A VNS-based Hierarchical Clustering Method

CHIEH-YUAN TSAI* AND CHUANG-CHENG CHIU
Industrial Engineering and Management Department, Yuan-Ze University
No. 135 Yuan-Tung Road, Chungli, Taoyuan, Taiwan, R.O.C.
*

Abstract: - Traditional hierarchical clustering methods adopt a greedy strategy to merge objects progressively and construct a clustering dendrogram. However, their clustering quality might not be reliable because only local optimal information is referred during a dendrogram construction. To conquer the problem, this paper proposes a global optimal strategy to guide the dendrogram construction. The strategy aims to find an optimal circular traveling order that minimizes the total traveling distances for visiting all objects along the branches of the dendrogram, which is viewed as a traveling salesman problem (TSP). The TSP problem is solved using the variable neighborhood search (VNS) method because of its parameter-free advantage. Then, the clustering dendrogram is constructed based on the information provided by the order. Through our experiments, the clustering quality of our proposed method is superior to traditional hierarchical clustering methods.

Key-Words: - Hierarchical clustering, Traveling salesman problem, Global strategy, Variable neighborhood search.

1 Introduction

Clustering is an important data exploratory task in data mining [5]. Clustering aims at grouping objects into clusters so that the objects within a cluster have high similarity but are dissimilar to the objects in other clusters. Traditional clustering methods can be divided into two main categories of partitional and hierarchical clustering methods. Partitional clustering methods separate all objects into K clusters where the number of clusters, K, is pre-assigned according to application purpose. A partitional clustering method aims at optimizing the clustering result according a defined objective function. Minimizing the sum of distances between all objects and their corresponding cluster centers is one of the most common objective functions. K-means [11], a typical partitional clustering method, is based on an iterative scheme to reaching optimization. Besides K-means, several metaheuristic techniques have been applied to solve this partitional clustering problem, such as genetic algorithm [6], ant colony optimization [8], and particle swarm optimization [15].

Alternatively, hierarchical clustering methods conduct a series of successive merging process. A hierarchical clustering method starts by considering each object as a cluster and progressively merges them until one cluster remains. At each merging stage, two clusters which have the highest similarity are merged. The merging process can be expressed as a tree-like structure, called a dendrogram. The hierarchy of nested clustering tree can be broken at different levels to yield different numbers of clusters. A dendrogram construction example using an agglomerative clustering method is shown as Fig. 1. In Fig. 1, clusters A and B are first merged as a cluster F. After that, cluster G is generated by merging clusters D and E, then cluster H is generated by merging clusters C and F. Finally, clusters G and H are merged as cluster I.

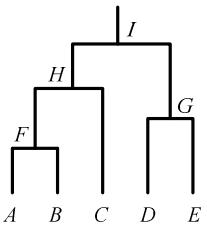


Fig. 1. Constructing a dendrogram using a hierarchical clustering method

In traditional hierarchical clustering methods, agglomerative merger process is performed based on one of three classical greedy cluster-merging strategies: single-link [13], complete-link [7], and average-link [16]. The greedy strategies are used to decide which two clusters are most similar to be merged together at each stage of the agglomerative merger process. In the single-link strategy, the dissimilarity between two clusters is defined as the minimum distance from any object in a cluster to any object in another cluster. In the complete-link strategy, by contrast, the dissimilarity between two clusters is defined as the maximum distance from any object of one cluster to any object of another

cluster. In the average-link strategy, the dissimilarity between two clusters is defined as the average distance from any object of one cluster to any object of another cluster. Although, the implementation of hierarchical clustering methods using the three greedy strategies is easy and efficient, their clustering quality may be low and not reliable because only local optimal information is referred but not global one during a dendrogram construction. For example, the single-link strategy suffers from a chaining effect problem [12], while the complete-link strategy is sensitive to outliers [3].

