CHAPTER 1

INTRODUCTION

The advancement of biomedical devices in health-care has led to the emergence of temporal data that describes the state of patient's health. These data varies over time and are stored as Electronic Health Records (EHR's).The data in EHR's are liable to several complexities such as missing value, irregular observations and time-constrained attribute set. Although, clinical time series data suffers from these temporal complexities, they also possess useful hidden knowledge. This knowledge can be used for developing a Clinical Decision Making System (CDMS) that can assist the physician in decision making task such as diagnosis, monitoring, prognosis and drug discovery.

Temporal data mining refers to the process of discovering knowledge from time series data (Batal et al. 2016; Fu 2011). The importance of temporal data mining in clinical domain was investigated in many research studies (Moskovitch & Shahar 2015a; Batal et al. 2013; Moerchen 2006). However, the presence of complexities such as missing values, irregular observations and large attribute set in clinical time series data challenges the knowledge mining process. Missing values commonly occur in EHR's due to an unobserved patient data or due to a data recording error. In clinical domain, data are often considered as irregular since the state of patient's health is usually observed at irregular time intervals.

Most of the data acquired from EHR's contains the results of laboratory and physical examination such as platelet count, blood pressure, cholesterol level, total protein, ElectroCardioGraphy (ECG), ElectroEncephaloGraphy (EEG) and wearable sensors. The results of these examinations are stored as numeric measurements at different points in time. The examinations are referred as attributes in clinical data. In general, clinical data contains large attribute set, since a patient may have undergone several common lab examinations in addition to the disease specific lab examination.

This chapter provides descriptions about the emergence of temporal data and the need for temporal data pre-processing and temporal data mining in clinical domain.

1.1 TEMPORAL DATA IN CLINICAL DOMAIN

Temporal data commonly refers to any data that varies over time (Batal et al. 2016; Miguel et al. 2013; Han & Kamber 2001). Temporal data arises in various domains like business, retail, health-care, transport, banking etc. In health-care, the temporal data is generated from medical equipment's like anatomical sensors, physiological sensors, hospital data and incidence records. The anatomical sensor describes data like movement, position and location. The physiological sensor describes data like blood pressure, heartbeat rate, ECG and EEG. The consultation records and diagnosis reports represents the hospital data and laboratory examinations. The incidence records refer to data that describes bacterial surveillance and pharmacologic surveillance. These temporal data is stored as EHR.

The clinical time series data acquired from EHR consists of set of observations that describes the patient's state of health. The EHR data sample shown in Figure 1.1 stores the laboratory examinations result for each patient with respect to the day of observation. P_1, P_2 and P_3 in the Figure 1.1, refers to

the patients who have undergone four laboratory examinations namely Albumin (ALB), Alkaliphophatase (ALP), Glutamic Oxaloacetic Transaminase (GOT) and Total Cholesterol (T-CHO). Figure 1.1 describes the EHR report for the patient's P_1, P_2 and P_3 who have undergone these laboratory examinations on the following days 19-02-08, 10-03-08, 16-04-08 and 15-05-08.

