

A Genetic Algorithm Clustering Ensemble Approach to Learning Relational Database

Abstract

Clustering is an unsupervised learning algorithm. k-Means algorithm is one of the well-known and promising clustering algorithms that can converge to a local optimum in few iterative. In our work, we will be hybridizing k-Means algorithm with Genetic Algorithms to look for the solution in the global search space in order to converge to a global optima. The problem for clustering is that when the number of clusters increases up to the same number of total records in the dataset, it leads to a scenario in which a cluster only contains a single record, and thus the cluster purity is maximized to the maximum value, 1. However, it will be useless since the common regularities among records will not be seen. Therefore, choosing the best number of clusters is trivial. Instead of choosing an inappropriate number of clusters and risking the main purpose of the clustering process, a Genetic Algorithm based k-Means ensemble is proposed in order to find the consensus result of several runs of clustering task using different number of clusters, k .