CHAPTER

5 HMM-Based Hybrid Meta-Clustering in Association With Ensemble Technique

CHAPTER OUTLINE

5.1 INTRODUCTION

For model-based clustering approaches, we consider that each time series is generated by some kind of model or by a mixture of underlying probability distributions. Temporal data are considered similar when the models characterizing individual data or the remaining residuals after fitting the model are similar, based on the likelihood distance measures. The model type is often specified a priori, such as the Gaussian (Banfield and Raftery, 1993) or Hidden Markov models (Panuccio et al., 2009). The model structure (e.g., the number of hidden states in a Hidden Markov Model [HMM]) can be determined by model-selection techniques. While model parameters can be estimated by using maximum likelihood algorithms, for example, the expectation-maximization (EM) algorithm (Bilmes, 1998). The entire process of such approaches therefore aims to estimate the optimal parameters of components models with maximum log-likelihood.

HMM, an important model-based approach for temporal data clustering, has been studied for the last decade. Although HMM has outstanding ability in capturing temporal features whose values change significantly during the observation period, thereby satisfying the Markov property, it still suffers from the critical model-selection problems of finding the appropriate HMM model configuration and selecting the intrinsic number of clusters.

Temporal Data Mining via Unsupervised Ensemble Learning. http://dx.doi.org/10.1016/B978-0-12-811654-8.00005-1

In this chapter, we therefore present HMM-based hybrid meta-clustering in association with ensemble technique for temporal data. In this approach, clustering ensemble technique is used to tackle both the model-selection and initialization problems, and the hybrid meta-clustering aims to, at once, improve clustering results and reduce computational cost. The proposed approach yields favorable results, having been demonstrated on a set of data sets.

The rest of chapter will be presented as follows. First, we state the motivation of our proposed approach and follow it with model description. Then, the experimental results based on the various data sets including HMM-generated data set, synthetic time series (Cylinder-Bell-Funnel [CBF]), a collection of time series benchmarks, and motion trajectories database (CAVIAR) are reported in the simulation section. Finally, we conclude our proposed approach and discuss future work.

5.2 HMM-BASED HYBRID META-CLUSTERING ENSEMBLE

In this section, we briefly analyze HMM model-based clustering algorithms that are related to our proposed model. Based on their weaknesses and strengths, we then describe our motivation for proposing our model. Then, the model description is presented in detail.

5.2.1 MOTIVATION

Generally, standard model-based clustering algorithms can be classified within two major approaches; model-based partitioning and hierarchical clustering. For the standard HMM-clustering algorithms, the HMM-based K-models should be categorized into the model-based partitioning approach, and HMM-based agglomerative and divisive clustering should be categorized into the model-based hierarchical approach. According to early studies on these algorithms, we can deduce that each has a different weakness when applied to temporal data clustering. For instance, HMM-based K-models suffer from the model-selection and initialization problems inherent in conventional K-means algorithms. Although HMM-based agglomerative and divisive clustering can avoid the initialization problem, they still incur high computational costs in temporal data—clustering tasks. However, a model-based hybrid partitioning-hierarchical clustering and its variants (Zhong and Ghosh, 2003) have been proposed to combine the strengths of partitioning and hierarchical approaches.

As mentioned in Section 3.2.4, HMM-based hybrid partitional-hierarchical clustering (Zhong and Ghosh, 2003) is essentially an improved version of HMM-based agglomerative clustering, keeping some hierarchical structure. By associating with HMM-based K-models clustering, the complexity of input data for the agglomerative clustering is relatively reduced, requiring less computational cost. However, the initial flat-partitioned clusters are still produced by HMM-based K-models as a generative of K-means. Therefore, this hybrid clustering still suffers from the initialization problem existed in K-means algorithm and leaves the major model-selection problem (selecting an intrinsic number of clusters) unresolved. On the other hand, HMM-based hierarchical meta-clustering (Zhong and Ghosh, 2003) makes further improvements on the previous model, which provides two major benefits for temporal data clustering. Firstly, no re-estimation of merged models is required because the composite model can be represented by the parameters of its children. Therefore, computational complexity is significantly reduced. Second, a composite model more efficiently captures the character of the complex structure of cluster than a single model which is more difficult to define and train. However, the meta-data as inputs for agglomerative clustering are obtained by using HMM-based K-models, which still causes the initialization problem and still suffers from the model-selection problem.

Although both algorithms have a number of strength for temporal data—clustering tasks, they still crucially suffer from the initialization sensitivity and model-selection problem existed in most of common model-based approaches. Hence, motivated by our earlier studies on unsupervised ensemble learning (Yang and Chen, 2006, 2007) for temporal data—clustering problems, a novel model-based clustering approach is proposed by associating HMM-based hierarchical meta-clustering with ensemble technique, where initialization problem would be addressed by combining the multiple partitions obtained by HMM-based K-models into a single robust consensus partition as inputs of agglomerative clustering, and the appropriate cluster number can be automatically determined by applying proposed consensus function, Dendrogram-based Similarity Partitioning Algorithm (DSPA) (Yang and Chen, 2006) to the multiple partitions of target temporal data during the ensemble process. Moreover, this proposed algorithm still inherent excellence from original HHM-based hierarchical meta-clustering such as that composite model has an outstanding ability to characterize the complex-structured clusters, and no parameter re-estimation is required for the new merged pair of clusters, serving to reduce computational cost.

