CHAPTER 2

LITERATURE REVIEW

This chapter presents a review on the works carried out by the researchers in the field of temporal reasoning, temporal data abstraction and temporal data classification in clinical domain. EHR's stores medical information that includes the results of laboratory and physical examinations undergone by a patient. The application of temporal reasoning, temporal data abstraction and temporal data classification in clinical data analysis has been studied in several research studies (Batal et al. 2016; Orphanou et al. 2014; Stacey & McGregor 2006).Temporal reasoning process infers the state of a system and its associated events at different time points (Augusto 2005; Adlassnig et al. 2006; Zhou 2009; Chittaro & Montanari 2000). Temporal data abstraction refers to the process of transforming low level quantitative descriptions in time series data into high level qualitative descriptions (Orphanou et al. 2014; Keravnou & Shahar 2005; Sacchi et al. 2007; Carrault et al. 2003). Temporal data classification is a task in temporal data mining, which aims at building a trained classification model for time-stamped data (Moskovitch & Shahar 2015; Fu 2011; Moerchen 2006).

2.1 TEMPORAL REASONING IN CLINICAL DOMAIN

In clinical domain, temporal reasoning refers to the task of inferring states of patient health at different time points and the events that triggers the change of these states. Combi & Shahar (1997) have presented a study about temporal reasoning and temporal data maintenance to develop time-oriented

medical systems. Investigations into these two concepts summarize the challenging research works identified in collaborating temporal reasoning and temporal data maintenance. Allen (1983) has presented thirteen temporal relations that define associations between different time intervals. These temporal relations have been adopted in many clinical research studies to perform temporal reasoning.

Augusto (2005) presented a survey that illustrates the importance of temporal reasoning in clinical decision support task. Temporal reasoning supports several clinical tasks such as prevention, diagnosis, treatment and prognosis (Augusto 2005; Zhou 2009). Prevention refers to the task of predicting the risk factors using time series analysis (Baumert et al. 2005). Diagnosis refers to the task of discovering temporal patterns that describes the behavior of diseases (Kahn et al. 1991; O'Connor 2001). Treatment refers to the task of monitoring the clinical activities associated with the therapies such as chemotherapy over a period of time (Duftschmid et al. 2002; Long 1996; Haimowitz & Kohane 1996). Prognosis refers to the task of forecasting the adverse effects of health care on the patients under medications (Schmidt & Gierl 2003; Zupan et al. 2001). The applications of temporal reasoning in clinical decision support systems aid the physician in performing the clinical tasks.

Adlassnig et al. (2006) have presented a detailed discussion on the applicability of temporal representation and reasoning to clinical tasks such as monitoring and treatment. The authors examined several concepts related to temporal databases, handling uncertainties in clinical data and reasoning on temporal clinical data for mining. Khan et al. (1991a) have presented a program named TOPAZ that uses clinical knowledge about time stamped associations to infer temporal clinical events. TOPAZ summarizes the temporal events in the patient health records that can be used in temporal reasoning tasks. To

overcome the temporal reasoning problems faced in developing a clinical expert system an extension to the time-oriented databank model based databases was presented by Khan et al. (1991b).

2.2 TEMPORAL DATA ABSTRACTION IN CLINICAL DATA ANALYSIS

Temporal abstraction is a method of representing point based time series data to interval based summarized descriptions. The process of temporal abstraction has been proven to be effective in discovering knowledge from time series (Moskovitch et al. 2015a). Temporal abstraction is considered as a preprocessing step in the temporal knowledge mining process (Moerchen 2006; Batal et al. 2009; Batal et al. 2013). Shahar (1997) describes the temporal abstraction in clinical domain as the process of interpreting clinical temporal parameters and events in the time stamped data as states and trends. The task of temporal abstraction also plays a vital role in temporal reasoning that aids the process of time series data analysis.

Orphanou et al. (2014) have presented a survey on temporal abstraction and temporal Bayesian networks to perform clinical tasks such as diagnosis and prognosis in clinical domains. Stacey & Mcgregor (2006) have presented a detailed survey on temporal abstraction based clinical data analysis. The authors have discussed several works that illustrate the development of temporal abstraction systems such as RESUME (Shahar & Musen 1996), TRENDX (Haimowitz & Kohane 1996), VIE-VENT (Miksch et al. 1996), ECHO (Semrl 1999) and RASTA (O'Connor 2001). Batal et al. (2009) have proposed a segmented time series feature (STF) mine algorithm for extracting the discriminative temporal abstraction patterns from frequent temporal patterns using qualitative temporal abstraction mechanism and Allen's temporal relations (Allen, 1983). The multivariate features obtained are used to learn a classification model.