To conquer this problem, a new hierarchical clustering method based on a global optimal strategy is proposed in this paper. From Fig. 1, visiting the five data objects once along the branches of the dendrogram must follow the circular traveling order *A-B-C-D-E-A* in which starting and ending points are the same object *A*, shown as Fig. 2. The traveling distance based on the circular traveling order is defined as the sum of the dissimilarities between *A* and *B*, *B* and *C*, *C* and *D*, *D* and *E*, *E* and *A*. For different circular traveling orders, their traveling distances along their respective dendrograms will be different. A circular traveling order has the less traveling distance, the corresponding dendrogram can be constructed using the less cost. Therefore, instead of using traditional greedy strategies, our proposed method firstly determines an optimal circular traveling order that has the minimum traveling distance. Our proposed method then constructs the clustering dendrogram based on the optimal circular traveling order, so that we can ensure the dendrogram is constructed using the minimum cost.

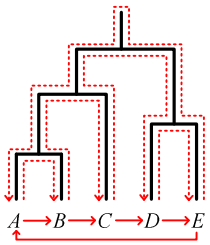


Fig. 2. Visiting the five objects along the dendrogram based on a circular traveling order *A-B-C-D-E*

2 A TSP-based hierarchical clustering method

This section presents our proposed TSP-based hierarchical clustering method in detail, including the optimal circular traveling order determination and the clustering dendrogram construction. The process of our proposed method is shown as Fig. 3.

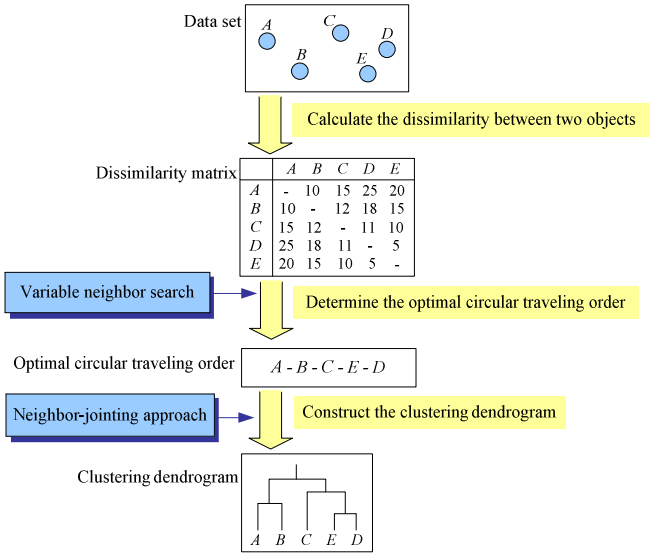


Fig. 3. The process of our proposed TSP-based hierarchical clustering method

2.1 Optimal circular traveling order determination

An object in a dataset $DS=\{x_1,\ldots, x_i,\ldots, x_n\}$ is represented as a vector of d features $x_i=\{x_{i1},\ldots, x_{ia},\ldots, x_{id}\}$ where x_{ia} represents the ath feature of the ith object x_i. The dissimilarity between two objects x_i and x_j, termed as dist(x_i, x_j), can be obtained by calculating their Euclidean distance, which is defined as Equation (1). The smaller the dist(x_i, x_j) value, the more similar the two objects x_i and x_j are. Moreover, the dissimilarities between all pairs of objects in *DS* are recorded in a dissimilarity matrix *M* where the size of *M* equals to $n \times n$.

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{a=1}^{d} (x_{ia} - x_{ja})^2} \qquad (1)$$

Our proposed hierarchical clustering method merges clusters according to an optimal circular traveling order for visiting the dendrogram with minimum cost. We consider searching the optimal order as a traveling salesman problem (TSP), one of well-known NP-hard combinational optimization problems [9]. The objective of TSP is to determine the shortest circular route passing through all cities that each city is visited exactly once. In our proposed method, each object is considered as a city. Besides, the dissimilarities between all pair-wise objects are regarded as the distances between all pair-wise cities, and are calculated using Equation (1). Accordingly, the found solution in TSP can be considered as the

optimal circular traveling order. The objective function of TSP is described as Equation (2):

$$Minimize \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \times \mathrm{dist}(x_i, x_j) \quad (2)$$

