CHAPTER 7

MISSING VALUE IMPUTATION IN UNEVENLY SPACED CLINICAL TIME SERIES DATA

Clinical trials generate time-stamped data that record set of observations on state of patient's health. These data are liable to missing values since there are situations, where the patient observations are neither done regularly nor updated correctly. Missing value imputation in clinical time series data becomes challenging when the observations are done at irregular intervals. The objective of this contribution is to impute missing values in an unevenly spaced clinical time series data using a Tolerance Rough Set induced Bio-Statistical (TRiBS) framework. The proposed framework adopts an Inverse Distance Weight (IDW) interpolation technique and improves it using the concept of tolerance rough set and Particle Swarm Optimization (PSO). The classical IDW interpolation suffers from two major drawbacks while interpolating an unknown data point: first, in selecting the known data points and second, choosing an optimal influence factor. TRiBS framework overcomes the first limitation of IDW using tolerance rough set and the second using PSO.

7.1 MISSING VALUE IMPUTATION TECHNIQUES: BACKGROUND

This section provides a detailed description about the statistical and machine learning techniques used in imputing the missing values.

7.1.1 Statistical Techniques

Ford (1983) has presented a hot-deck imputation method in which the available complete records act as donors for the records that contain missing values. This method attempts to impute missing values from observed values with similar pattern hence it is also termed as similar response pattern imputation. Andridge & Little (2010) have provided a comprehensive review about the various versions of hot deck imputations and its applications. Perez et al. (2002) in their work have illustrated the usage of various imputation techniques like mean, hot deck and multiple imputation for predicting the outcome in the intensive care units.

Van der Heijden et al. (2006) presented a mean imputation method that can be conditional or unconditional. In unconditional imputation, the overall mean of the attribute corresponding to the missing value from the observed data set is used to impute the missing values. Conditional mean replaces the missing values with the mean of the specific subgroup to which it belongs. If the attributes are categorical then the missing data are replaced by its modal value. Sullivan & Andridge (2015) have developed a non-ignorable proxy pattern mixture hot deck multiple imputation method which suits all types of missingness. The authors have combined the ideas of hot deck imputation using distance-based donor selection and a parametric non-ignorable imputation procedure which are based on the assumption of Missing Not At Random (MNAR) category. A donor quality metric named minimum mean distance has been proposed and a sensitivity parameter was used to identify the missingness mechanism.

Inverse Distance Weight (IDW) interpolation presented by Shepherd and Donald (1968) finds the missing attribute value using the weighted average values from known data points. IDW suffers from many drawbacks which include the distance calculation, selection of neighbourhood

and its weight factor. To overcome these limitations several enhancements related to the distance computation and influence factor optimization were incorporated to the traditional IDW (George et al. 2008; Cressman 1959; Gandin 1970; Barnes 1964; Sen & Sahin 2001).

Little et al. (1987) have presented multiple imputation method for handling missing data. The process involves three steps, first the missing values are imputed 'm' times in order to represent the uncertainty about which value to impute, second step involves the analyses of the 'm' complete data sets and as the third step combines the results to yield a single combined estimates, e.g., p-values, regression coefficients, standard errors that formally incorporate missing data uncertainty.

Van der Heijden et al. (2006) have presented a work that illustrates the effectiveness and impact of handling missing data in clinical diagnostic studies. For handling the missing data the authors have considered the following methods namely complete case analysis, missing-indicator method, multiple imputation, single imputation of unconditional and conditional mean. For experimentation the authors have considered diagnostic patient data for pulmonary embolism. The experimental results shows, that imputing missing data improves the prediction results compared to the complete case analyses and missing-indicator method. It has been observed that multiple imputations attain effective results but when there is low number of missingness the single imputation method works effectively.

Little & Rubin (1987) used regression imputation to construct a regression model with the complete observations. This model is used to predict the value of the missing data. Expectation Maximization (EM) methodis a two-step iterative process proposed by Dempster et al. (1977) in which a complete data set is obtained by imputing the missing values by performing the E-steps (expectation) and the M-steps (maximization)

repeatedly until convergence occurs. Initially the E-step computes the expected value of the sum of missing data variables with an assumption that the value for the population mean, and variance-covariance matrix are known. The expected value of the sum of a variable is used in the M-step to estimate the population mean and covariance. The process iterates until the values of the estimates do not change. The major advantage of EM method over mean and hot deck imputation is that it preserves the relationship between the variables whereas mean and hot-deck imputations reduce the variance and the absolute value of the covariance.

