Normalized entropy adjustment for model selection in automated clustering

Ildar Baimuratov

October 11, 2022

Abstract

In this research, the problem of model selection for automated clustering is considered. A normalized entropy based convex function for cluster labels partition evaluation is proposed. As this function does not depend on data points, it is a metaheuristic and has to be accompanied with a clustering quality index. To make the normalized information entropy convex, the method of adjustment for chance with a new randomness model is applied, as the randomness models for the adjusted mutual information are inapplicable. The adjusted normalized entropy used for selecting between clustering models optimized with clustering indices is compared with model selection based on the constant usage of an index, and with model selection based on vanilla entropy measures. Silhouette, Davies-Bouldin, and Calinski-Harabasz clustering indices are considerd. The results are evaluated with an accuracy of the cluster number selection regarding the true labels. The experiments on 8 datasets show that in average, the error of the adjusted normalized entropy is 0.375, while the best of the clustering indices has the error 0.625, and vanilla normalized entropy has the error 1.