CHAPTER

# Introduction 1

## CHAPTER OUTLINE



## 1.1 BACKGROUND

The unsupervised classification or clustering provides an effective way to condensing and summarizing information conveyed in data, which is demanded by a number of application areas for organizing or discovering structures in data. The objective of clustering analysis is to partition a set of unlabeled objects into groups or clusters where all the objects grouped in the same cluster should be coherent or homogeneous. There are two core problems in clustering analysis; that is, model selection and proper grouping. The former is seeking a solution that estimates the intrinsic number of clusters underlying a data set, while the latter demands a rule to group coherent objects together to form a cluster. From the perspective of machine learning, clustering analysis is an extremely difficult unsupervised learning task since it is inherently an ill-posed problem and its solution often violates some common assumptions (Kleinberg, 2003). There have been many researches in clustering analysis (Jain et al., 1999), which leads to various clustering algorithms categorized as partitioning, hierarchical, density-based, and model-based clustering algorithms.

Actually, temporal data are a collection of observations associated with information such as the time at which data have been captured and the time interval during which a data value is valid. Temporal data are composed of a sequence of nominal symbols from the alphabet known as a temporal sequence and a sequence of continuous real-valued elements known as a time series. The use of temporal data have become widespread in recent years, and temporal data mining continues to be a rapidly evolving area of interrelated disciplines including statistics, temporal pattern recognition, temporal databases, optimization, visualization, high-performance computing, and parallel computing.

**Temporal Data Mining via Unsupervised Ensemble Learning.** http://dx.doi.org/10.1016/B978-0-12-811654-8.00001-4

However, the recent empirical studies in temporal data analysis reveal that most of the existing clustering algorithms do not work well for temporal data due to their special structure and data dependency (Keogh and Kasetty, 2003), which presents a big challenge in clustering temporal data of various and high dimensionality, large volume, very high-feature correlation, and a substantial amount of noise.

Recently, several studies have attempted to improve clustering by combining multiple clustering solutions into a single-consolidated clustering ensemble for better average performance among given clustering solutions. This has led to many real-world applications, including gene classification, image segmentation (Hong et al., 2008), video retrieval, and so on (Jain et al., 1999; Fischer and Buhmann, 2003; Azimi et al., 2006). Clustering ensembles usually involve two stages. First, multiple partitions are obtained through several runs of initial clustering analysis. Subsequently, the specific consensus function is used in order to find a final consensus partition from multiple input partitions. This book is going to concentrate on ensemble learning techniques and its application for temporal data clustering tasks based on three methodologies: the model-based approach, the proximity-based approach, and the feature-based approach.

The model-based approach aims to construct statistical models to describe the characteristics of each group of temporal data, providing more intuitive ways to capture dynamic behaviors and a more flexible means for dealing with the variable lengths of temporal data. In general, the entire temporal data set is modeled by a mixture of these statistical models, while an individual statistical model such as Gaussian distribution, Poisson distribution, or Hidden Markov Model (HMM) is used to model a specific cluster of temporal data. Model-based approaches for temporal data clustering include HMM (Panuccio et al., 2009), Gaussian mixture model (Fraley and Raftery, 2002), mixture of first-order Markov chain (Smyth, 1999), dynamic Bayesian networks (Murphy, 2002), and autoregressive moving average model (Xiong and Yeung, 2002). Usually, these are combined with an expectation-maximization algorithm (Bilmes, 1998) for parameter estimation.

The proximity-based approach is mainly based on the measure of the similarity or distance between each pair of temporal data. The most common methods are agglomerative and divisive clustering (Jain et al., 1999), which partition the unlabeled objects into different groups so that members of the same groups are more alike than members of different groups based on the similarity metric. For proximity-based clustering, either the Euclidean distance or an advanced version of Mahalanobis distance (Bar-Hillel et al., 2006) would be commonly used as the basis for comparing the similarity of two sets of temporal data.