Sacchi et al. (2007) have proposed an algorithm apriori-like, which extracts the temporal association rules based on the relationship between complex temporal patterns from time series data. The temporal patterns are obtained by applying the knowledge based temporal abstraction framework and Allen's temporal operators are used for obtaining the temporal precedence. The work was experimented on two different kinds of data, first set includes clinical data monitored during hemodialysis sessions and the second consists of DNA microarray gene-expression data for mining genetic regulatory relationships. Sacchi et al. (2007) have compared the performance of two types of temporal abstractions namely qualitative and quantitative temporal abstraction. In the former, domain specific knowledge has been used to derive a small set of symbolic representation for the temporal data whereas the later derives numeric abstraction using statistical process. The result indicated that quantitative procedure are informative and improves prediction accuracy than qualitative procedure for intensive care unit data.

Ho et al. (2003) have proposed a temporal abstraction approach for extracting knowledge from Hepatitis dataset collected from Chiba University hospital. The temporal abstraction process extracts the states and trends for each patient for a particular laboratory test within a specified episode. It was inferred that various machine-learning methods can be applied to the abstracted data in order to extract knowledge that can be used by physicians.

2.3 TEMPORAL DATA CLASSIFICATION

Classification on time series data is challenging as they possess several temporal intervals and abstracted interpretations in addition to the time stamped data points. In general, a time series data can be categorized as univariate or multivariate based on the presence of single variable or multiple variables. Moskovitch et al. (2015a) have discovered symbolic frequent Time Intervals Related Pattern (TIRP) using an algorithm named Karmalego. The

proposed algorithm includes a data structure and new method for TIRP candidate generation using the Allen's temporal relations (Allen 1983). The authors have experimentally proved the improvement in the execution speed of their Karmalego algorithm over other state of art time interval pattern mining techniques like ARMADA, IEMiner and H-DFS.

Moskovitch et al. (2015b) have proposed a framework, Karmalegosification (KLS) for classification of multivariate time series data. Temporal abstraction was carried out to transform time point series into time interval series. In order to avoid finding all of the TIRPs while mining a single entity, modified Karmalego was proposed which consists of Singlekarma and Singlelego algorithms. Various discretization strategies such as Symbolic Aggregate Approximation (SAX), Equal-Width Discretization (EWD) have been used for transforming time series data into time intervals series.

Moskovitch et al. (2015c) have presented a supervised discretization technique to improve the accuracy of the classification process. The authors have described the procedures for determining optimal-cut offs in continuous data inorder to create discrete symbols. They have derived symbolic time intervals and frequent Temporal Interval Relation Pattern (TIRP) using the methods discussed in their previous work (Moskovitch et al. 2015a). The extracted patterns are then used to induce a classifier as discussed in Moskovitch et al. (2015b). The authors have compared the presented Temporal Discretization for Classification (TD4C) with unsupervised discretization namely EWD, Knowledge base and SAX. The performance of TD4C was found effective compared to the unsupervised methods.

Batal et al. (2013) have presented a temporal pattern mining technique named Minimal Predictive Temporal Patterns (MPTP), for performing classification of medical health records. MPTP algorithm combines

pattern selection and frequent pattern mining. The health records of Heparin induced Thrombocytopenia (HPT) patients, were used for experimentation. This MPTP framework has extracted useful features for classification, which was used in developing a clinical decision making system. Moerchen (2006) has presented an effective unsupervised algorithm to perform mining from the temporal concepts extracted by temporal language Time Series Knowledge Representation (TSKR) based on sequential pattern and itemset mining. The usage of TSKR in mining has overcome the limitations of Allen interval relations and it is demonstrated using a sport medicine dataset (Allen 1983).

Bodyansky et al. (2005) have presented a neuro-fuzzy network using the Kolmogorov's superposition theorem named Neuro-Fuzzy Kolmogorov's Network (NFKN). The Least Square Method (LSM) was used to train the output layer in NFKN and the gradient descent method was used to train the hidden layer. NFKN is highly suited for classification since it effectively handles the dimensionality complexity using a two level structure based on KST. However, the training process requires improvement in the convergence behavior and in extending the classification process to support multiple class labels. Petkovic et al. (2013) have used an Adaptive Neuro-Fuzzy Inference System (ANFIS) network presented by Jang (1996) to study the impact of Autonomic Nervous System (ANS) on the significant Heart Rate Variability (HRV) parameters. For analysis, they have extracted 14 parameters of HRV signal. They have done a detailed investigation to identify the HRV parameters that are affected by the ANS functions. Two ECG datasets were used for the analysis, namely Arrhythmia Database and epilepsy database. A comparative analysis between the ANFIS prediction method and linear regression model with respect to its regression error shows that the performance of ANFIS model is improved over the linear regression model.