where w_{ij} = 1 if the circular route passes through objects x_i and x_j; otherwise, w_{ij} = 0. A lot of researches have applied various metaheuristic techniques to solve TSP, including ant colony optimization [2], particle swarm optimization [17], and genetic algorithm [18]. This paper applies variable neighborhood search (VNS) to determine the optimal circular traveling order in TSP since VNS has the parameter-free advantage to avoid redundant manual parameter settings. VNS, proposed by Hansen and Mladenovi´c [4], is a metaheuristic technique that explicitly applies a search strategy based on systematically changing the neighborhood structures of a solution. A set of neighborhood structures $\{N_1, \ldots, N_t, \ldots, N_{max}\}$ has to be pre-defined before performing VNS where the region size of N_t is no less than the region size of N_{t-1}. At the beginning, an initial solution S is generated as the current solution, and the neighborhood index t is initialized as one. Three main steps, including shaking, local search and move, are operated in each neighborhood structure N_t. In the shaking step a solution S_s is randomly selected to perturb the solution searching process. The solution S_s becomes a starting point of N_t in the local search step. In the local search step, all neighbor solutions of S_s are generated according to the definition of N_t. Let S_l be the optimal local solution among all neighbor solutions. Then, S_l is compared with the current solution S in the move step. If S_l is better than S, S will be replaced by S_l and the algorithm starts again with t = 1. Otherwise, t is incremented by one and a new shaking step starts again using the (t+1)th neighborhood N_{t+1}. VNS iterates the above processes until a stopping criterion is met. The stop criteria of VNS are maximum CPU time allowed, maximal iterations reached, or maximum number of iterations between two improvements.

Since a solution in TSP can be represented as a circular city-visiting sequence, its neighbor solutions can be generated by exchanging the order of some cities in the sequence. It makes the neighborhood structures in VNS be successfully pre-defined. Two neighborhood structures N_1 and N_2 are defined in our proposed VNS. A two-point swap local search method is used to find all neighbor solutions in N_1, while a two-point inversion search method is used to find all neighbor solutions in N_2. In two-point swap local search, two objects in the circular sequence are swapped. For example, as shown in Fig. 4(a), a neighbor solution of the current solution *A-B-C-D-E* is *A-D-C-B-E* if objects *B* and *D* are exchanged. Similarly, two-point inversion local search selects two objects and reverses the circular sequence between the two objects. As illustrated in Fig. 4(b), a neighbor solution of the current solution *A-B-C-D-E* is *A-E-D-C-B* if the sequence between objects *B* and *E* is inversed.

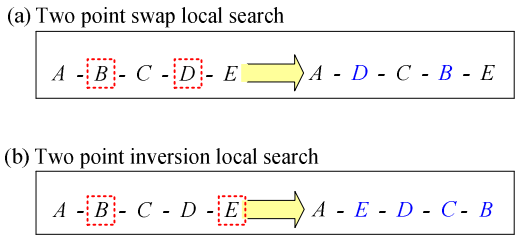


Fig. 4. An example of two-point swap and two-point inversion local search methods

In addition, the initial solution is generated randomly, and the stopping criterion is to check whether the maximum number of iterations between two improvements, termed as Q, has been met or not. The flowchart of our proposed VNS for finding the optimal circular traveling order is illustrated as Fig. 5. When the current solution S is not replaced by S_l found in neighborhood structures N_1 and N_2, S is considered as the local optimal solution among all found neighbor solutions in this iteration. Furthermore, if the current solution S is retained throughout all iterations until the stopping criterion is reached, it is considered as the optimal circular traveling order.

2.2 Clustering dendrogram construction

After the optimal circular traveling order is determined, a neighbor-jointing approach is developed to construct the dendrogram based on the information of the order. Let the optimal circular traveling order be $\pi = \{c_1, \cdots, c_{i-1}, c_i, c_{i+1}, \cdots, c_n, c_1\}$ where c_i is the ith visited object and the number of all objects is n. At the beginning of the jointing process, each object is considered as a single cluster respectively. The merger priority between two neighbor clusters depends on their dissimilarity. The smaller the dissimilarity, the higher the merger priority is. The dissimilarities between c_i and its two neighbor clusters c_{i-1} and c_{i+1} can be obtained from the dissimilarity matrix M. After two neighbor clusters are merged as a single cluster, the number of clusters in the order will be reduced by one. The

merging process will operate progressively n-1 times until all clusters are merged as one cluster, and the dendrogram is constructed completely by our proposed method. An example of the process of clustering dendrogram construction is illustrated as Fig. 6.