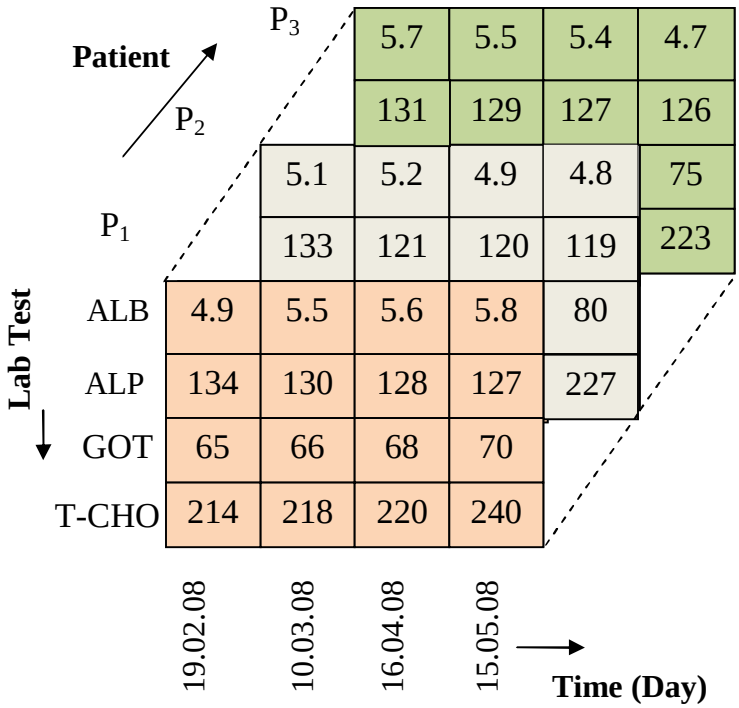


Figure 1.1 EHR Data Sample

Four types of patterns commonly occur in time series data namely trend, seasonal, cyclic and irregular variations. Figure 1.2 shows the types of pattern that commonly occurs in time series data (Cooray 2008). Trend describes the long term increase or decrease in the data. Trend is classified into upward trend, lower trend and no change. The upper trend describes the increase in the growth rate as shown in the Figure 1.2.a. The lower trend describes the decrease in the growth rate as shown in the Figure 1.2.b. The condition of no change in the trend represents that there is no increase or

decrease in the growth rate as shown in the Figure 1.2.c. Figure 1.2.d shows the seasonal pattern that refers to the patterns that are commonly occurring in fixed period (say month, year and week). The cycle pattern shown in Figure 1.2.e refers to the pattern that shows fluctuations that are not of fixed periods. The variations in cyclic patterns are recurrent rise and fall in time series that occurs normally greater than a year. Figure 1.2.f shows the irregular variations which are fluctuations that are unpredictable and are short in duration.

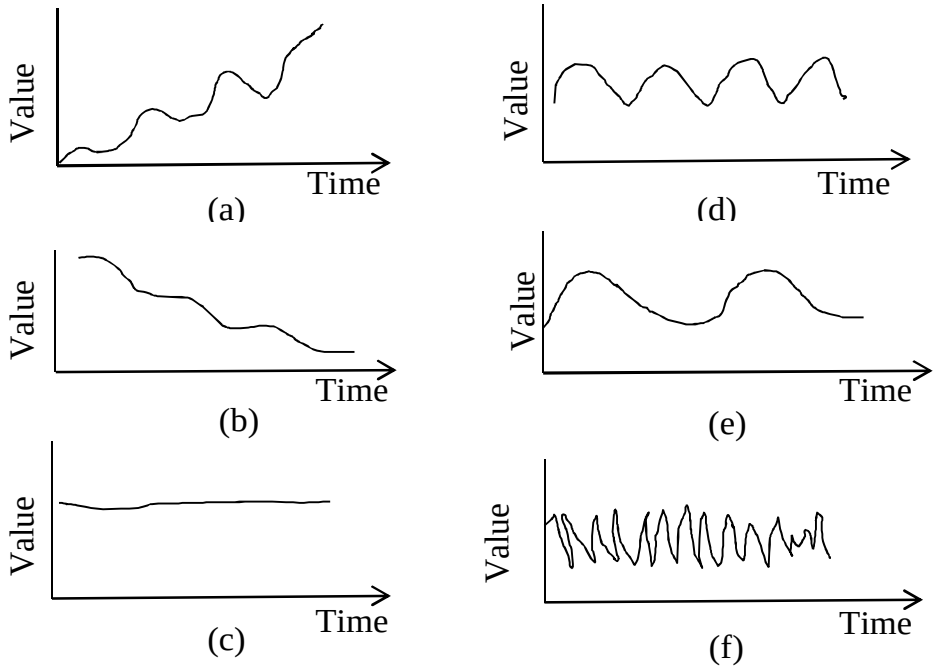
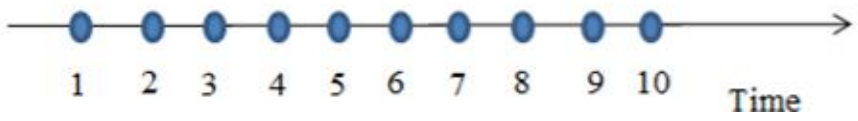


Figure 1.2 Time Series Patterns (a): Upward Trend, (b): Downward Trend, (c): No Change (d): Seasonal (e): Cycle (f): Irregular

1.1.1 Time Series Data Categories

Time series data can be categorized as univariate or multivariate. Univariate time series data is described using single temporal attribute whereas multivariate time series data consists of more than one temporal attribute. Similarly, a time series data can be considered as evenly spaced (regular) or unevenly spaced (irregular). In the evenly spaced time series data,

observations are done at regular intervals and in the unevenly spaced data observations are done at irregular intervals. Figure 1.3 depicts the evenly spaced and unevenly spaced time series.