Mcnameea et al. (2005) presented a Neuro-Fuzzy Inference System (NFIS) to simulate heart rate variations. They have developed a system to predict the changes in health conditions for patients in the Neurological Intensive Care Unit (NICU). They have demonstrated the NFIS model with both observed and simulated data from NICU patients. The experimental results indicate that the NFIS is capable of effectively predicting the changes in heart rate. Khanna et al. (2007) have proposed a clinical decision making process using four different mining techniques, namely association rule mining, decision tree, neural network and neuro-fuzzy along with the temporal constraints. These approaches extract temporal rules, validate it and store the rules in the knowledge base. For experimentation the authors have used time series Hepatitis and Thrombosis datasets (Hepatitis, 2005; Thrombosis, 1999).

Kamali et al. (2014) have proposed an approach for quantitative analysis of ElectroMyoGraphy (EMG) signals by decomposing it into Motor Unit Action Potential Trains (MUAPTs). The authors have presented supervised classification techniques that combine both time and time-frequency features of MUAPT to determine its class label. The time domain features include rise time, duration, spike duration, peak-to-peak amplitude, area, turns; and the frequency domain features include the characteristics extracted from the discrete wavelet transform. Single classifier scheme as well as multi-classifier techniques that use ensemble of Support Vector Machines (SVMs) as base classifiers are explored in their work. The subsets of features employed for each base classifier are optimized through wrapper method. The results show that time-frequency domain features are superior to those of time domain features.

Minas et al. (2010) have proposed a data mining system for the evaluation of risk factors related to heart. The system uses classification algorithm C4.5 but instead of using single splitting criteria, five different

splitting criteria's namely information gain, gini index, likelihood ratio chi-squared statistics, gain ratio, and distance measure were investigated to segregate the attributes. A heuristics process was used for pruning based on the statistical significance of splits. For classifying Myocardial Infarction (MI), Percutaneous Coronary Intervention (PCI), and Coronary Artery Bypass Graft Surgery (CABG) patients, three classification models were developed based on decision trees.

Xinmeng et al. (2012) proposed a new method to build decision trees by using maximum similarity of attributes as the splitting criteria. The splitting criterion was determined based on maximum similarity for constructing the decision tree. The algorithm also does pruning, which was not done in ID3 algorithm making it more accurate than ID3. Bellazzi & Zupan (2008) have presented a detailed review about the usage and challenges of predictive data mining in the medical domain. A detailed study about the merits and demerits of various classification methods discussed provides the guidelines required for carrying research studies in clinical data mining.

Liu et al. (2014) have presented a new hierarchical system framework that builds a temporal model for irregularly sampled time series data to support clinical decision making. The authors have presented algorithms to learn temporal models from the data. Moreover, these models accurately predict future values. The authors in their framework have used machine learning and data mining algorithms such as Linear Dynamical System (LDS) and Gaussian Process (GP) (Kalman 1963; Rasmussen et al. 2006). GP models irregular time series data that accurately predict future values (Rasmussen et al. 2006). GP makes observations as a function of time and there is no need to mark when the observations were made and whether they are regularly or irregularly spaced. LDS (Kalman 1963) defines a state-space process with linear transitions between two consecutive states taken at discrete time points.

However, in most of the real world applications time series is not discrete. Hence, GP was used at lower levels over time windows for modeling irregular time series data. LDS then tracks the transition in the GP process.

To perform analysis of an irregularly sampled data the following two methods were used: direct value interpolation (Pandit 1983; Adorf 1995; Dezhbaksh & Levy 1994; Kreindler 2006; Rehfeld et al. 2011) and windows-based segmentation (Chu1995; Keogh et al. 2001). The former assumes that all values are collected regularly with a pre-specified sampling frequency and converts time series with irregular observations to discrete time observation sequences. The later first segments time series to fixed-sized windows. From this summary statistic was calculated. Like LDS, Autoregressive model (AR) is a discrete time series model used to represent a stochastic process. Prediction was carried out by taking an initial sequence using AR or LDS model. The correctness of the system was proved using mean absolute prediction error and absolute percentage error. The framework was used with a univariate time series data. Complete Blood Count (CBC) laboratory time series data was used for experimentation. The results has proven an improvement in the prediction accuracy compared to the best performing baseline (AR, LDS, GP) and other window based segmentation method. The authors have concluded that their work has a limitation that it works only with univariate time series data and it was extended to support multivariate time series data.

Bahadori et al. (2012) have presented a Generalized Lasso Granger (GLG) method that discovers the temporal dependencies from irregular time series data. The authors also have presented a review on various techniques used in analyzing irregular time series. The general methods available for analyzing irregular time series are namely the repair approach, Lomb-Scargle Periodogram (LSP), wavelets and Kernel methods (Kreindler 2006; Rehfeld et al. 2011; Cuevas-tello et al. 2009; Scargle 1981). GLG uses kernel functions

to simplify the inner product for irregular time series. The authors have presented a theoretical analysis and simulated experiments with four synthetic datasets to prove the effectiveness of their proposed work. An application of the GLG method with the datasets of δ18O (a radio isotope of Oxygen) is provided to detect the moisture transfer patterns. GLG is likely to have lower absolute errors because it predicts the actual observations without additional repair error. GLG becomes more accurate when there is a decrease in probability of missing a data. However, the authors have concluded that GLG approach has limitations with respect to the scalability in data analysis.