3 Experiment results

3.1 Visual evaluation for clustering quality

Three 2-dimensional datasets [1] are used to evaluate the clustering quality of our proposed method because the object distribution in a 2-dimensional space can be easily observed by the sense of sight. Three traditional hierarchical clustering methods of single-link, complete-link and average-link merging strategies are taken as the comparisons in the following discussion. In the first dataset DS1 (filename: ulysses22.tsp), 22 objects are grouped into two clusters where an object located at the below region can be regarded as an outlier. The clustering results of the four different clustering methods are shown in Fig. 7. It is found that our proposed method obtains the desired clustering result successfully, while other three clustering methods are influenced by the outlier object located at the lower region.

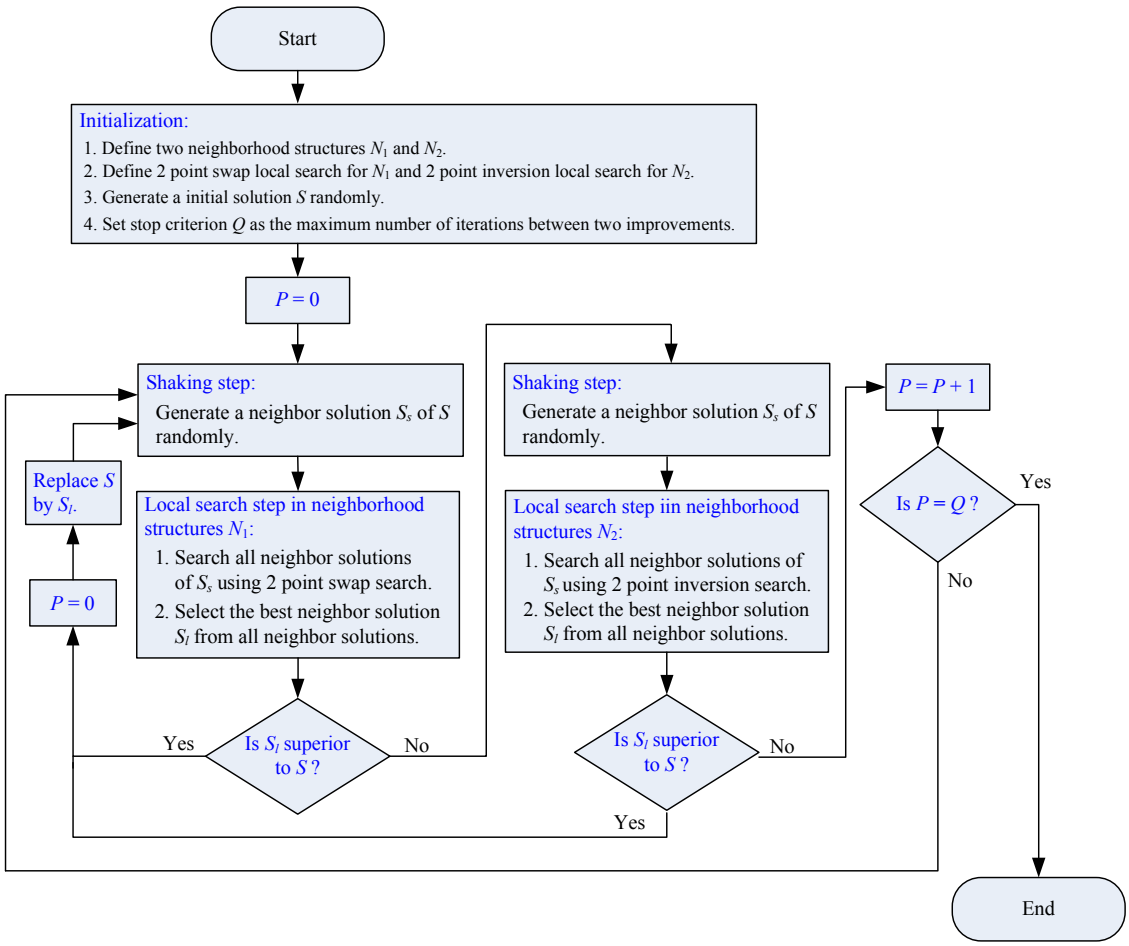


Fig. 5. The flowchart of our proposed VNS for searching optimal circular traveling order

