"Adapting the Right Measures for K-means Clustering"

A critique of the KDD'09 paper by Junjie Wu, Hui Xiong and Jian Chen

Justin Yap

MATH3346 Data Mining Talk
Mathematical Sciences Institute
The Australian National University

29 October 2009

Introduction

- Wu, Xiong and Chen evaluate the performance of 16 validation measures for K-means clustering (e.g. entropy, mutual information, classification error etc).
- Criteria based upon whether certain properties are satisfied, sensitivity to differences in the data and the ability to detect misclassification
- Measures are shown to be identical, equivalent or improvements upon other measures.
- Measures are normalised, and it is verified that this improves their performance.

Validation Measure Properties

- Mathematical properties:
    1. Symmetry (swapping actual classes and predicted clusters)
    2. N-Invariance (multiplying the confusion matrix by a constant)
    3. Convex-Additivity (convex combinations of partitions of data)
    4. Left-Domain-Completeness (0 when cols and rows of conf. matrix are statistically independent)
    5. Right-Domain-Completeness (1 when clustering matches classes)
- Sensitivity to differences in the data.
- Ability to capture the optimal cluster size.

Measure Equivalence and Normalisation

- Many validation measures were shown to be identical or equivalent.
- Validation measures normalised, e.g.

$$S_n = \frac{S - \min(S)}{\max(S) - \min(S)} \quad \text{or} \quad S_n = \frac{S - \mathrm{E}(S)}{\max(S) - \mathrm{E}(S)}$$

- Normalisation improves performance of most measures at detecting misclassification, and makes the measures more consistent with each other.

Critical Analysis – Normalisation

- Normalisation only involved a simple affine transformation. Nonlinear monotonic transformations were not considered.
- By observing that normalised measures are more correlated with each other, the authors conclude that the normalised measures are more **robust**. It is not obvious why this is so.

K-means Clustering

- A classification method discussed in this course.
- K-means tends to create clusters of equal sizes.
- Results in **misclassification** for data with imbalanced class sizes.

	Class 1	Class 2	Class 3	Total
Cluster 1	70	2	1	73
Cluster 2	52	12	3	67
Cluster 3	53	7	10	70
Total	109	21	20	

DCV

- Paper uses the Coefficient of Variation (CV) to measure class/cluster size imbalance. Given sizes $X = \{x_1, , \ldots, x_n\}$,

$$CV = \sigma(X)/\overline{X}.$$

- The difference in CV for the class size CV_0 and cluster sizes CV_1 from K-means, $DCV = CV_1 - CV_0$, gives a measure for this type of misclassification.
- For the matrix shown in the last slide, $DCV = 0.48 - 1.02 = -0.54$.

Detection of K-means Misclassification

- The validation measures were applied to clusters that were poorly and well classified where the class sizes were imbalanced (high DCV).
- Performance based upon whether the validation measures could correctly score the clustering.
- The **correlation** between the validation measures and DCV was calculated using various real and simulated data sets.

Best Performing Validation Measure

- The *normalised van Dongen criterion*:

$$VD_n : \quad \frac{2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij}}{2n - \max_i x_i - \max_j x_j}$$

  where n is the confusion matrix, $\{x_i\}$ are the class sizes.
- Chosen because it is easy to compute, satisfies the mathematical properties and performs well for imbalanced class distributions.
- Not sensitive to data differences, which can be a disadvantage.

Critical Analysis – Narrowness of Study

- Performance based upon **narrow** criteria (detection of misclassification by K-means due to imbalanced class distributions).
- Other classification methods and criteria may yield different results.
- No analysis related to accuracy, false positive rate, true negative rate etc.
- Single examples were used to justify conclusions.