A Tabu-Search-based Algorithm for Clustering

성창섭, 진현웅
한국과학기술원 산업공학과

This paper considers a problem of clustering to partition a given data set into a certain number of natural and homogeneous subsets such that each subset is composed of data similar(if possible) to one another but different from those of any other subset. The clustering problem is NP-complete, so that an efficient heuristic procedure is desired to find a near optimum solution at appropriate time.

A Tabu-Search-based heuristic algorithm is exploited for the problem. The algorithm is constructed by combining the Tabu Search method with two functional procedures of operation, *packing* and *releasing*, where the Tabu Search procedure is employed as a sub-module. The packing and releasing operations are employed to pack each pair of data and to separate such packed pairs, respectively.

The packing operation is considered here for a role to make a drastic improvement in solution search efficiency, and also to reduce any space composed of solutions which are regarded to be *bad*. In fact, once a data is determined to change its belonging cluster for any better solution, the operation makes it possible to move together all the data packed with the cluster-changing data. On the other hand, the releasing operation has a complementary role of recovering such reduced solution space. These complementary operations help a lot each other to search for any better solution space more thoroughly, so that they can be regarded as intensification and diversification strategies, respectively.

Another important subject for the algorithm is to determine the number of required packings for the given problem, which can be determined by use of the so-called *packing property* charactered in this paper. The proposed algorithm also incorporates the so-called secondary tabu list to prevent any cycle problem of the replication of clusterings rather than the replication of mathematical solutions.

The operational procedure of the algorithm is now briefed. The algorithm first performs the operation of packing all the given data to find an initial feasible solution (composed of a fixed number of clusters). Then, the Tabu-Search procedure is applied to the initial solution for finding an improved solution by considering every possible re-adjustment for any packed data to be made into different clusters, while each initial packing-pair relation remains during such re-adjustment. The search operation continues to find the best improved solution (the best among such improved solutions) by repeating the re-adjustment operation with all the packed-data, and then the initial packing-pair relation of the worst packed pair (having the longest data-to-data distance) in the best-improved (or initial) solution is released. The best improved solution with the released packing pair becomes an initial solution for another round of the Tabu Search. In this way, the search is iteratively repeated until all the initially-packed pairs are released.

The proposed algorithm is tested for its effectiveness in comparison with three other references including *K*-means algorithm, Simulated Annealing algorithm, and Tabu Search algorithm. The test results show that the proposed algorithm outperforms the Tabu Search algorithm which is known superior to the *K*-means algorithm and the Simulated Annealing algorithm.