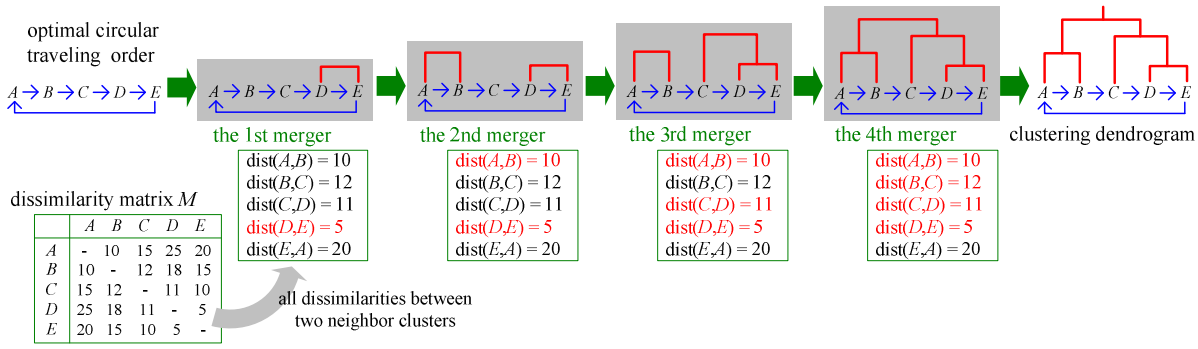
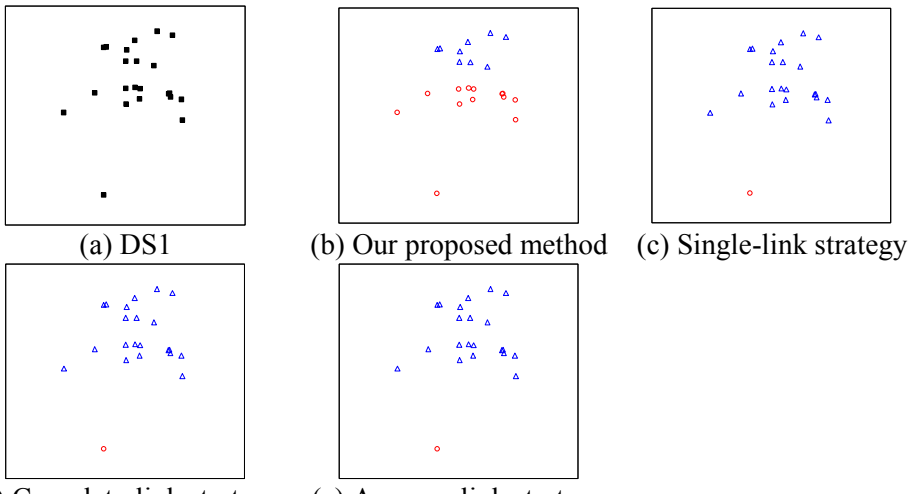


Fig. 6. An example of the process of clustering dendrogram construction

All 76 objects in DS2 (filename: pr76.tsp) are not only distributed arbitrarily but also vaguely grouped as three transverse ellipse clusters located in a top-down order. As shown in Fig. 8, the clustering result of our proposed method and the complete-link strategy are obviously superior to the one of the single-link and average-link strategies, since the two superior methods distinguish objects to three main clusters. For our method, most of objects located in the above region are grouped as the cluster marked with red circle whereas ones located in the below region are grouped as the cluster marked with blue triangle. For the complete-link strategy, however, the objects in the above region are fairly grouped as two clusters marked with red circle and green square. Similarly, the objects in the below region are averagely grouped as two "blue triangle" and "green square" clusters. It is blurred to categorize the objects in two regions into definite clusters. Therefore, the clustering quality of our method is superior to the clustering quality using complete-link strategy.

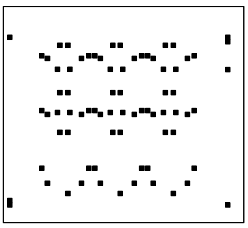
Finally, the 225 objects in the third dataset DS3 (filename: tsp225.tsp) are arranged to form three characters of T, S, and P in which each character is considered as a cluster. The clustering results using the four different hierarchical clustering methods for the dataset DS3 are shown as Fig. 9. It is found that our proposed method, complete-link strategy, and average-link strategy can obtain our desired clustering result in which each of the three characters can be identified by a specified cluster clearly.