5.2.2 MODEL DESCRIPTION

The proposed approach is modified version of HMM-based hierarchical meta-clustering described in Section 3.2.4, which consists of three modules; that is, HMM-based K-models clustering, clustering ensemble, and HMM-based agglomerative clustering. In this approach, the ensemble learning technique is implemented to overcome the initialization problem caused by HMM-based K-models clustering during initial clustering analysis. By associating with proposed consensus function (DSPA), the intrinsic number of clusters would be automatically determined during the clustering ensemble phase. In addition to Fig. 5.1, an algorithm description is given below:

1. Produce a set of partitions by applying HMM-based K-models clustering on the target data under different initialization conditions and randomly select a K value from a preset range, $K^* - 2 \leq K \leq 2K^* + 1$ $(K > 0)$.
2. Combine the collection of multiple partitions obtained by HMM-based K-models clustering to form the consensus partitions based on three consensus functions (Cluster-based Similarity Partitioning Algorithm -CSPA, Hyper Graph-Partitioning Algorithm -HGPA, and Meta-CLustering Algorithm -MCLA) described in Section 4.5.1.1.

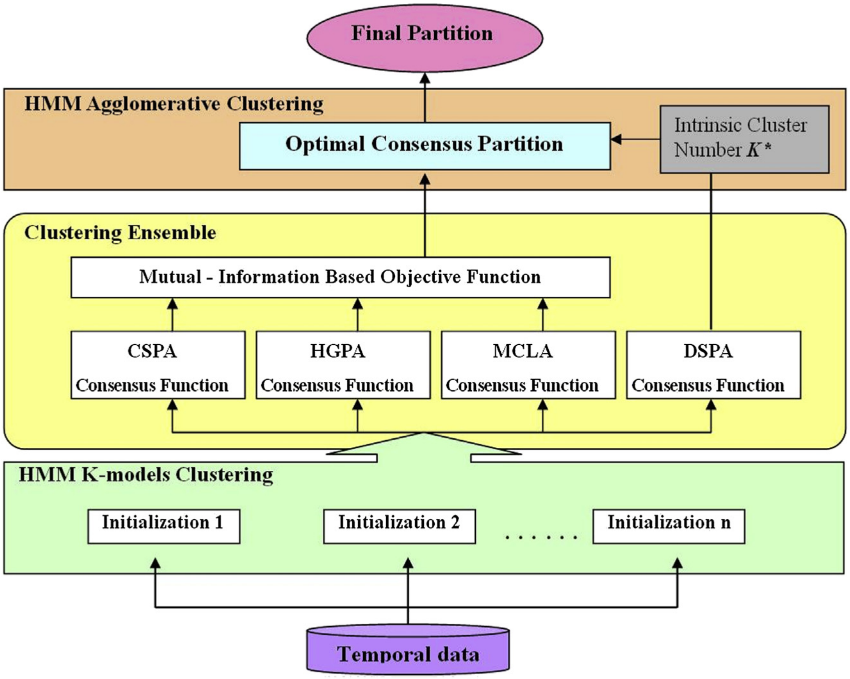

FIGURE 5.1

HMM-based hybrid meta-clustering associating with ensemble technique.

3. Apply DSPA consensus function detailed in Section 4.5.1.2 separately onto the collection of multiple partitions in order to determine the intrinsic number of clusters K'.
4. Select an optimal consensus partition $P_{optimal}$ from three candidates by using the mutual information—based objective function detailed in Section 4.5.2.
5. Re-estimate HMM model parameters for the optimal consensus partition $P_{optimal}$ by using EM algorithm.
6. Calculate the intercluster distances of optimal consensus partition by using a symmetric version of BoundaryKL distance measure $D^{boundary}(\lambda_i, \lambda_j) = \frac{1}{|x|}\sum_{x\in C_i\cup C_j}(\log p(x|\lambda_i) - \log p(x|\lambda_j))$ where $(C_i,C_j) \in P_{optimal}$ represent cluster i and j in optimal consensus partition $P_{optimal}$, respectively. The closest pair of clusters is merged to form a composite model $\lambda_{i,j} = \{\lambda_i, \lambda_j\}$, and the distance between two composite models is defined as $D(\lambda_a, \lambda_b) = \frac{1}{|\lambda_a|\times|\lambda_b|}\sum_{\lambda'_a\in\lambda_a}\sum_{\lambda'_b\in\lambda_b} D^{boundary}\left(\lambda'_a, \lambda'_b\right)$. The process of merging clusters is repeated until the determined number of clusters K' reached.

A pseudocode is also given.

```
Input:
• a data set $X = \{x_1, x_2, \ldots, x_N\}$.
• an integer $K^*$ (intrinsic cluster number)
• an integer $T$ (number of partitions)
• an integer $M$ (number of HMM model states)
• the HMM-based K-models clustering HMMkm
• the expectation-maximization (EM) algorithm EM
• A NMI based objective function NMI
• Clustering ensemble ENSEMBLE

for $t = 1$ to $T$
    $K = RAND\{K^* - 2 \le K \le 2K^* + 1\}$
    $P_t = HMMkm(K, M, X)$
end for
$P = \{P_1, P_2, \ldots P_T\}$
$[P_{CSPA}, P_{HGPA}, P_{MCLA}] = ENSEMBLE(P)$
$K' = DSPA(P)$
$[P_{optimal}, K_{optimal}] = NMI(P_{CSPA}, P_{HGPA}, P_{MCLA})$
$\lambda_{optimal} = EM(P_{optimal}, M)$
$K_{agg} = K_{optimal}$
while $K_{agg} > K'$
    $D(\lambda_a, \lambda_b) = \min_{(\lambda_a, \lambda_b) \in \lambda_{optimal}} \frac{1}{|\lambda_a| \times |\lambda_b|} \sum_{\lambda'_a \in \lambda_a} \sum_{\lambda'_b \in \lambda_b} D^{boundary}\left(\lambda'_a, \lambda'_b\right)$ where
    $D^{boundary}(\lambda_i, \lambda_j) = \frac{1}{|x|} \sum_{x \in C_i \cup C_j} (\log p(x|\lambda_i) - \log p(x|\lambda_j))$ and $(C_i, C_j) \in P_{optimal}$
    Update clustering structure
    $\lambda_{a,b} = \{\lambda_a, \lambda_a\}$;    $C_{a,b} = \{C_a, C_a\}$;
    $\lambda_{optimal} = \lambda_{optimal} / \{\lambda_a, \lambda_a\} \cap \lambda_{a,b}$
    $P_{optimal} = P_{optimal} / \{C_a, C_a\} \cap C_{a,b}$
    $K_{agg} = K_{agg} - 1$;
end while
$P_{agg} = P_{optimal}$
Output:
the final clustering $P_{agg}$
```

5.3 SIMULATION

In this section, we evaluate the performance of our proposed clustering ensemble model for solving initialization and model-selection problems. The proposed ensemble model results are reported and compared with similar methods on HMM-generated data set, synthetic time series named CBF, time series benchmark, and CAVIAR database of motion trajectories.

