CHAPTER

Conclusions, Future Work 8

The works presented in this book principally focus on solving the fundamental problems of temporal data clustering tasks in close association with ensemble-learning techniques. As described earlier, there are three methodologies for temporal data clustering; model-based clustering, proximity-based clustering, and feature-based clustering. Each approach favors differently structured temporal data or types of temporal data with certain assumptions. There is nothing universal that can solve all problems, and it is important to understand the characteristics of both clustering algorithms and the target temporal data, so that the right approach can be selected for a given clustering problem. However, there are very limited amounts of prior information for most clustering tasks, making the selection of a proper clustering algorithm for certain characteristics of temporal data extremely difficult. Therefore, motivated by the divide-and-conquer principle (Chen, 2005), the research has been carried out over a range of aspects of the three categorized temporal-clustering approaches: model-based clustering, proximity-based clustering, and feature-based clustering. Using a wide range of background knowledge and supportive study, we identify each approach's different characteristics based on their strengths and weaknesses. Model-based approaches such as Hidden Markov Model (HMM) have outstanding ability in modeling the dynamic behaviors of temporal data, but they critically suffer from detecting the intrinsic number of clusters as model-selection problem and high computational costs. Although the proximity-based approach gives a simple method for directly applying conventional static data−clustering algorithms to temporal data with appropriate similarity measures, it still presents a real challenge in general temporal data−clustering tasks due to several distinct characteristics of temporal data, such as its high dimensionality, complex time dependency, and large volume (Sun and Giles, 2000). Representation-based approaches reduce the computational complexities for high dimensional temporal data. However, in normal circumstances, the process of feature extraction always causes the loss of information in the original temporal data, and extracted features cannot fully capture the characteristics of temporal data. Furthermore, the model-selection problem remains critical. In order to overcome the problems inherent in these temporal-clustering approaches, three clustering ensemble models have been specifically proposed by author, and the

Temporal Data Mining via Unsupervised Ensemble Learning. http://dx.doi.org/10.1016/B978-0-12-811654-8.00008-7

experimental results reported in this book demonstrate the robustness and feasibility of the proposed solutions.

Initially we proposed a novel *HMM-based meta-clustering ensemble* model from the perspective of a model-based approach, which consists of three modules; flat partitional clustering, a clustering ensemble, and finally, agglomerative meta-clustering. In the flat partitional clustering module, various partitions of target temporal data are generated by HMM-based K-models clustering (Butler, 2003) with different model initializations. In the clustering ensemble module, three consensus functions Cluster-based Similarity Partitioning Algorithm (CSPA), Hyper-Graph partitioning Algorithm (HGPA), and Meta-CLustering Algorithm (MCLA) (Strehl and Ghosh, 2003) are individually applied, combining the collection of multiple partitions in forming the final partition candidates from different perspectives. Moreover, the mutual information-based objective (Strehl and Ghosh, 2003) is used to determine an optimal partition, the final ensemble partition, from these candidates. Then, applying the proposed Dendrogram-based Similarity Partitioning Algorithm (DSPA) consensus function (Yang and Chen, 2006) on target temporal data, the intrinsic number of clusters is automatically determined. In the final module of HMM-agglomerative clustering, by use of the symmetric version of *Boundary*KL distance measure, the intercluster distance is calculated and compared. The closest pair of clusters is merged to form a composite model. The process of merging clusters is repeated until a determined number of clusters is reached. This model-based clustering ensemble approach results in four major benefits:

1. No parameter re-estimation is required for the new merged pair of clusters, significantly reducing computation costs, which has been typically justified in a similar model-based hybrid-clustering approach proposed by Zhong and Ghosh (2003). As described in Section 5.2.2, the clusters of optimal consensus partition obtained from clustering ensemble are treated as meta-data, then standard HMM-based agglomerative clustering is applied to group the meta-data, where the distance between meta-data as intercluster distance is calculated and compared by using the symmetric version of *Boundary*KL distance measure, the closest pair of clusters is merged to form a composite model concatenating the model parameters of each clusters instead of re-estimating the parameters of merged clusters.
2. In comparison to single model such as HMM-based hybrid partitional-hierarchical clustering, the composite model is better equipped to characterize complexly structured clusters in order to produce a robust and accurate clustering results, which has been demonstrated on a various temporal data sets including HMM-generated data set shown in Table 5.1, a general synthetic data set (CBF) shown in Table 5.2, and a collection of time series benchmarks shown in Table 5.4.
3. The model initialization problem is solved by implementing the ensemble technique, which has been typically investigated by an experimental study described in Section 5.3 on both of HMM-generated data set and a general