(a)

(b)

Figure 1.3 Time Series Categories (a) Evenly Spaced (Regular) Time Series (b) Unevenly Spaced (Irregular) Time Series

Clinical time series data is often considered to be multivariate and unevenly spaced; since patients may have undergone many common laboratory examinations in addition to the disease specific laboratory examinations and these observations are usually done at irregular intervals.

1.2 TEMPORAL DATA PRE-PROCESSING

Temporal data pre-processing aims in improving the quality of time series data to make it highly suitable for data analysis and data mining. Data acquired directly from any sources are often incomplete, inconsistent and noisy (Han & Kamber 2001). Incomplete data commonly refers to the missing value of an attribute. The noisiness in the data arises due to the outliers or deviated data. Inconsistencies in the data refer to the common contradictions. Data pre-processing techniques overcomes the difficulties that arise with incomplete, noisy and inconsistent data during analyzing or mining. Hence, they are usually considered as pre-requisite step in mining process. The

temporal data pre-processing techniques can be grouped into three categories: data cleaning, data integration and transformation, temporal data representation and attribute selection.

1.2.1 Data Cleaning

Data cleaning is the process of filling missing value, identifying and correcting the corrupted or noisy data that occurs in the dataset. The occurrence of missing data is obvious in many real-life applications where there is periodic record maintenance. The impact of missing data and its management have been studied in several research studies (Cismondi et al. 2013; Ding & Ross 2012; Enders 2010; Scheuren 2005; Schafer 2007; Dempster et al. 1977; Little & Rubin 1987). Treating these missing values is considered as a vital task, since it improves the effectiveness of knowledge discovery process (Enders 2010; Ford 1983). In healthcare domain, clinical data are liable to have missing values since the observations are done for each patient at irregular intervals and the number of observations done varies for every patient.

Missing data can be classified into two categories based on its pattern and relationship between observed variable with missing data (Enders 2010; Little & Rubin 1987). First category corresponds to six patterns namely univariate pattern, unit non-response pattern, monotone pattern, general pattern, planned missing pattern and latent variable pattern. Second category classifies missing data as Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) (Little & Rubin 1987). The two common strategies for handling missing values are ignoring (deletion) and imputation (Enders 2010). There are several missing value imputation techniques namely mean, median, nearest neighbour, hot-deck, maximum likelihood, regression (Dempster et al. 1977; Little & Rubin 1987; Enders 2010).

The applicability of these missing imputation techniques in non-time series data differs from time series data, due to the presence of temporal patterns like trend, seasonal, cyclic and irregular variations in time series data. Several research works has been carried out to illustrate the importance of imputing the missing values in time series (Little & Rubin 1987). Clinical time series data is characterised by the temporal patterns, which identifies the change in the temporal sequence of observed laboratory or physical examinations for a particular disease. Thus, missing value imputation in clinical time series data becomes challenging when the observations are done irregularly.

1.2.2 Data Integration and Transformation

The presence of redundancy in the dataset makes the knowledge discovery process inefficient. Data integration focuses on reducing the redundancies and eliminating the inconsistencies that occur in the data source. Data transformation maps the input to the specified format that is appropriate for mining. The following activities are involved in data transformation namely smoothing, aggregation, discretization and normalization (Han & Kamber 2001). The smoothing process eliminates noise from data using techniques like binning, regression and clustering. In aggregation, the data are summarized using aggregate operations. Normalization is the process of mapping a large input scale to smaller range. Min-max, z-score and decimal-scaling are commonly used normalization techniques. Discretization is the process of replacing the numeric values in the attributes using conceptual or interval labels.

1.2.3 Temporal Data Representation

Temporal data representation aims at reducing the number of data points in the original time series resulting into smaller sections with no

compromise in the analytical and mining results (Lovric et al. 2014). The following are the activities carried out in time series data representation: time series dimensionality reduction, time series segmentation and temporal data abstraction.