5.3.1 HMM-GENERATED DATA SET

In the first phase of experiments, we evaluate the performance of our approach in comparison to various relative algorithms. Standard HMM-clustering algorithms including K-models, agglomerative and divisive clustering, HMM-based hybrid clustering, and HMM-based hybrid meta-clustering (detailed in Section 3.2.4), are concurrently applied on the HMM-generated data set.

Taking a similar approach to the work (Smyth, 1997), a total of 200 HMM-generated data are produced from a mixture of four HMM components, and each continuous HMM model generates 50 sequences with an identical length of 200. The generation function is stated as,

$$p(X|\lambda) = 0.25p(X|\lambda_1) + 0.25p(X|\lambda_2) + 0.25p(X|\lambda_3) + 0.25p(X|\lambda_4) \tag{5.1}$$

$$\lambda_k = \left(\pi_k, A_k, \mu_k, \sigma_k^2\right) \tag{5.2}$$

where each HMM model has two hidden states with transition parameters $A_1 = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$, $A_2 = \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}$, $A_3 = \begin{bmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{bmatrix}$, $A_4 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$, the emission distribution corresponding to each state is distributed as a single Gaussian, mean $\mu_k^1 = 3$, variance $\sigma_k^{12} = 1$ for state 1, and mean $\mu_k^2 = 0$, variance $\sigma_k^{2^2} = 1$ for state 2. The initial state probabilities π_k are randomly generated by uniform distribution.

The clustering performance depends heavily on the selection of input parameter K. However, this important input parameter K cannot be automatically determined by the most of algorithms themselves and needs to be selected manually. In contrast, our proposed clustering ensemble model runs the HMM-K-models for initial clustering analysis with number of states—four and a randomly selected K value as cluster number from a preset range $K^* - 1 \leq K \leq K^* + 2$ $(K > 0)$ instead of manually selected number of clusters, which produces 10 partitions on the target data set. These input partitions are used to construct the final partition, and the cluster number of final partition is automatically determined by DSPA consensus function. Fig. 5.2 represents the dendrogram produced by applying DSPA consensus function on 10 input partitions, where correct number of clusters represented by four different colored subtree is automatically detected by cutting the dendrogram at a range of threshold values corresponding to the longest cluster's lifetime.

In comparison with the standard model-selection approach, we also apply our approach on the target data set by trying the fixed cluster size in a range of $2 \leq K \leq 10$ on restart and using the Bayesian information criterion (BIC). As illustrated in Fig. 5.3, the optimal number of clusters is selected as seven with a minimum value of BIC, which failed to determine the correct number of clusters ($K^* = 4$).

By prespecifying the number of states $M = 2$ and cluster number $K^* = 4$, clustering algorithms including HMM-K-models, HMM-agglomerative, HMM-divisive clustering, HMM-based hybrid clustering, and HMM-based hybrid meta-clustering are also applied to the target synthetic data with different model initializations. In contrast, our proposed clustering ensemble model runs the HMM K-models for initial

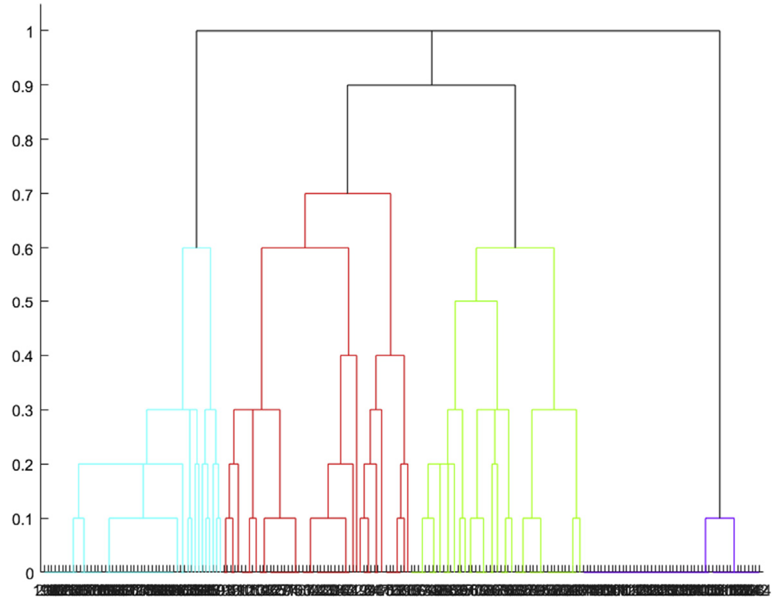

FIGURE 5.2

Dendrogram (HMM-generated data set).

