Right-Protected Data Publishing with Provable Distance-Based Mining

Abstract:

Protection of one's intellectual property is a topic with important technological and legal facets. We provide mechanisms for establishing the ownership of a dataset consisting of multiple objects. The algorithms also preserve important properties of the dataset, which are important for mining operations, and so guarantee both right protection and utility preservation. We consider a right-protection scheme based on watermarking. Watermarking may distort the original distance graph. Our watermarking methodology preserves important distance relationships, such as: the Nearest Neighbors (NN) of each object and the Minimum Spanning Tree (MST) of the original dataset. This leads to preservation of any mining operation that depends on the ordering of distances between objects, such as NN-search and classification, as well as many visualization techniques. We prove fundamental lower and upper bounds on the distance between objects post-watermarking. In particular, we establish a restricted isometry property, i.e., tight bounds on the contraction/expansion of the original distances. We use this analysis to design fast algorithms for NN-preserving and MST-preserving watermarking that drastically prune the vast search space. We observe two orders of magnitude speedup over the exhaustive schemes, without any sacrifice in NN or MST preservation.