The feature-based approach is indirect temporal data clustering, which begins with the extraction of a set of features from raw temporal data, so that all temporal data can be transformed into a static feature space. Then, classical vector-based clustering algorithms can be implemented within the feature space. Obviously, feature extraction is the essential factor that decides the performance of clustering. Generally, feature-based clustering reduces the computational complexities for higher dimensional temporal data.

## 1.2 PROBLEM STATEMENT

Although the clustering algorithms have been intensively developing for last decades, due to the natural complexity of temporal data, we still face many challenges for temporal data clustering tasks.

How to select an intrinsic number of clusters is still a critical model selection problem existing in many clustering algorithms. In a statistical framework, model selection is the task of selecting a mixture of the appropriate number of mathematical models with the appropriate parameter setup that fits the target data set by optimizing some criterion. In other words, the model selection problem is solved by optimizing the predefined criterion. For common model selection criterion, Akaike information criterion, AIC (Akaike, 1974), balances the good fit of a statistical model based on maximum log-likelihood and model complexity based on the number of model parameters. The optimal number of clusters is selected with a minimum value of AIC. Based on Bayesian model selection principles, the Bayesian information criterion, BIC (Schwarz, 1978), is a similar approach to AIC. However, while the number of parameters and maximum log-likelihood are required to compute the AIC, the computation of BIC requires the number of observations and maximum log-likelihood instead. Monte-Carlo cross validation (Smyth, 1996) divides a data set into training and test sets at certain times in a random manner. In each run, the training set is used to estimate the best-fitting parameters while the testing set computes the model's error. The optimal number of clusters is selected by posteriori probabilities or criterion function. Recently, the Bayesian Ying-Yang machine (Xu, 1996) has been applied to model selection in clustering analysis (Xu, 1997). It treats the unsupervised learning problem as a problem of estimating the joint distribution between the observable pattern in the observable space and its representation pattern in the representation space. In theory, the optimal number of clusters is given by the minimum value of cost function. In addition, other criterions of model selection include minimum message length (Grunwald et al., 1998), minimum description length (Grunwald et al., 1998), and covariance inflation criterion (Tibshirani and Knight, 1999). However, recent empirical studies (Zucchini, 2000; Hu and Xu, 2003) in model selection reveal that most of the existing criterions have different limitations, which often overestimate or underestimate the cluster number. Performance of these different criterions depends on the structure of the target data set, and no single criterion emerges as outstanding when measured against the others. Moreover, a major problem associated with these model selection criterions also remains: the computation procedures involved are extremely complex and time consuming.

How to significantly reduce the computational cost is another importance issue for temporal data clustering task due to the fact of that temporal data are often collected in a data set with large volume, high and various dimensionality, and complex-clustered structure. From the perspective of model-based temporal data clustering, Zhong and Ghosh (2003) proposed a model-based hybrid partitioning-hierarchical clustering and its variance such as HHM-based hierarchical meta clustering. In the first approach, one is an improved version of model-based agglomerative clustering,

