory

2. Simulation in SAS and R
   - SAS
   - R

Prediction Strength

- Assessing number of clusters (how many, how well)
- model selection based on prediction strength
- Comparing to GAP, Calinski and Harabasz (CH), Krazanowski and Lai(KL)

Selection-based Statistics

Prediction Strength

$$ps(k) = \min_{1\leq j\leq k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i\neq i'\in A_{kj}} D[C(X_{kr},k),X_{te}]_{ii'}$$

GAP

$$GAP_n(k) = E_n{}^*[log(W_k)] - log(W_k)$$

CH

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

KL

$$DIFF(k) = (k-1)^{2/p}W_{k-1} - k^{2/p}W_k$$

$$KL(k) = |\frac{DIFF(k)}{DIFF(k+1)}|$$

SAS

Example (DATA A)

```
DATA A(DROP=I J);
ARRAY D(10);
    DO I=1 TO 200;
DO J=1 TO 10;
        D(J)=RANUNI(0);
END;
        OUTPUT;
END;
    RUN;
```

Example (FASTCLUS)

```
PROC FASTCLUS DATA=A OUT=CLUST MAXC=1;
VAR D1-D10;
RUN;
```

R

Example (dataA)

```
dataA <- data.frame(matrix(NA, nrow = 200, ncol = 10))
dataA[,1:10] = runif(200, min = 0, max = 1)
```

Example (dataA)

```
prediction_strength(dataA)
gap_statistic(as.matrix(dataA))
```

- Similar steps can be done for other simulated data.

References

Tibshirani, Robert, and Guenther Walther. (2005)
cluster validation by prediction strength
Journal of Computational and Graphical Statistics 14.3 (2005): 511-528.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. (2001)
Estimating the number of clusters in a dataset via the Gap statistic
Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.