1.2.3.1 Time series dimensionality reduction

Lin et al. (2007) have provided a hierarchy of various time series representation methods that reduces the dimension of a time series. Accordingly, the widely adopted techniques in time series representations are Piecewise Aggregate Approximation (PAA), Piecewise Linear Representation (PLR), Perceptually Important Points (PIP), Symbolic Aggregate Approximation (SAX), Shape Description Alphabet (SDA), supervised discretization process, subsequence clustering and Multiple Abstraction Level Mining (MALM). The techniques PAA, PLR, PIP represent a sampling method of numeric representation of time series, whereas techniques like SDA, supervised discretization, subsequence clustering and MALM uses symbolic form to represent the time series. The other techniques include Fourier transforms (Keogh et al. 2001) and wavelets (Chan & Fu 1999).

Wu et al. (2000) have presented a comparative analysis on two time representation methods namely discrete wavelet transforms and discrete Fourier transform. Further, the combination of wavelet and Fourier transform has been studied by Kawagoe & Ueda (2002). Though, there are several works in the literature that aims in summarizing an evenly spaced temporal data, but still abstraction in an unevenly spaced data remains challenging area of research.

1.2.3.2 Time series segmentation

Time series segmentation is the process of partitioning the time series into segments that are inter-similar. Depending on the application, the

goal of the segmentation varies such as to locate stable periods of time, to identify change points or to simply compress the original time series into a more compact representation. There are three main classical segmentation algorithms namely sliding window, top-down and bottom-up segmentation (Lovric et al. 2014; Keogh et al. 2004). Sliding window performs segmentation by finding the left boundary of the first potential segment and then it approximates the data to the right with increasing longer segments until the error of the potential segment is greater than the pre-defined threshold value. The very next point of the identified segment is considered as the left boundary for the new potential segment.

In top-down method, the time series is recursively partitioned until a stopping criterion is met (Li 1998). It works by considering every possible partitioning of the times series and splitting them at the best location. Both sub-segments are then tested to see if their approximation error is below some user-specified threshold. If not, the algorithm recursively continues to split the subsequence until all the segments have approximation errors below the threshold. Bottom-up method starts the segmentation process by dividing the time series into a large number of very small segments with equal length (Keogh & Pazzani 1998; Hunter & McIntosh 1999). Then, two successive pair which causes the smallest increase in error are found and merged into one new bigger segment. This step is repeated until some approximation error condition is satisfied or a predefined number of segments is found.

1.2.3.3 Temporal data abstraction

In clinical data analysis, temporal data abstraction and mining is a challenging area of research. Temporal abstraction refers to the process of transforming quantitative (numeric) values in each attribute to qualitative interpretations. The commonly used temporal interpretations are trend and value abstractions (Moskovitch & Shahar 2015a; Orphanou et al. 2014;

Shahar & Musen 1996; Batal et al. 2009). The trend is the overall growth rate of an attribute and state represents the range or mean value of an attribute. These descriptions reduce the dimensionality of large time point based data to a small interval based representations, thereby simplifying the knowledge discovery process (Orphanou et al. 2014; Batal et al. 2009). The concept of temporal abstraction and time series representation in clinical domain has been investigated in many research studies (Keravnou & Shahar 2005; Adlassnig et al. 2006; Stacey & McGregor 2006; Shahar & Musen 1993; Shahar & Musen 1996). Stacey & McGregor (2006) have presented a detailed study about the temporal abstraction methods in clinical data analysis, such as RESUME (Shahar & Musen 1993), TRENDX (Haimowitz & Kohane 1996), VIE-VENT (Miksch 1996), RASTA (Connor 2001), ECHO (Semrl 1999) and CAPSUL (Chakravarty & Shahar 2000).

1.2.4 Attribute Selection

Temporal attribute selection is the process of determining minimal attribute subset for a specific problem domain. The minimal attribute subset generated maintains the high accuracy of the original attribute set (Pradipta & ParthaGarai 2013; Komorowski et al. 1999; Jensen 2007; Dash & Liu 1997). In EHR data, an attribute represents the laboratory or physical examination undergone by the patient. Clinical time series data are considered to be highly multivariate because the patient's will undergo many common laboratory examinations before taking up disease specific laboratory examinations. Thus, it is required to identify the clinical examination that is not relevant to diagnosing a particular disease. This will reduce the complexity of the mining process. Attribute selection techniques can be categorized into three groups namely filter approach, wrapper approach and embedded approach (Saeys et al. 2007, Geurts et al. 2005; Jensen 2007). This categorization is done based

on the evaluation procedures. Figure 1.4 shows the categorization of attribute selection techniques (Jensen 2007).