Full Information Maximum Likelihood (FIML) is a model based approach in which parameter estimation are handled in a single step (Enders & Bandalos 2001). It is also referred as raw-data maximum likelihood, which reads in the raw data one case at a time, and maximizes the maximum likelihood function one case at a time; finally the results are combined to produce an overall estimate of the maximum likelihood function.

Olinsky et al. (2003) presented a comparative study that illustrates the efficacy of mean imputation, regression imputation, multiple imputation, EM, maximum likelihood and FIML imputation for missing data in structural equation modelling. The results indicated FIML to be far superior to other methods and could be used for the determination of parameters in structural equation modelling, however if a complete set of imputed data is required then maximum likelihood is found to be superior.

Strike et al. (2001) have evaluated the performance of mean imputation, hot deck imputation and listwise deletion with respect to bias and precision for software cost modelling. The results show that listwise deletion is suitable when there is less percentage of missingness. However the precision shows variation, when the missingness increases or it is non-ignorable. In the case of mean imputation the precision tends to be higher and

bias is nearly zero for missing at random (MAR) and missing completely at random (MCAR) type of missingness but accuracy decreases for non-ignorable missing data. Hot deck imputation with z-score standardization and Euclidean distance outperforms list-wise deletion and mean imputation as it shows less bias and higher precision even for non-ignorable missingness.

7.1.2 Machine Learning Techniques

To impute missing values using machine learning techniques, a data model is built from the complete data available for each of the attribute and later the constructed model is used to predict the missing values. The methodology remains the same for several supervised machine learning algorithms such as decision trees, probabilistic, and association rules (Farhangfar et al. 2007). K-Nearest Neighbour method imputes missing values from the computed nearest neighbour of suitable distance. Artificial neural networks (ANN) are capable of extracting information and patterns from the data. ANN's are thus flexible in modeling many types of nonlinear relationships and finds its application in imputing missing values (Nordbotten 1966; Silva-Ramírez et al. 2011; Junninen et al. 2004; Gheyas & Smith 2010)

Nordbotten(1966) have presented a work that uses a ANN to impute missing values. The training model is built using non-missing records and the trained model is used for imputing missing records. Silva-Ramírez et al. (2011) developed a multilayer perceptron method of imputation for handling data which are missing completely at random. The performance of this method was almost similar to other methods like hot deck, KNN, mean/mode and regression for quantitative variables but it outperformed the others for categorical variables. Junninen et al. (2004) have presented an evaluation study on missing data imputation using air quality datasets. Missing patterns were simulated and the dataset was evaluated with regression-based imputation, linear, spline and nearest neighbour

interpolation, multivariate nearest neighbour, Self-Organizing Map (SOM), Multi-Layer Perceptron (MLP), and hybrid methods. The comparison results illustrate that SOM and MLP outperforms the multivariate nearest neighbour and other techniques. Gheyas & Smith et al. (2010) have proposed two missing value imputation algorithms based on an ensemble model of Generalized Regression Neural Networks (GRNN) namely GRNN-ensemble for multiple-imputation and GRNN ensemble for single imputation. The key advantages of these algorithms lies in the fact that they are local approximators and are non-parametric algorithms which avoid assumptions made on distributions.

Decision tree method treats the missing variable as the target and the remaining variables as predictors. Decision tree imputation employs learning algorithm such as ID3 to build a decision-tree classifier using the rows which are complete. The decision tree rule is then applied on the row with missing value to predict the missing value. Rahman & Islam (2013) have presented two techniques for imputing missing values using decision trees and decision forests. These techniques forms segments in the dataset which contains records with high similarities and attribute correlations. Missing values are then imputed based on the similarity between the segments and missing record. The precision of the imputed results are increased because the authors have considered all the attributes based on their correlation strength.