which keeps some hierarchical structure. However, associating with HMM-based K-models clustering, the complexity of input data to the agglomerative clustering is relatively reduces. Therefore, this approach requires less computational cost. Moreover, the HHM-based hierarchical meta clustering further reduces the computational cost due to no re-estimation of merged component models as a composite model. However, both of them are still quite time consuming in comparison to most proximity-based and representation-based approaches. Furthermore, the aforementioned model selection problem is still unavoidable. From the perspective of proximity-based temporal data clustering, K-means algorithm is effective in clustering large-scale data sets, and efforts have been made in order to overcome its disadvantages (Huang, 1998; Ordonez and Omiecinski, 2004), which potentially provides a clustering solution for temporal with large volume. Sampling-based approach such as Clustering LARge Applications (CLARA) (Kaufman and Rousseeuw, 1990) and Clustering Using REprisentatives (CURE) (Guha et al., 1998) reduces the computational cost by applying an appropriate sampling technique on the entire data set with large volume. Condensation-based approach such as Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) (Zhang et al., 1996) constructs the compact summaries of the original data in a Cluster Feature (CF) tree, which captures the clustering information and significantly reduces the computational burden. Density-based approach such as Densit-based Spatial Clustering of Application with Noise (DBSCAN) (Ester et al., 1996) Ordering Points To Identify the Clustering Structure (OPTICS) (Ankerst et al., 1999) is able to automatically determine the complex clustered structure by finding the dense area of data set. Although each algorithm has a good performance for clustering large volume of data set, most of them have the difficulty to deal with temporal data with various length. From the perspective of representation-based temporal data clustering, the computational cost can be significantly reduced by projecting the temporal data with various length and high dimensionality into a uniform lower dimensional representation space, where most of the existing clustering algorithms can be applied. However, our previous study (Yang and Chen, 2011a) has shown that no single representation technique could perfectly represent all the different temporal data set, each of them just capture limited amount of characters obtained from temporal data set.

How to thoroughly extract the important features from original temporal data is concerned with the representation methods. Nowadays, various representations have been proposed for temporal data (Faloutsos et al., 1994 Dimitrova and Golshani, 1995; Chen and Chang, 2000; Keogh et al., 2001; Chakrabarti et al., 2002; Bashir, 2005; Cheong et al., 2005; Bagnall et al., 2006; Gionis et al., 2007; Ding et al., 2008; Ye and Keogh, 2009), and its variants such as the multiple-scaled representation (Lin et al., 2004) continues to be proposed for improving the temporal data clustering performance. Nevertheless, one representation tends to encode only those features well presented in its representation space, which inevitably causes the loss of other useful information conveyed in the original temporal data. Due to the high complexity and varieties of temporal data, to our

knowledge, there is no universal representation that perfectly characterizes miscellaneous temporal data. Therefore, a representation is merely applicable to a class of temporal data where their salient features can be fully captured in the representation space but such information is hardly available without prior knowledge and a careful analysis. Therefore, the alternative approaches have been developed for use of different representations. They are simply lumping different representations together to constitute a composite representation (Li and Wrench, 1983; Attili et al., 1988; Openshaw et al., 1993; Colombi et al., 1996) and directly apply clustering analysis on the composite representation. Although such an approach could reduce the computational cost, it leads to a higher dimensional feature vector of redundancy, and different representations get involved in different measure criteria, and so it is a nontrivial issue on how to normalize different representations to form a composite representation.

## 1.3 OBJECTIVE OF BOOK

A large of number of recent studies have shown that unsupervised ensemble approaches improve clustering quality by combining multiple clustering solutions into a single consolidated clustering ensemble that has the best performance among given clustering solutions. As a result, ensemble learning techniques may give an optimal solution of dealing with temporal data clustering problems. During the last 10 years, the author was trying to solve the problems of temporal data clustering via ensemble learning approaches and produced some excellent research works including proposing and developing several innovation methods and algorithms. However, as he works on the subject, he has been aware that there is just a little literature of presenting unsupervised ensemble learning with a focus on its application to temporal data clustering. Therefore, he published this book to fill the need for a comprehensive guide on the subject. This book will not only give an overview of temporal data mining, in-depth knowledge of temporal data clustering, and ensemble learning techniques in an accessible format designed to appeal to students and professional researchers with little mathematical and statistical background but also have a rich blend of theory and practice with three proposed novel approaches:

- HMM, an important model-based approach for temporal data clustering, has been studied. The author proposes a novel approach based on the ensemble of HMM-based partitioning clustering associated with hierarchical clustering refinement in order to solve problems in finding the intrinsic number of clusters and model initialization problems which exist in most model-based clustering algorithms.
- Proximity-based approaches pose real challenges of computational cost in temporal data clustering due to the large volume of temporal data. In this book, author presents an unsupervised ensemble learning model of iteratively

constructed partitions on the subtraining set obtained by a hybrid sampling scheme. The proposed approach not only reduces the computational cost of temporal data clustering but also obtains a more general framework for any type of data clustering analysis, which provides a potential solution for large temporal data clustering tasks.