There has been many works in the literature that addresses the task of mining in time series data. However, these methodologies have restrictions to work with multivariate time series data observed at irregular intervals because they are either tuned to support regular time series data or irregular univariate time series data. Clinical observations are often irregular and multivariate. Hence, mining in such clinical time series data is a challenging area of research.

2.4 RESEARCH DIRECTIONS

Comparing to the works discussed in literature, the proposed research contributions are different in following ways:

The first research contribution presents a Temporal Rough Set induced Neuro-Fuzzy (TRiNF) mining framework for classifying unevenly spaced clinical time series data. In this work, the temporal complexities that occur in clinical time series data have been pre-processed using an enhanced DES forecasting method presented by Wright (1986). The forecasted results are used to impute the missing values and to derive the temporal patterns. The temporal patterns obtained for each clinical attribute are used in the attribute selection and classification process which differs from traditional

methodologies that use actual observed value. A temporal rough set induced attribute selection process is presented to identify the relevant attributes. Temporal trend pattern obtained for each clinical attribute has been used to fuzzify the inputs for temporal pattern induced neuro-fuzzy classifier. The work was experimented with Hepatitis and Thrombosis time series dataset.

The second research contribution presents a Q-Back Propagated Time-Delay Neural Network (Q-BTDNN) classifier that builds a temporal classification model. This classification model has been used in developing CDMS for diagnosing the severity of gait disturbances in PD. A reinforced Q-learning back-propagation (Q-BP) algorithm has been presented to train the TDNN in an incremental way. During the training process, the network weights are adjusted based on the reinforced back-propagated error signal. The temporal ordering among the observed gait patterns of each subject (person) are considered in diagnosing the severity conditions of the gait disturbances in PD. The experimental result proves the efficiency of Q-BP in terms of its improved classification accuracy.

The third research contribution focused on imputing the missing values that occur in unevenly spaced clinical time series data. This work has proposed an enhancement to the Inverse Distance Weight (IDW) interpolation by using the concept of tolerance rough set analysis and PSO for missing value imputation. The IDW has two major limitations in choosing the number of known data points (nearest neighbourhood) and finding the optimal value for influence factor parameter. The proposed work overcomes these limitations by using the tolerance rough set analysis concept for selecting the relevant known points and PSO techniques to find the optimal value for influence factor parameter. From the experimental results it has been inferred that the proposed framework has effectively reduced the error rate and

improved the accuracy of the imputed results. The work was experimented with Hepatitis and Thrombosis time series dataset.

The fourth research contribution presents a Statistical Tolerance Rough Set induced Decision tree (STRiD) mining framework for developing CDMS that performs classification on clinical time series data. This work has enhanced the DES method by incorporating fuzzy inference system in smoothing constant estimations. A forecasting model was constructed using the proposed Fuzzy Inference Double Exponential Smoothing (FIDES) method. The obtained forecasted model is used for missing value imputation and temporal pattern derivation. The FIDES chooses optimal smoothing constant value for trend and estimation based on interval spacing's among each observation. Attribute selection and classification process is performed based on the attributes temporal pattern and not on the actual observed value. In attribute selection, the significant attribute set has been formed using a temporal pattern based tolerance rough set approach which differs from the first contribution in the way of forming tolerance class instead of equivalence class. The classification model is built using a decision tree classifier that uses a temporal pattern induced gain ratio as splitting criteria. The work was experimented with Hepatitis and Thrombosis time series dataset.

The fifth contribution presents a Forecast-Error Approximation based Bottom-Up (FeAB) segmentation approach that effectively reduces the dimensionality (length) complexity of clinical time series data. The novelty of the work lies in two aspects. First, in effectively using the forecasted results obtained from Hanzak updated DES in bottom-up segmentation and second, incorporating the temporal summarized segments with Time Delay Neural Network (TDNN) to classify unevenly spaced data. Although, classifiers like TDNN have dynamic response to time series data, the challenge arises due to the irregularities in clinical data. The work consists of two stages, namely

temporal summarization and classification. In temporal data summarization, clinical time series data is divided into sequence of temporal interpreted segments using the proposed FeAB segmentation. FeAB adopts a Double Exponential Smoothing (DES) technique to derive the growth rate, mean and forecast-error for each clinical observation. The obtained forecast-error has been used to compute the merge-cost for FeAB segmentation. TDNN has been used in the classification process to build a classification model for the segmented time series. The work was experimented with Hepatitis and Thrombosis time series dataset.