Bayesian Principal Component Analysis (BPCA) is a probabilistic model based on Bayesian principle component analysis for estimating the missing values (Oba et al. 2003). This method consists of three processes: principal component regression, Bayesian estimation and EM algorithm. The method divides the data set into complete and incomplete set. Bayes theorem is used to calculate the PCA and the Bayesian estimation calculates posterior distribution of model parameter and input matrix containing samples. The iterative EM method is employed to compute the unknown parameter.

Cismondi et al. (2013) employs a fuzzy logic based classification model which identifies whether the missing values are recoverable (independent) or non recoverable (independent) based on the dependency of the missing values with other variables. According to the classification result recoverable values are imputed using mean imputation and non recoverable values are deleted. Ding & Ross (2012) proposed a gaussian mixture model based K- nearest neighbour method which outperformed methods based on likelihood, regression, bayesian and multiple imputation for handling missing scores in biometric fusion. Qu et al. (2009) have presented an approach that integrates principal component analysis and maximum likelihood estimation to form a probabilistic Principal Component Analysis (PCA) for imputing missing values in time series data. PCA is used for extracting dominant parts and the maximum likelihood estimation is used for imputing the missing values using the likelihood function derived from the sampled data.

An analysis of various missing data imputation methods based on both statistical and machine learning methods was carried out by Jerez et al. (2010). The adopted statistical methods are: hot deck methods, multiple imputation and mean methods. The adopted machine learning methods are: multilayer perceptron, SOM, K-nearest neighbor. The results concluded that the machine learning methods outperformed statistical methods with significant improvement in prediction accuracy.

Most of the imputation techniques discussed in the literature can work effectively with the time series data that are regular. Since clinical time series data is considered to be unevenly spaced, the direct application of these techniques may degrade their performance. Hence, this work proposes an enhancement to the IDW interpolation by using the concept of TR analysis and PSO for missing value imputation. The traditional 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 TR analysis concept for selecting the relevant known points and PSO techniques to find the optimal value for influence factor parameter.

7.2 TRiBS FRAMEWORK

The proposed framework is shown in the Figure 7.1. The components of the proposed framework are significant data selection, influence factor optimization and imputation

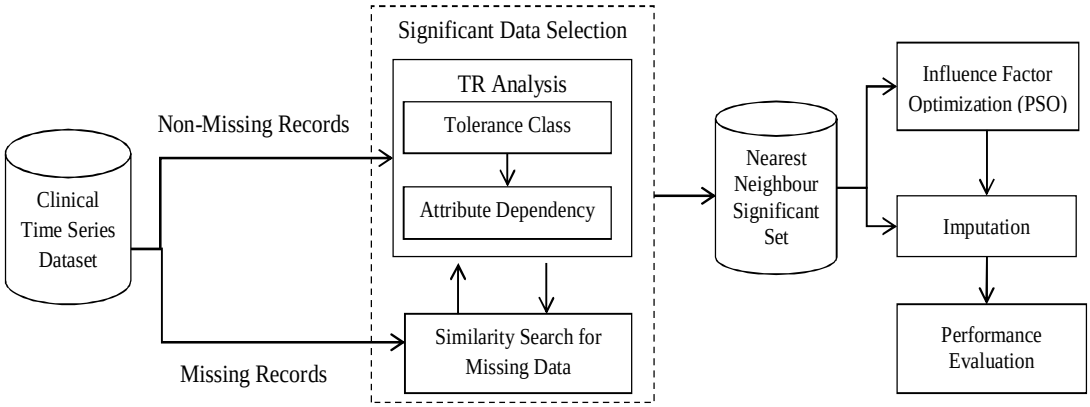

Figure 7.1 TRiBS Framework

Imputing missing value in the clinical data becomes challenging due to the uneven spacing's in the observed data.

7.2.1 Significant Data Selection (SDS)

The classical IDW (Shepard 1968) method finds the value for any unknown point using the measured known data significant to it, which forms the neighbourhood set. The size of the neighbourhood set chosen, has a direct influence in the accuracy of the determined value. The IDW has limitation in

choosing the number of known values for interpolation. The proposed work overcomes this limitation using significant data selection process, which aims in identifying the nearest significant known data points. This process is done in two steps: First, attribute dependent set generation and Second, similarity search. The concept of tolerance rough set analysis is used in significant data selection process.