- Feature-based approach to temporal data clustering is proposed through a weighted ensemble of a simple clustering algorithm with minimum user-dependent parameters, such as k-means with different representations, in order to address both proper grouping with minimum computational cost and selecting an intrinsic number of clusters as model selection problems in clustering analysis as a whole. This proposed approach takes into account the diversity of partitions generated by certain clustering algorithm on different representations, initial and reconciles them in an optimal way. Furthermore, the proposed weighted consensus function not only enables automatic model selection for clustering analysis but also provides a generic technique for the optimal solution of combining multiple partitions.

Furthermore, readers will realize that each of temporal clustering approaches favor 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.

## 1.4 OVERVIEW OF BOOK

This book is enlightening for students and researchers wishing to study on temporal data mining and unsupervised ensemble learning approaches. It is not only to enumerate the existing techniques proposed so far but also to classify and organize them in a way that may be of help for a practitioner looking for solutions to a concrete problem. Furthermore, author also provides some of novel unsupervised ensemble learning approaches for temporal data clustering in this book. This book is organized as follows:

In Chapter 2, a review of temporal data mining is carried out from three aspects. Initially, representation of temporal data are reviewed, followed by similarity measures of temporal data mining based on different objectives, and then five mining tasks including prediction, classification, clustering, search and retrieval, and pattern discovery are briefly described at the end of chapter.

In Chapter 3, we present a comprehensive survey on temporal data clustering algorithms from different perspectives, which includes partitional clustering, hierarchical clustering, density-based clustering, and model-based clustering. Their strengths and weakness are also discussed for temporal data clustering tasks. Moreover, based on the internal, external, and relative criteria, most common clustering validity indices are described for quantitative evaluation of clustering quality.

In Chapter 4, a systemic literature of ensemble leaning is presented in two parts. First, we discuss the ensemble learning from three aspects: ensemble learning algorithms, combining methods, and diversity of ensemble learning. By giving in-depth knowledge about unsupervised ensemble learning, we further discuss the consensus functions and objective functions of clustering ensemble approaches.

In Chapter 5, HMM model-based framework is detailed with related works. We discuss the problems of existing HMM model-based clustering algorithms and present a novel HMM-based ensemble clustering approach. Such approach is designed to solve the problems in finding the intrinsic number of clusters and model initialization sensitivity. This approach has been compared with several similar approaches and evaluated on synthetic data, time series benchmark, and motion trajectory database and yields promising results for clustering tasks.

In Chapter 6, we initially have a brief analysis on sampling-based ensemble approaches including both boosting and bagging and identify the major differences between both of ensemble learning approaches. Then, inspired by both boosting and bagging, an iteratively constructed clustering ensemble model is proposed by combining the strengths of both boosting and bagging. The proposed approach is also evaluated on synthetic data, time series benchmark, and real-world motion trajectory data sets, and experimental results show satisfactory performance for a variety of clustering tasks.

In Chapter 7, we present a weighted clustering ensemble of multiple partitions produced by initial clustering analysis on different temporal data representations. This approach is designed to solve the problems in finding the intrinsic number of clusters, sensitivity to initialization, and combination method of ensemble learning. It also provides a tradeoff solution between computational cost and accuracy for temporal data clustering. To demonstrate effectiveness, the proposed approach is applied to a variety of temporal data clustering tasks, including benchmark time series, motion trajectory, and time-series data stream clustering. The experimental results and their analyses are stated. A detailed discussion of future works concludes this chapter.

In Chapter 8, the work presented in the book is summarized. The three proposed ensemble models are reviewed and analyzed, and then final conclusions are drawn. Unsolved problems are also discussed with regard to their potential for future research work.