synthetic data set (CBF) in comparison with other similar model-based clustering algorithms. As shown in Tables 5.1 and 5.2, the higher averaged classification accuracy with smaller standard deviation obtained by the proposed approach just demonstrated its insensitivity to model initialization.

4. The appropriate cluster number (model selection) can be automatically determined by applying a proposed consensus function DSPA (Yang and Chen, 2006) on multiple partitions of the target temporal data during the ensemble process. This automatic model selection ability has been systematically testified on HMM-generated data set shown in Fig. 5.2, a general synthetic data set (CBF) shown in Fig. 5.5, and a collection of time series benchmark shown in Table 5.4. Compared with standard model selection approach Bayesian information criterion (BIC) on HMM-generated data set shown in Fig. 5.3 and a general synthetic data set (CBF) shown in Fig. 5.6, the better performance of our proposed ensemble model has become obvious.

Although this algorithm is able to reduce the computational cost in the final module of agglomerative clustering due to no parameter re-estimation for composite models, it is still quite time consuming in comparison with other proposed algorithms, which has been systematically analyzed in computational complexity and further demonstrated on CAVIAR database shown in Table 7.4.

Therefore, we proposed an *Iteratively constructed clustering ensemble* in order to reduce the computational cost and provide a meaningful combination of input partitions by a hybrid sampling technique. Basically this approach iteratively constructs multiple partitions on the subset of whole input instances, selected by a hybrid combination of boosting and bagging sampling schemes. Weights over instances are updated for each iteration, and part of the subtraining set is chosen according to weights over instances as a selection probability. The rest of the subtraining set is constructed using random sampling from the input space. Then, a basic clustering algorithm is applied to partition the subtraining set, and the final output of the clustering ensemble is obtained by aggregating this iterative construction. Four major benefits become very clear:

1. By subsampling, computational costs are significantly reduced during the training process, which has been demonstrated by a comparison of three proposed clustering ensemble models based on computational complexity and execution time on motion trajectories database (CAVIAR) shown in Table 7.4.
2. There is an improved diversity of input partitions obtained from the training subset. As shown in Figs. 6.1 and 6.3, a *Normalized Mutual Information (NMI)* criterion has been used to measure diversity of input partitions produced by the proposed approach, restarted K-means, bagging and boosting, respectively, on Gaussian-generated data set and a general synthetic data set (CBF).
3. For complex-structured data set such as time series and large data set, the major strengths of both boosting and bagging are greatly employed for solving clustering problems, where dealing with clustered data with unbalanced populations, arbitrary shapes, and large volume are simplified by sampling

techniques, and certain difficult problems are divided into several interacted simple tasks. It has been initially demonstrated on a Gaussian-generated 2D-data set as the motivation described in Section 6.2.1 and a general synthetic data set (CBF) with a visualization and better understanding on the experiment results shown in Figs. 6.1 and 6.3, respectively. Then a set of experiments on time series benchmarks shown in Table 6.2 and motion trajectories database (CAVIAR) shown in Fig. 6.6 were carried out in Section 6.3. The experimental results have demonstrated that the proposed algorithms generally have a better performance than either boosting or bagging algorithm.

4. We can directly apply most conventional static data clustering algorithms as a base learner. As a general framework for ensemble learning, hierarchical, K-NN, and K-means have been employed as the base learner of the proposed clustering ensemble model; each of them has shown the promising results on a collection of time series benchmark in Table 6.2.

However, this proposed clustering ensemble model has the major difficulty of dealing with the data sets with various lengths, which requires the data sets to be uniform length, and combining the input partitions with different number of clusters, where the input partitions must have identical number of clusters due to the limitation of majority voting combination.