The following notations have been used in the discussion and equations. Let I= (𝕌, A) be an Information system, $\mathbb{U} = \{x_1 . . . x_i . . . x_j . , x_n\}$ is called as an universe, x_i, x_j be the i^{th} and j^{th} objects in the universe '𝕌', *A* is the Knowledge in a universe which is the non-empty finite set of attributes, a be the attribute, $a \in A, a(x_i), a(x_j)$ is the value of x_i, x_j for the attribute a, τ is the tolerance value (0 to 1) , $a(t_j)$ and $a(t_i)$ is the observation period of x_i, x_j for the attribute a, a_{max} and a_{min} is the maximum and minimum value of a, $a_{t_{max}}, a_{t_{min}}$ is the maximum and minimum duration, $a<(x_i(t_i), x_j(t_j)), \tau >$ value of x_i, x_j for the attribute at time t_i and t_j, B and Q are two attributes such that $B \epsilon A, Q \epsilon B$; $\bar{a}$ represents the mean.

Generally, an information system in rough sets is represented as *I=(𝕌,A)*, where *𝕌={x_1, . . . , x_i, . . x_j . , x_n}* is called as an universe, which is a nonempty set of finite objects and *A* is the knowledge in an universe which is the non-empty finite set of attributes (Pawlak 1982). The equivalence classes of the B-indiscernibility relation (*IND(B)*) is also denoted as $[x]_B$ and $X \subseteq \mathbb{U}$ represents an elementary portion of knowledge that can be extracted. This concept of equivalence class becomes complex, when the attributes hold real value data. So to address the real value data problem, the concept of tolerance rough set has been chosen (Komorowski et al. 1995).

The tolerance rough set forms tolerance classes based on the similarity in attributes. Lower and upper approximations are the two main

operations that are used in characterizing the knowledge. Lower approximation of set contains all elements that surely belong to the set. Upper approximation of set contains all elements that possibly belong to set. Based on these approximations three regions are characterized namely positive, negative and boundary region. Positive region contains all the objects of '𝕌' that can be classified to equivalence classes of '𝕌/B' where $B \subseteq A$. Negative region contains all the objects, which is certainly non-member of X. Boundary region contains all objects which is possibly member of X. The degree of dependency of the attributes is an important factor to be considered in the attribute selection process (Pawlak 1982; Komorowski et al. 1995). The degree of dependency between two attribute sets is defined using the positive region.

In the first step of significant data selection process, dependent attributes are identified for each clinical attributes. A temporal similarity measure is used to identify the similarity for each attribute based on its observation time. The similarity for each attribute is computed based on its observation time using a temporal tolerance similarity measure. Temporal tolerance similarity measure, $SIM_{a<(x_i(t_i), x_j(t_j)),\tau>}$ for the attribute a and its i^{th} and j^{th} object observed at time t_i and t_j is defined in the Equation (7.1).This temporal based similarity measures are then used to compute temporal tolerance relation, lower approximations to construct positive regions defined in the Equation (7.2), (7.3) and (7.4). Finally, the significance of the attribute is computed using temporal tolerance based degree of dependency defined in the Equation (7.5).

$$SIM_{a<(x_i(t_i), x_j(t_j)),\tau>} = 1 - \bar{a}\left\{\left(\frac{a(x_i)-a(x_j)}{a_{\max}-a_{\min}}\right), \left(\frac{a(t_j)-a(t_i)}{a_{t_{\max}}-a_{t_{\min}}}\right)\right\} \quad (7.1)$$

$$tempTOL_{\tau}(B)=\{ (x_i , x_j) \in U \mid \forall a \in B, (x_i , x_j) \in SIM_{a<(x_i(t_i), x_j(t_j)),\tau>}\} \quad (7.2)$$

$$\underline{B_\tau}X = \{x|SIM_{B,\tau}(x) \subseteq X\} \quad (7.3)$$

$$POS_{B,\tau}(Q) = \cup_{x\in U/Q} \underline{B_\tau}X \quad (7.4)$$

$$K = \gamma_{B,\tau}(Q) = |POS_{B,\tau}(Q)|/|U| \quad (7.5)$$

An attribute a in A where a$\in A$ is said to be dependent on other attributes a_i in A, $a_i \in A$ when its tolerance based degree of dependency value meets the dependency threshold ($D\tau$) . The value of $D\tau$ is selected by the domain based expert guidance. The complete data is grouped into two dataset categories namely missing records and non-missing records. TR analysis generates tolerance classes with the non-missing records. Lower approximation, positive region and degree of dependency are computed for each clinical attribute. Based on the degree of dependency the attribute dependent set is generated for each clinical attribute. The identified set represents the dependencies among the clinical attributes. In the second step, a similarity search process starts for each missing record. The similarity search identifies the number of significant known points. The identified significant points are considered for interpolation. For each missing record, its dependent attribute set is extracted. The similar records are grouped based on the tolerance classes for the non-missing records of the identified dependent attributes.