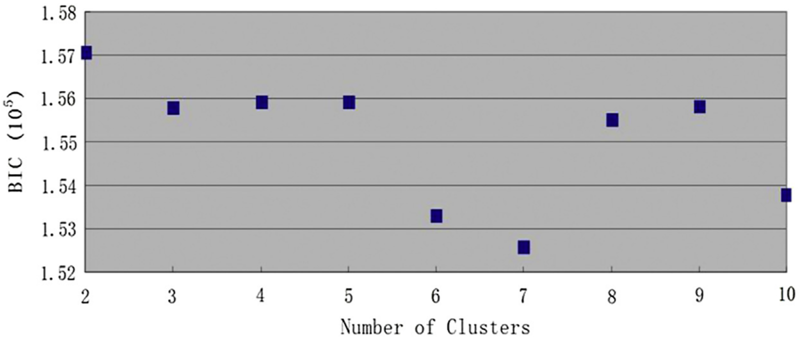

FIGURE 5.3

BIC on different number of clusters (HMM-generated data set).

clustering analysis with a randomly selected K value as cluster number from a preset range $K^* - 1 \leq K \leq K^* + 2$ $(K > 0)$ under different initialization conditions which produces 10 partitions on the target data set to produce the final partition. Each clustering algorithm is run 10 times. Average classification accuracy is presented in the form of mean ± standard deviation and illustrated in Table 5.1.

The top part of the Table shows the results for algorithms without model-selection ability from where the higher valued standard deviation in the

Table 5.1 Classification Accuracy (%) of Our HMM-Based Hybrid Meta-Clustering Ensemble on HMM-Generated Data Set

Clustering Algorithms	Average Classification Accuracy (Mean ± Std)
HMM K-model	73.2 ± 5.5
HMM agglomerative	71.1 ± 2.1
HMM divisive	68.5 ± 3.9
HMM hybrid clustering	73.8 ± 3.9
HMM hybrid meta-clustering	74.2 ± 2.6
Our approach	**83.1 ± 1.8**

Bold denotes best result

classification rate indicates that the performances of these clustering algorithms are unstabilized due to the model initialization problem. In contrast, the last row of Table 5.1 shows the averaged classification accuracy of proposed clustering ensemble model on the HMM-generated data with the higher averaged classification accuracy and smaller standard deviation, which indicates that our proposed ensemble model with model-selection ability is insensitive to the initialization.

5.3.2 CBF DATA SET

In the second experiment, we are going to evaluate the performance of our approach for the general temporal data—clustering tasks by using a synthetic time series. This data set has been used as a benchmark in temporal data mining (Keogh and Kasetty, 2003). As illustrated in Fig. 5.4, this data set is a 1-D time series named CBF consisting of three classes of data, cylinder (c), bell (b), or funnel (f). Although this data set is originally designed for supervised classification, we can use it for the purpose of testing the proposed unsupervised clustering approach. The data are generated by three time series functions:

$$c(t) = (6 + k)x_{[a,b]}(t) + \varepsilon(t), \tag{5.3}$$

$$b(t) = (6 + k)x_{[a,b]}(t)(t - a)/(b - a) + \varepsilon(t), \tag{5.4}$$

$$f(t) = (6 + k)x_{[a,b]}(t)(t - b)/(b - a) + \varepsilon(t). \tag{5.5}$$

where k and $\varepsilon(t)$ are drawn from the normal distribution $N(0,1)$, a and b are two integers randomly drawn from intervals [16, 32] and [48, 128], and $x_{[a,b]}(t)$ is defined as 1 if $b \leq t \leq a$ and 0 otherwise. Three stochastic functions in (5.3-5.5) randomly generate a time series of 128 frames corresponding to three classes: cylinder, bell, and funnel. In our simulations, we generate 100 samples for each class and the whole data set contains 300 samples in total.

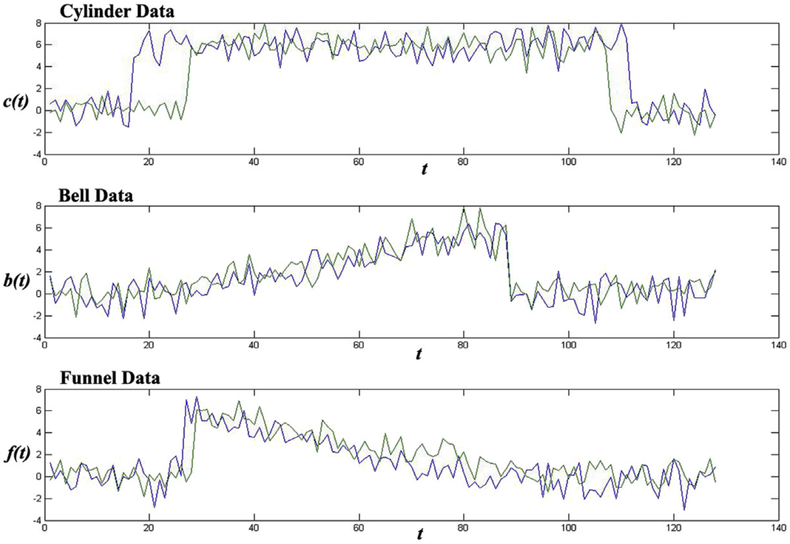

FIGURE 5.4

Cylinder-bell-funnel data set.

Following the same experiment setup in the first part of simulation, the performance of model selection based on the DSPA consensus function is compared with standard model-selection approach by applying our approach on the CBF data set with all cluster size ($2 \leq K \leq 10$) and using BIC model-selection criteria to detect the optimal number of clusters. In order to achieve the best parameter setup based on the target data set, the stated number of HMM models is set to seven by an exhaustive search. As illustrated in Fig. 5.5, the DSPA consensus automatically detects the correct number of clusters ($K^* = 3$) again represented in three different colored subtree. However, the wrong number of clusters ($K = 4$) is obtained by standard model-selection approach with a minimum value of BIC shown in Fig. 5.6. As a result, we trust that our approach based on Dendrogram-based Similarity Partitioning Algorithm (DSPA) consensus function has a better performance of model selection than the standard approach.