Finally, we proposed a *Weighted clustering ensemble with multiple representations* in order to provide an alternative solution to solve the common problems such as selection of intrinsic cluster numbers, computational cost, and combination method raised by both former proposed clustering ensemble models from the perspective of a feature-based approach. The approach consists of three phases of work. First, temporal data are transformed into a different feature space and become the input for the clustering algorithm. Second, the clustering algorithm is applied for clustering analyses. Finally, clustering ensemble on different representations are employed, and the weighted consensus function, based on three different clustering validity criterion—Modified Huber's T Index (Theodoridis et al., 1999), Dunn's Validity Index (Davies and Bouldin, 1979), and NMI (Vinh et al., 2009)—is carried out to find out an optimal single consensus partition from multiple partitions based on different representations. Then, a final agreement function is used to construct the final partition from the candidates yielded by the weighted consensus function based on different clustering validity criterion. This proposed representation-based clustering ensemble model results in four major benefits:

1. Through representation, the complex structures of temporal data with variable length and high dimensionality are transformed into lower-fixed dimensional feature spaces, significantly reducing computational burden, which has been demonstrated on the motion trajectories database (CAVIAR) in terms of execution time shown in Table 7.4.
2. We see a high capability for capturing the properties of temporal data as well as the synergy of reconciling diverse partitions with different representations, which has been initially demonstrated on a synthetic 2D-data set as the

motivation described in Section 7.2.1 with a visualization and better understanding on the experiment results shown in Fig. 7.1. Moreover, a set of experiments on time series benchmark shown in Table 7.1 and motion trajectories database (CAVIAR) shown in Fig. 7.5 demonstrated the benefit of using different representations in comparison of solely using single representation.

3. The weighted consensus function has outstanding ability in automatic model selection and appropriate grouping for complex temporal data, which has been initially demonstrated on a complex Gaussian-generated 2D-data set shown in Fig. 7.2 as the motivation described in Section 7.2.1, then a set of experiments on time series benchmarks shown in Table 7.1 in comparison with standards temporal data clustering algorithms, Table 7.2 in comparison with three state-of-the-art ensemble learning algorithms, Table 7.3 in comparison with other proposed clustering ensemble models on motion trajectories database (CAVIAR).
4. There is enhanced flexibility in association with most of existing clustering algorithms. As a general framework for ensemble learning, K-means, hierarchical, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) have been employed as the base learner of this proposed clustering ensemble model; each of them has shown the promising results on a collection of time series benchmark shown in Table 7.1. Also the proposed clustering ensemble model has been successfully applied for online time-series data streaming clustering, which has demonstrated on the Physiological Data Modeling Contest Workshop data set in Table 7.6 and Fig. 7.7.

Although there are some achievements made on the temporal data mining during last decade, there remain several open theoretical questions we can try to answer and research directions to follow in the future.

1. From the perspective of representation-based temporal clustering, the exploration of effective yet complementary representations in association with the clustering ensemble is a difficult task when applied to various structured temporal data.
2. For model-based temporal clustering, it is clearly important to choose a suitable model family, for example, the HMM, a mixture of first-order Markov chain (Smyth, 1999), dynamic Bayesian networks (Murphy, 2002), or the autoregressive moving average model (Xiong and Yeung, 2002). The choice is made according to the best representation of differently structured temporal data. Subsequently constructed is the suitable similarity measure applied to the specified model family. Moreover, Expectation Maximization (EM) algorithm (Chang, 2002) is used for model parameter estimation, causing problems of local optima and convergence difficulty.
3. With a discrete optimization problem approach, during each run of the clustering ensemble, the base learner constructs a "best" partition on the subset of the target data set (subsampling) by optimizing a predefined clustering quality

measure. However, the appropriate partition will better approximate the underlying data space of the target data set (ground truth) than will the "best" partition, which is treated as an over fitting problem.

4. The clustering objective function (clustering quality measure) is the core of any clustering algorithm. It is extremely difficult to design such internal criterion without supervision information.
5. An effective data clustering approach requires a minimum amount of user-dependent parameters. However, most current clustering algorithms always require several key input parameters in order to produce optimal clustering results. They are, therefore, unfeasible for use in real-world applications.