COMBINING DISSIMILARITY SPACES TO IMPROVE ANN SEARCH

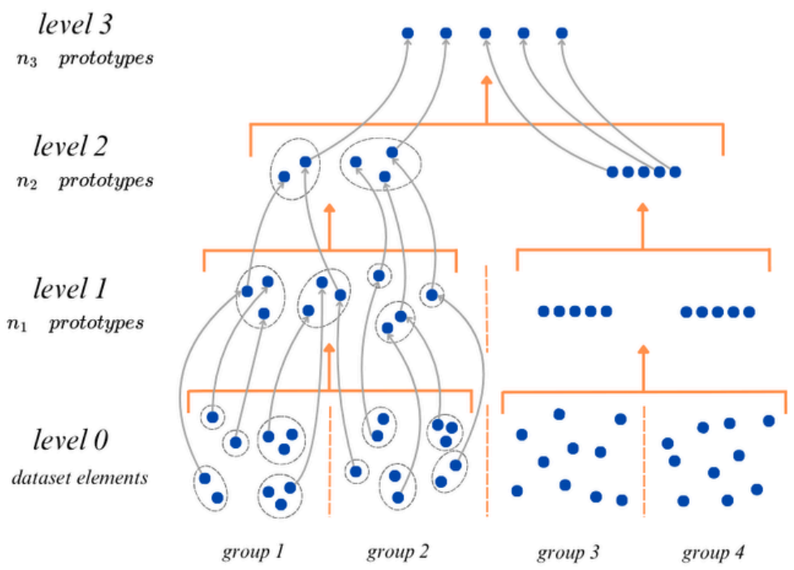
Elena García-Morato | elena.garciamorato@urjc.es

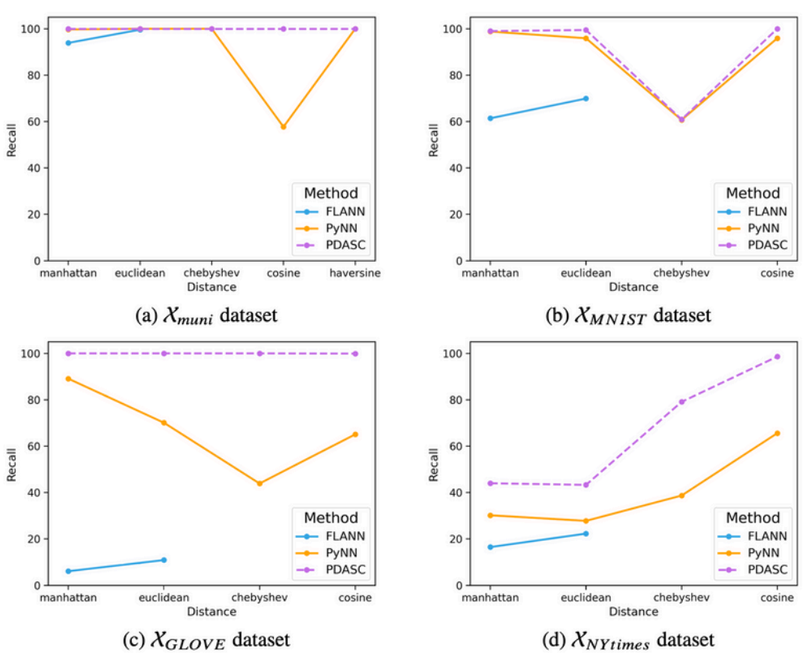
Hypothesis: Not all distance functions are equally effective for a given ANN search problem, and using the default Euclidean metric may lead to suboptimal performance.

Goal: Foster a deeper understanding of the impact of alternative distance functions in high-dimensional data contexts.

Contributions:

PDASC: A distributed indexing algorithm specifically designed for ANN search with arbitrary distance functions. It builds a hierarchical, multi-level index optimised for distributed environments, enabling scalable search even with non-metric or domain-specific distances. This flexibility is achieved through the integration of clustering algorithms, such as k-medoids, that inherently support arbitrary dissimilarity measures.





(a) $\mathcal{X}_{muni}$ dataset (b) $\mathcal{X}_{MNIST}$ dataset

(c) $\mathcal{X}_{GLOVE}$ dataset (d) $\mathcal{X}_{NYtimes}$ dataset

Distance-aware ANN: An extensive experimental evaluation of PDASC was conducted, parametrized with different distance functions and compared against state-of-the-art methods. The results demonstrate that the choice of distance function, when aligned with the characteristics of the dataset, has a significant impact on ANN search performance.

Combined-space ANN: Motivated by the *Dissimilarity Representation for Pattern Recognition*, which posits that distinct dissimilarity spaces encode complementary structural information about the data (Pekalska & Duin, 2005), an Ensemble Learning-based method that aggregates ANN solutions from multiple dissimilarity spaces is proposed.

Conclusions: **By integrating multiple dissimilarity representations, the proposed approach yields more robust and accurate ANN search results, effectively exploiting complementary information provided by diverse distance functions.**

sisap

Universidad Rey Juan Carlos