In order to compare the performance between our approach and relative HMM-based clustering algorithms, five clustering algorithms evaluated in the first part of the simulation are also applied to the CBF data set with the optimal number of states $M = 7$ and cluster number $K^* = 3$. For the proposed approach associated with ensemble technique, HMM k-models clustering initially produces various partitions of CBF data with different initialization and random selection of cluster numbers from a range $K* - 1 \leq K \leq K* + 2$ $(K > 0)$. Then, three consensus functions (CSPA, HGPA, and MCLA) are applied to yield respective consensus partitions. Subsequently, the mutual information–based objective function determines the optimal consensus partition. The fourth consensus function DSPA is used to

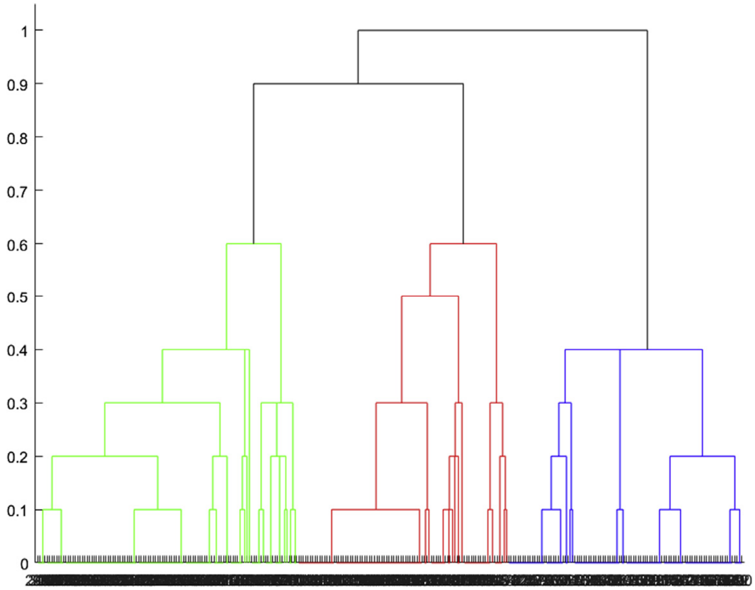

FIGURE 5.5

Dendrogram (Cylinder-bell-funnel data set).

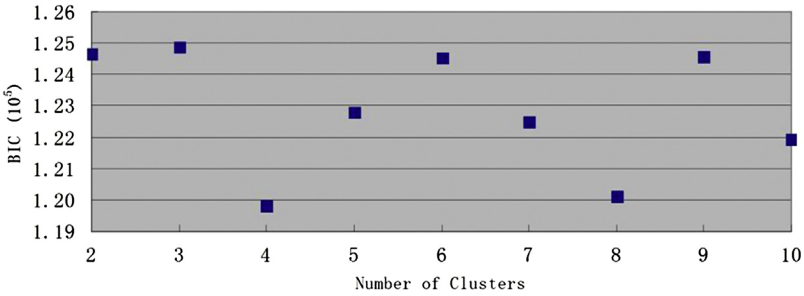

FIGURE 5.6

BIC on different number of clusters (Cylinder-bell-funnel data set).

automatically select the cluster number K^*. Finally, both the optimal consensus partitions obtained from the ensemble of HMM k-models clustering and the selected cluster number K^* are used as the input of HMM-agglomerative clustering to produce the final partition for the CBF data. We run each of clustering algorithms 10 times on the CBF data to obtain its average classification accuracy. As shown in Table 5.2, our approach once again yields a favorable result on the CBF data set when compared to the relative clustering algorithms, even given the best parameter setup (optimal number of states and correct number of clusters), which once again demonstrates the efficiency of our approach to solve model-selection and initialization problems for general temporal data—clustering tasks.

Table 5.2 Classification Accuracy (%) of Our HMM-Based Hybrid Meta-Clustering Ensemble on CBF Data Set

Clustering Algorithms	Average Classification Accuracy (Mean ± Std)
HMM K-model	65.2 ± 4.3
HMM agglomerative	71.1 ± 3.1
HMM divisive	64.5 ± 3.1
HMM- based hybrid clustering	69.8 ± 2.9
HMM hybrid meta-clustering	70.2 ± 2.5
Our approach	**77.9 ± 1.5**

Bold denotes best result

5.3.3 TIME SERIES BENCHMARKS

The previous experiment examines the legitimacy of the fundamental concept under certain assumptions, where the data sets are either HMM-generated data set or simple synthetic time series. In order to evaluate the performance of our approach for more general temporal data—clustering tasks, we implement another experiment by using a collection of several time series data set.

Time series benchmarks of 16 synthetic or real-world time series data sets (Keogh, 2003) have been collected to evaluate a number of classification and clustering algorithms in the context of temporal data mining. In this collection, the ground truth, that is, the class label of time series in a data set, is available, and each data set is further divided into the training and testing subsets in advance for the evaluation of a classification algorithm. The information on all 16 data sets is tabulated in Table 5.3, including the number of classes, the number of time series, and the length of time series in every data set. Although the outcome of clustering analysis can be used for miscellaneous tasks, we focus on only clustering-based classification tasks in simulations. By given the ground truth, the classification rate is defined as the ratio of the number of time series of the same class label that are grouped together into the same cluster to the overall number of time series of the data set.