In filter based approach the algorithm performs selection independently without any learning algorithm. In wrapper based approach, a learning algorithm is used for attribute selection. This method searches through the attribute subset space using the estimated accuracy from an induction algorithm as a measure of subset suitability. Although wrappers may produce better results, they are expensive to run and can break down with very large numbers of attributes (Jensen 2007). This is due to the use of learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets. The embedded approach performs the search for relevant attributes while constructing the classification model. Embedded methods are less computationally expensive than wrapper methods (Saeys et al. 2007).

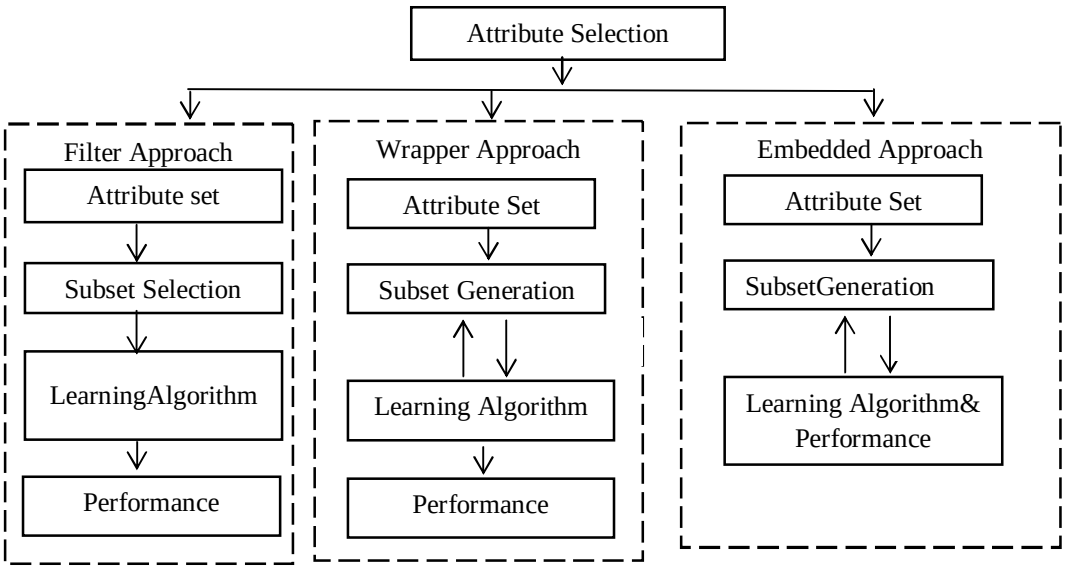


Figure 1.4 Attribute Selection Categorization

Generally, high dimensional time series data set contains many insignificant and irrelevant attributes (features). The process of eliminating

irrelevant attributes can improve the accuracy of the mining system. The relevancy of an attribute is determined by its closeness to the attributes that influence the task of decision making. Dash & Liu (1997) have presented detailed feature selection techniques for classification process.

1.3 TEMPORAL DATA MINING

Temporal data mining refers to the process of extracting useful hidden knowledge from temporal data. There are several works in literature that discusses the challenges in considering the temporal aspects of the data during temporal data mining (Moskovitch & Shahar 2015b; Batal et al. 2009). Moerchen (2006) has discussed the pros and cons of several temporal data mining techniques that exist in the literature. The temporal data mining tasks includes: clustering, classification, rule discovery.

1.3.1 Clustering

Cluster analysis is the process of grouping set of observations into subsets or groups based on their similarities. The members within the cluster show high similarity. The following are the traditional categorization of clustering methods: hierarchical, partitioning, density based and grid-based method (Han & Kamber 2001). Clustering is a challenging research area and several clustering techniques exist in the literatures.

In temporal data mining, clustering technique is widely used to discover patterns in time series data. The patterns are commonly referred as frequently appearing or surprising patterns (Fu 2011; Keogh et al. 2001). The other popularly used term for pattern discovery is motif discovery, anomaly detection or finding discords. Das et al. (1997) have used distance based clustering to discover patterns from time series.

