Cluster Metrics

Danny McClanahan

January 6, 2017

Precision and Recall

Homogeneity

- How similar is each element in a given cluster?
  - How well does the cluster predict the elements within it?

Specificity

- How different is each cluster from the others?
  - How significant is the difference / how reliably do we get a significant difference across one or more features between two elements in separate clusters?

Examples

1. Jaccard Similarity
2. F-measure
3. Pearson corr/cross prods
   - gives direction/strength of association between two datasets
   - a *low* strength means a *more* specific (and therefore a "better" clustering)

Descriptivity

- How accurately does the clustering model the sample population?
  - How likely is the clustering's assumed generative model (usually a GMM) to have produced the output?
  - Can we characterize which instances the clustering models the sample well and which it does not?

Examples

1. Kolmogorov-Smirnov (K-S)
2. **(?)** Fisher information

Compression

- How well does the clustering compress the dataset?
- How complex is the data within a single cluster?
  - How much information does each cluster contain about the rest of the data (or vice versa)?

Examples

1. Shannon Entropy Estimation
2. Minimum Description Length (MDL)
3. (Estimation of) Kolmogorov Complexity / Entropy
4. Absolute Mutual Information