Quality- and Diversity-Aware Clustering for Subclass-Inclusive Classification

Abstract

Clustering plays a fundamental role in unsupervised learning, with K-means being one of the most widely used algorithms due to its simplicity and effectiveness in partitioning data into meaningful groups. However, the performance of K-means is highly sensitive to the choice of initial seeds, commonly referred to as seed selection. Poor initialization can result in suboptimal clustering outcomes. These issues are especially pronounced in high-dimensional and large-scale datasets.

Motivated by these limitations, the thesis explores the role of seed selection in improving clustering quality and examines its downstream impact on robust classification. We introduce a *seed selection strategy* for K-means based on a repulsive point process. This method jointly captures both the *diversity* and *quality* of potential centroids, improving clustering outcomes. It ensures that selected seeds are both well-dispersed and representative, thereby improving clustering performance. We demonstrate that the proposed method significantly outperforms traditional methods.

The second contribution of the thesis focuses on extending the proposed seed selection technique to large-scale datasets. Unlike conventional subsampling-based scalable clustering methods that risk information loss, our approach leverages the entire dataset during initialization without sacrificing computational efficiency. We introduce a *scalable seed selection* framework that retains the benefits of diversity and quality-aware selection while significantly reducing runtime. Experiments on 8 benchmark datasets reveal that our approach consistently outperforms the state-of-the-art scalable clustering algorithms.

Finally, we address the *hidden stratification* issue to enhance classification. Many real-world classification tasks involve latent subclasses within annotated classes: the subgroups that are not explicitly labeled. The standard classification models, which are trained on coarse-grained class labels, often fail to capture these intra-class variations, leading to poor performance on underrepresented or challenging subgroups. To address this, we utilize our clustering framework to uncover these hidden subclasses and integrate the discovered subclass information into classification, which is subclass-inclusive. In summary, this thesis makes three primary contributions: (1) a novel probabilistic point process-based seed selection method for K-means clustering, (2) a scalable seed selection technique for large datasets, and (3) a subclass-aware classification framework that incorporates hidden strata uncovered by diversity-aware clustering.

Keywords: K-means clustering, seed selection, scalable clustering, hidden stratification, subclass-aware classification.