In this experiment, we employ K-means, dynamic time warping (DTW)-based K-means, HMM K-model, HMM hybrid clustering, HMM hybrid meta-clustering, and our approach on the benchmark collection, where the performance of the K-means algorithm is provided by benchmark collectors. The class number of each data set, K^*, used in the algorithms includes K-means, DTW-based K-means, HMM K-model, HMM hybrid clustering, and HMM hybrid meta-clustering. In contrast, there is not the user input of class number in our approach due to its automatic model-selection ability. For our approach, we did not use the information of the class number of each data set and simply select K value of cluster number

Table 5.3 Time Series Benchmark Information

Data Set	Number of Class K^*	Size of Data Set (Training + Testing)	Length
Syn control	6	300 ± 300	60
Gun point	2	50 ± 150	150
CBF	3	30 ± 900	128
Face (all)	14	560 ± 1690	131
OSU leaf	6	200 ± 242	427
Swedish leaf	15	500 ± 625	128
50 words	50	450 ± 455	270
Trace	4	100 ± 100	275
Two patterns	4	1000 ± 4000	128
Wafer	2	1000 ± 6174	152
Face (four)	4	24 ± 88	350
Lightning-2	2	60 ± 61	637
Lightning- 7	7	70 ± 73	319
ECG	2	100 ± 100	96
Adiac	37	390 ± 391	176
Yoga	2	300 ± 3000	426

from a preset range $K^* - 2 \leq K \leq K^* + 2$ $(K > 0)$. In order to compare model-selection ability based on our proposed DPSA consensus function with the standard model-selection approach, we also modified our approach with BIC model selection, which determines the optimal number of clusters from a range of clusters size, with a minimum value of BIC. For HMM-based clustering algorithms, the state number of HMM models are critical for the performance of modeling the time series. However, these information are not given for time series benchmarks. Therefore, we could only employ an exhaustive search to determine optimal number of states for each of data set included in the time series benchmarks, which are correspondingly 6,2,4,9,2,6,3,10,8,8,9,6,7,8,2,3.

Table 5.4 lists all the results achieved in this experiment. For each algorithm, we run the experiment 10 times with best parameter setup, and the best result is reported in the table. It is observed from Table 5.4 that there is not any algorithm that had had outstanding performance than others. Comparably, our approach outperforms other methods, as it has the best performance on 6 of 16 data sets. DTW-based K-means surprisingly achieves best results for five data sets, the HMM K-model, HMM hybrid clustering, and HMM hybrid meta-clustering algorithms win on two, one and two data sets, respectively. Given the fact that our approach based on both of BIC model selection and DSPA consensus function are capable for finding a cluster number in a given data set, we also report their model-selection performance with the notation that * is added behind the classification accuracy if the algorithm finds

Table 5.4 Classification Accuracy (%)[a] of Clustering Algorithms on Time Series Benchmarks

Data Set	K-Means	DTW-Based K-means	HMM K-Model	HMM Hybrid Clustering	HMM Hybrid Meta-Clustering	Our Approach (BIC Model Selection)	Our Approach
Syn control	67.9	69.8	69.1	69.8	71.1	70.8*	**73.2***
Gun point	50.0	**65.6**	43.8	51.8	50.0	56.0*	65.2*
CBF	62.6	**80.9**	60.1	63.2	65.2	70.0*	64.3*
Face (all)	36.0	**49.4**	37.8	36.4	39.2	23.6	31.4
OSU leaf	37.8	35.1	44.2	40.8	**45.1**	35.0	38.0
Swedish leaf	40.6	48.1	38.6	47.6	**49.2**	28.3	42.5*
50 words	42.0	37.2	40.8	38.9	41.0	39.1	**46.2***
Trace	48.5	63.4	50.9	56.3	59.8	61.8*	**63.9***
Two patterns	32.2	**56.3**	33.1	35.2	38.1	27.2	50.6*
Wafer	62.5	47.5	63.9	**65.1**	62.9	54.2*	53.4
Face (four)	66.9	**70.7**	69.1	64.2	61.9	49.7	58.8
Lightning-2	61.1	62.1	57.7	63.2	66.8	51.6	**67.6***
Lightning-7	48.4	50.5	**51.2**	47.3	45.3	49.5	50.0*
ECG	69.8	62.8	**70.3**	61.6	63.3	55.5	65.8
Adiac	38.4	39.6	38.9	42.0	40.2	38.9*	**43.2***
Yoga	51.7	56.3	48.5	44.3	47.1	34.8	**63.8***

[a] *Notation of correct cluster number determined*

the correct cluster numbers. As a result, our approach based on proposed DPSA consensus function is able to find the correct cluster number on 11 of 16 data sets, and BIC model-selection criteria manage to find the correct cluster number for six data sets only, as shown in Table 5.4. The results indicate the challenge of standard model-selection criteria such as BIC in clustering temporal data of high dimensions. However, our approach based on proposed DPSA consensus function achieves promising results in terms of model selection and proper grouping. It is worth mentioning that all HMM-based algorithms reported in the experiments take a much longer time in comparison with K-means algorithm. The computational burden becomes quite significant when applying HMM-based clustering algorithms on the temporal data set with large volume and high dimensionality, which would be a major weakness of such approachs.

5.3.4 MOTION TRAJECTORY

For evaluating our approach in the real-world application, we apply our approach to the CAVIAR database for trajectory clustering analysis. The CAVIAR database is a benchmark designed for video content analysis (CAVIAR, 2002). From the manually annotated video sequences of pedestrians, a set of 222 high-quality motion trajectories are achieved for clustering analysis without concerning the affection of a visual-tracking algorithm. Fig. 5.7 illustrates the collection of all the motion

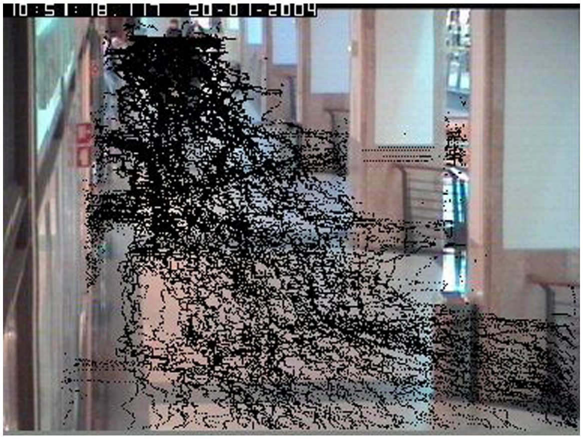

FIGURE 5.7

All motion trajectories in the CAVIAR database.

trajectories in the database. A motion trajectory is a 2D spatiotemporal data of the notation $\{(x(t), y(t))\}_{t=1}^{T}$, where $(x(t), y(t))$ are the coordinates of an object tracked at frame t.