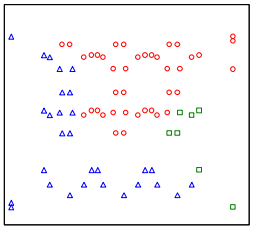
(a) DS1 (b) Our proposed method (c) Single-link strategy

(d) Complete-link strategy (e) Average-link strategy

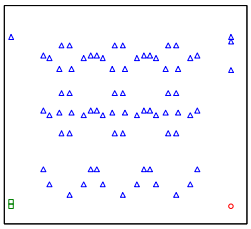
Fig. 7. Clustering results for DS1 using the four different clustering methods



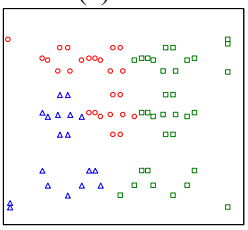
(a) DS2



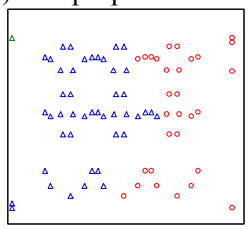
(b) Our proposed method



(c) Single-link strategy

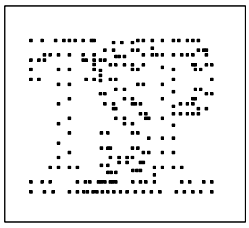


(d) Complete-link strategy

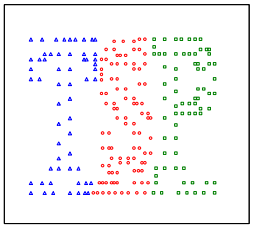


(e) Average-link strategy

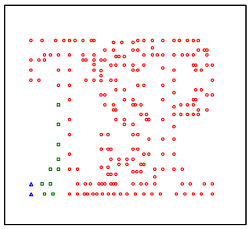
Fig. 8. Clustering results for DS2 using the four different clustering methods



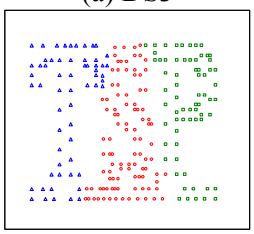
(a) DS3



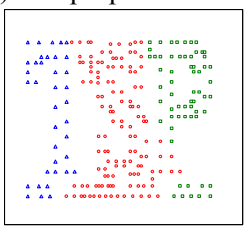
(b) Our proposed method



(c) Single-link strategy



(d) Complete-link strategy



(e) Average-link strategy

Fig. 9. Clustering results for DS3 using the four different clustering methods

Table 1. The properties of four datasets for clustering quality evaluation

Dataset	Number of objects	Number of features	Number of clusters
Glass identification	214	9	6
Iris plant	150	4	3
Pima Indians diabetes	768	8	2
Wine recognition	178	13	3

Table 2. M-B index values using four clustering methods for the four datasets

Dataset	Our proposed method		Single-link strategy	Complete-link strategy	Average-link strategy
	Mean	Standard deviation			
Glass identification	9.638	0.175	8.465	9.533	9.258
Iris plant	6.872	0.084	4.824	6.134	6.331
Pima Indians diabetes	11.253	0.389	9.290	10.667	7.644
Wine recognition	3.731	0.056	3.727	3.820	3.384
Average	7.874	0.176	6.577	7.539	6.654

(Note: Our proposed clustering method runs ten trails for each dataset.)

3.2 Cluster validity index analysis for the clustering quality

Four multi-dimensional datasets obtained from UCI Machine Learning Repository [14] including glass identification, iris plant, Pima Indians diabetes, and wine recognition datasets serve as the benchmark in this experiment. The properties of these four datasets are shown as Table 1. Moreover, we adopt M-B cluster validity index [10], defined as Equation (3), to measure the quality of a clustering result. Assume a dataset with n objects, $\{x_i|i=1,\ldots,n\}$, is partitioned into K clusters and z_k is the center of the kth cluster.

$$L(K) = (\frac{1}{K} \times \frac{E_1}{E_K} \times D_K)^2 \quad (3)$$

where $E_K = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ki} \times \text{dist}(x_i, z_k)$ and $D_K = \arg \max_{i,j=1,\cdots,K} \{\text{dist}(z_i, z_j)\}$. Here, if the x_i is the member of the kth cluster, $u_{ki} = 1$; otherwise, $u_{ki} = 0$. The larger the $L(K)$ value of a clustering method, the better the quality of its clustering result is. Because the initial solution of VNS in our proposed clustering method is randomly generated, the mean and standard deviation of the M-B index values are obtained after running ten trails for each dataset, and then are compared to other methods. As shown in Table 2, our proposed clustering method outperforms other methods in all datasets. It is noted that the standard deviation rises when the number of objects increases because the complexity for finding the optimal circular cluster-merging order also grows. Overall, our proposed clustering method is quite reliable in terms of clustering quality.

