Entropy adjustment for model selection in automated clustering

Ildar Baimuratov

September 23, 2022

Abstract

In this research, the problem of model selection for automated clustering is considered. To overcome the theoretical impossibility of a universal clustering measure, and to select a clustering model in an agnostic manner, it is proposed to evaluate a clustering only as a partition, not depending on data points. We consider entropy as such measure, but to prevent entropy-based model selection from trivial results we apply the method of correction for chance. We propose a new randomness model for entropy adjustment, as randomness models for adjusted mutual information are inapplicable to it. We illustrate experimentally that the optimum of adjusted entropy is not trivial and compare adjusted entropy with Silhouette and Davies-Bouldin scores. With relatively small true numbers of clusters adjusted entropy outperforms other metrics.