Given that there is no prior knowledge on the "right" number of HMM model states and clusters for this database, we manually determine the optimal state number as two by an exhaust search and run the HMM K-models as base leaner 10 times by choosing a K value from an interval between 5 and 10 with random HMM model initialization, then 10 partitions generated are fed to our clustering ensemble to yield three consensus partition candidates, finally the optimal consensus partition is obtained from three candidates by a refining process based on HMM-agglomerative clustering algorithm, and the optimal cluster number as eight in the final partition is automatically determined by proposed DPSA consensus function.

Without the ground truth, human visual inspection has to be applied for evaluating the results, as suggested in the study by Khalid and Naftel (2005). As observable in Fig. 5.8, coherent motion trajectories in a similar path have been properly grouped together while dissimilar ones are distributed into different clusters. From the camera viewpoint, the trajectories corresponding to "move horizontally" following the same path are grouped very well into a single cluster shown in Fig. 5.8A, while the "move vertically" trajectories at different location are grouped into two clusters shown in Fig. 5.8C and F. Trajectories corresponding to "walk and watch" movement are properly grouped into a single clusters, as shown in Fig. 5.8G. Fig. 5.8D—E indicate that trajectories corresponding to most activities of "enter and enter the store" are properly grouped together via multiple clusters in light of motion path. Finally, Fig. 5.8H illustrates the cluster roughly corresponding to the activity "pass in front". It is worth to mention that our approach is insensitive to distinguish the trajectories following a similar motion part with opposite direction. For example,

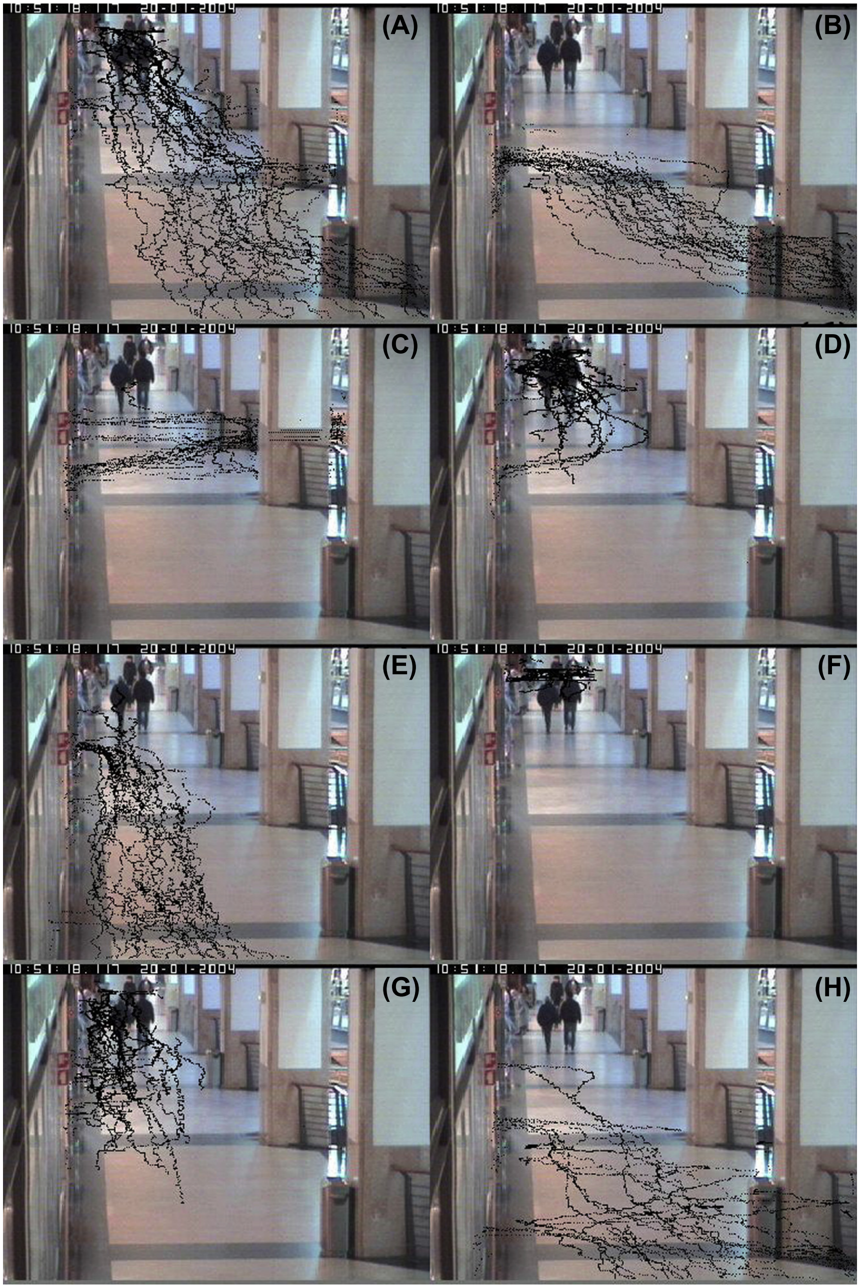

FIGURE 5.8

A clustering analysis of all moving trajectories on the CAVIAR database made by HMM-based meta-clustering ensemble model. Plots in A—H correspond to 8 clusters of moving trajectories in the final partition.

the trajectories corresponding to “move up” and “move down” are merged into same cluster of “move horizontally” shown in Fig. 5.8A, while “move left” and “move right” trajectories at different locations are grouped into two clusters of “move vertically” shown in Fig. 5.8C and F. However, results of the clustering analysis obtained by our approach still can be used to infer different activities based on the motion path at a semantic level.