1.3.2 Classification

Classification is the process of deriving model that describes the data and its targets. In data mining, classification is carried out in two steps namely learning step and classification step. In learning step the classification model is constructed and in the classification step the constructed model is used to predict the targeted value for the given data record. The following are the traditional classification techniques: decision tree, neural network, naive Bayesian classifier and Support Vector Machine (SVM). However, these traditional classifiers could not be used to directly classify a time series data. Few pre-processing steps are required to extract meaningful temporal patterns from the time series data. The classification model is then built using the obtained temporal patterns.

1.3.3 Rule Discovery

Rule discovery in time series data aims at extracting frequent patterns, associations and correlations based on the temporal characteristic of an attribute. In clinical domain, rule discovery approach is adopted to identify the commonly occurring patterns among the patients for diagnosing a particular disease. Association rule mining is widely used technique for identifying the frequent pattern and associations. Agrawal (1994) have presented apriori algorithm for discovering the association rules among the attributes based on candidate set generation. However, the limitations of this algorithm have been further addressed in several literatures. There are other algorithms such as frequent pattern growth and frequent item mining using vertical data format for discovering associations among the attributes (Han & Kamber 2001).

1.4 RESEARCH MOTIVATION

The emergence of temporal data has been drastically increased due to the advancement of biomedical equipment in clinical domain. This led the clinician to monitor the state of patient health periodically and to capture it as Electronic Health Records (EHR's). Clinical time series data that have been obtained from these EHR stores enormous medical knowledge which can be used to develop CDMS for assisting the physician in clinical diagnosis. Though, clinical data contain useful medical knowledge, they are also liable to temporal complexities such as irregular observations, missing values and time constrained attributes. The research work aims at using mining techniques to extract the knowledge from clinical time series data. The extracted knowledge has been used to develop CDMS that aids the physician in clinical diagnosis.

1.5 RESEARCH OUTCOME

This section outlines the research contributions and societal contribution of this research work.

1.5.1 Research Contributions and Findings

This research work presents mining frameworks that help in building temporal classification models for CDMSs. The research contributions are as follows: first, Temporal Rough Set Induced Neuro-Fuzzy (TRiNF) classifier; Second, Q-Backpropagated Time Delay Neural Network (Q-BTDNN) classifier; third, missing value imputation for temporal classification using Tolerance Rough Set induced Bio-Statistical (TRiBS) framework; fourth, Statistical Tolerance Rough Set induced Decision Tree (STRiD) classifier and fifth, Forecast-Error Approximation based Bottom-Up (FeAB) segmentation for Time Delay Neural Networks (TDNN).

1.5.2 Societal Contribution

CDMS can aid the physician to monitor and diagnose the state of patient health that is captured in EHR's. The mining frameworks proposed in this research work constructs classification models for developing the CDMSs. In this research work, CDMSs have been developed to diagnose Parkinson's, Hepatitis and Thrombosis in collagen diseases. The CDMSs developed in this work can be used by the physicians to perform the following clinical activities: to diagnose and monitor the severity of Parkinson's disease based on the gait disturbances, to diagnose Hepatitis B and C and to diagnose and monitor the severity of Thrombosis in collagen diseases. The proposed mining frameworks can be extended to develop CDMS for diagnosing other diseases with minor domain specific changes.

1.6 ORGANISATION OF THESIS

The thesis pertaining to this research work is organized into nine chapters. The first chapter provides an overview about the emergence of temporal data and temporal data mining in clinical domain. The second chapter presents the related works from literature that includes temporal data mining in clinical data analysis. Chapter three presents the system framework. Chapter four presents an overview about the datasets used in this research work. Chapter five presents the design and development of TRiNF classifier for constructing temporal classification model along with experimental results. Chapter six presents a classifier, Q-BTDNN, for analysing gait disturbances for Parkinson's disease diagnosis. An approach for imputing missing values in clinical time series data using TRiBS framework is presented in chapter seven. Chapter eight presents a framework, STRiD, for constructing temporal classification model using concept of FIDES forecasting, tolerance rough set and decision tree. Chapter nine presents FeAB segmentation for TDNN classification using the concept of forecasting in bottom-up segmentation. Conclusion and the scope for future work are discussed in chapter ten.