4 Conclusions

Traditional hierarchical clustering methods adopt a greedy strategy to progressively merge objects and construct a clustering dendrogram. Their clustering quality is not reliable because only local optimal information is referred during a dendrogram construction. To conquer this problem, this paper proposes a global optimal strategy to guide the dendrogram construction. The strategy aims to find an optimal circular traveling order that minimizes the total traveling distances for visiting all objects along the branches of the dendrogram. This optimal problem is transferred as a TSP problem. This paper applies VNS method to solve the TSP problem because of its parameter-free advantage. Then, the dendrogram is constructed based on the optimal circular cluster-merging order found in TSP. Through our experiments, the clustering quality of our proposed clustering method is superior to traditional hierarchical clustering methods. In the future, the performance of our proposed VNS will be further increased. In particular, we will embed the sampling and generalization techniques into the original local search step of our proposed VNS in order to enhance the performance in local search step. Consequently, it will make our proposed hierarchical clustering method is applied for a large database well.

References:

[1] A Library of Sample Instances for the TSP, http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/tsp/

[2] Dorigo, M., Gambardella, L.M., Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem, *IEEE Transactions on Evolutionary Computation*, Vol. 1, 1997, pp. 53-66.

[3] Frakes, W.B., Baeza-Yates, R.A., *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, New Jersey, 1992. pp. 13-27.

[4] Hansen, P., Mladenovi´c, N., Variable Neighborhood Search: Principles and Applications, *European Journal of Operational Research*, Vol. 130, 2001, pp. 449-467.

[5] Jain, A.K., Murty, M.N., Flynn, P.J., Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, 1999, pp. 264-323.

[6] Jiang, T., Song D.M., Cluster Analysis Using Genetic Algorithms, *Proceedings of the 3rd International Conference on Signal Processing*, 1996, pp. 1277-1279.

[7] King, B., Step-Wise Clustering Procedures, *Journal of the American Statistical Association*, Vol. 69, 1967, pp. 86-101.

[8] Kuo, R.J., Wang, H.S., Hu, T.-L., Chou, S.H., Application of Ant Colony System for Clustering Analysis, *Proceedings of the 9th Annual International Conference on Industrial Engineering Theory, Applications and Practice*, 2004, pp. 55-59.

[9] Lawler, E.L., Lenstra, J.K., Rinnooy, A.H.G., Shmoys, D.B., *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*, John Wiley & Sons, New York, 1986.

[10] Maulik, U., Bandyopadhyay, S., Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, 2002, pp. 1650-1654.

[11] McQueen, J., Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.

[12] Nagy, G., State of the Art in Pattern Recognition, *Proceedings of the IEEE*, Vol. 56, 1968, pp8 36-862.

[13] Sneath, P.H.A., Sokal, R.R., *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W. H. Freeman, San Francisco, 1973.

[14] UCI Machine Learning Database Repository, http://www.ics.uci.edu/~mlearn/MLSummary.html

[15] van der Merwe, D.W., Engelbrecht, A.P., Data Clustering Using Particle Swarm Optimization, *Proceedings of the 2003 Congress on Evolutionary Computation*, 2003, pp. 215-220.

[16] Voorhees, E.M., Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval, *Information Processing and Management*, Vol. 22, 1986, pp. 465-476.

[17] Wang, K.P., Huang, L., Zhou, C.G., Wei, P., Particle Swarm Optimization for Traveling Salesman Problem, *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics*, 2003, pp. 1583-1585.

[18] Xiong**,** S., Li, C., A Distributed Genetic Algorithm to TSP, Proceedings of the 4th World Congress on Intelligent Control and Automation, 2002, pp. 1827-1830.