As mentioned in the previous experiments, we experienced that the HMM-based clustering algorithms are quite time consuming, especially on the temporal data with large volume and high dimensionality. In order to compare the computational cost of HMM-based clustering ensemble with other proposed approaches presented in the latter chapters, we initially record the execution time of applying our proposed HMM-based clustering ensemble presented in this chapter on the CAVIAR database for trajectory clustering analysis, which results in 4623.44 s.

For further evaluation, we have performed one more additional experiment in classification by adding different amounts of Gaussian noise $N(0,\sigma)$, to the range of coordinates and removed five segments of trajectory of the identical length at random locations in order to simulate common scenarios that a moving object tracked is occluded by other objects or the background so that a tracking algorithm has to produce a trajectory with missing data and recorded that motion trajectory data are interfered by external noise. Corrupted trajectories are used for testing where a decision is made by finding a cluster whose corresponding HMM model represents the tested trajectory in terms of the maximum log-likelihood to see if its clean version belongs to this cluster. Apparently, the classification accuracy depends largely on the quality of clustering analysis. Fig. 5.9 shows performance evolution in the presence of missing data measured by a percentage of the trajectory length added

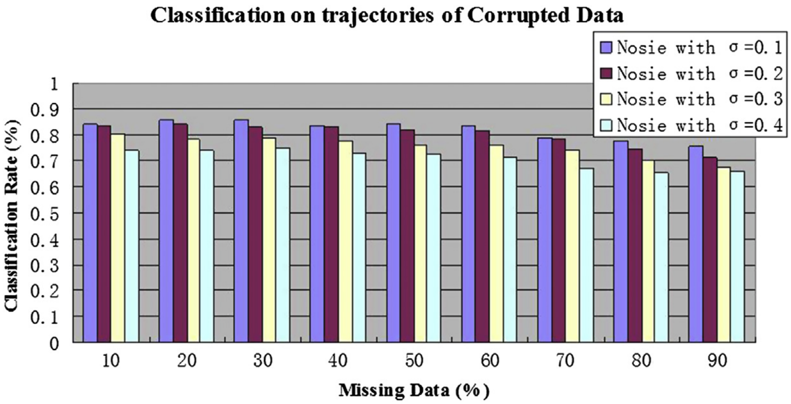

FIGURE 5.9

Performance of the HMM-based meta-clustering ensemble model on CAVIAR with corrupted data appears.

by different amounts of noise. It is evident that our approach performs well in real-world situations.

In summary, all the previously mentioned simulation results suggest that our model leads to robust clustering analysis. Therefore, the results may be used for higher level video content analyses.

5.4 SUMMARY

In this chapter, we have presented HMM-based hybrid meta-clustering in association with ensemble technique for temporal data. Based on our experimental results, we conclude that our proposed model yields robust clustering results and hence is suitable for application in an unknown environment, such as one with no prior information about the cluster number.

Essentially, our approach is proposed by modifying the HMM-based hierarchical meta-clustering described in previous Section 3.2.4 in order to overcome the problems of HMM model initialization sensitivity and selection of cluster number existed in the original approach. Therefore, our approach still inherent two major benefits from original approach by combining the strengths of partitioning and hierarchical approaches, which has been justified by Zhong and Ghosh (2003). First, no parameter re-estimation is required for the new merged pair of clusters which maximally reduces computation cost. As described in Section 3.2.2, 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 BoundaryKL distance measure based on log-likelihood, the closest pair of clusters is merged to form a composite model concatenating the model parameters of each cluster instead of re-estimating the parameters of merged clusters. Second, the composite model is better equipped to characterize complexly structured clusters in comparison of single model such as HMM-based hybrid partitional-hierarchical clustering. The robust and accurate clustering performance obtained by our proposed approach has been demonstrated on 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. Moreover, our approach is also beneficial from insensitivity of HMM model initialization and automatic selection of cluster number by associating clustering ensemble techniques. As demonstrated in the early experiments by running each of clustering algorithms (HMM-K-models, HMM-agglomerative, HMM-divisive clustering, HMM-based hybrid clustering, HMM-based hybrid meta-clustering, and our approach) 10 times on the HMM-generated data set and CBF data set, our approach results the highest averaged classification rate, and the smallest valued standard deviation indicates that our approach is insensitive to the HMM model initialization in comparison of other HMM-based clustering algorithms. Furthermore, the experimental results reported in Table 5.4 demonstrate that DPSA consensus function involved in our approach obtains an efficient model-selection ability of automatically determining the cluster number; it is able to find the correct cluster number on 11 out of 16 data sets, which significantly outperforms BIC model-selection criteria that manage to find the correct cluster number for six data sets only.

However, there are some important open theoretical questions which must be considered. For model-based temporal data clustering, it is quite difficult to choose a suitable model family such as HMM, mixture of first-order Markov chain (Smyth, 1999), dynamic Bayesian networks (Murphy, 2002), or autoregressive moving average model (Xiong and Yeung, 2002), in terms of which it better represents differently structured temporal data without any prior information. For HMM-based clustering, the state emission probability is modeled as multivariate Gaussian. Here, how to select the number of Gaussian components for individual state emission function would come into question. In general terms, the multivariate Gaussian has been found to offer better performance than single Gaussian (Butler, 2003), but its use is limited for the considerable reasons of high computational demands and overfitting on the limited training data set available. On the other hand, how to determine the number of states would be critical for HMM model configuration. Moreover, the model-based clustering algorithms are usually combined with an EM algorithm for parameter estimation, even though the EM always causes the problems of local optima and convergence difficulty. For clustering algorithm, the algorithm efficiency is still a critical issue for temporal data with huge volume and high dimensionality, the computational cost of HMM-based clustering combining with ensemble techniques becomes even more expensive. As demonstrated on the CAVIAR database in simulation section, our approach is quite time consuming which could be a major limitation of applying our approach on the real-world or online applications. Therefore, how to find a tradeoff solution between computational cost and clustering performance becomes an urgent topic for model-based clustering.