Algorithm 1: Significant_data_selection (Y,W, Z, τ, A, Ma, Dτ)

Input: Y- Complete dataset, W- non-missing records, Z- missing records, τ-tolerance factor, A – complete attribute set, Ma- missing attribute set, Dτ -dependency threshold

Output: Attribute Dependent set AD, Nearest significant set (NS).

1. Attribute_dependency (Y, τ, A, Ma, Dτ)

2. Similarity_ search (Z, τ, AD_i, Ma)

Algorithm 1.a: Attribute_dependency (Y, W, τ, A, Dτ)

1. m=size (A)
2. For i=1 to m do
3. D= A_i
4. R={ }
5. For j=i+1 to m do
6. $K = \gamma_{A_j,\tau}(A_i)$ // compute temporal degree of dependency using equation (7.5)
7. If k> Dτ then
8. $R = R \cup A_i$// generate dependent attribute set of A_i
9. end if
10. End for
11. AD_i = R
12. End for
13. return AD

Algorithm 1.b: Similarity_ search (W, Z, τ, AD, A)

1. n= size(A)
2. For i= 1 to n
3. H = dependent attribute set (AD) for A_i
4. T= all non-missing records of H in X
5. TC=computed tolerance classes ($tempTOL_\tau(H)$) for T using equation (7.2)
6. NS_i = Maximum set in TC

7. End for

8. Return NS

The similarity search process of SDS selects the significant data points that forms neighbourhood set for interpolation process. Figure 7.2 illustrates the similarity search process in SDS. Let AD be the attribute dependency set that was found in SDS for the attributes in A. The process of finding the nearest significant set for interpolation is described in the Figure 7.2.

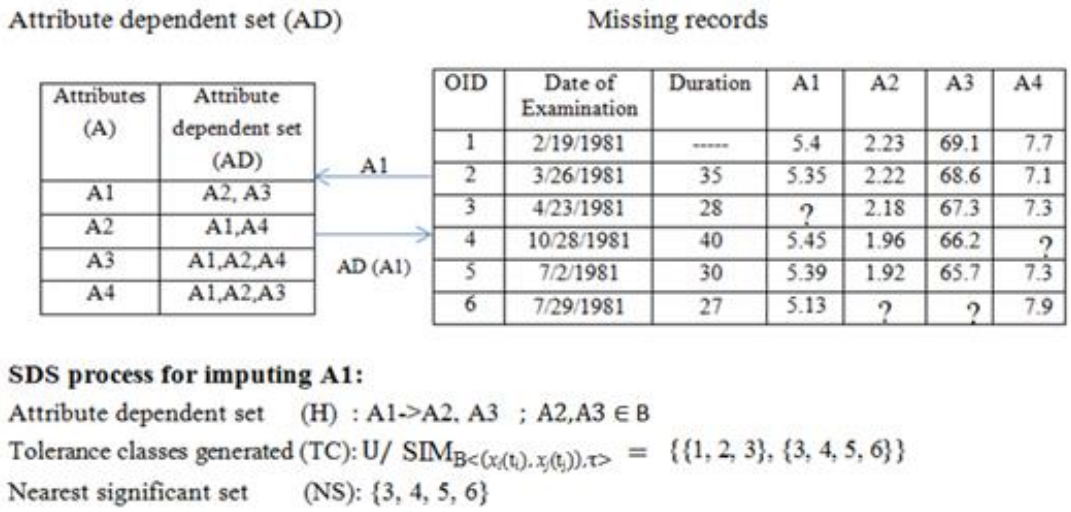

Figure 7.2 Illustration of Similarity Search Process in Significant Data Selection Process

The observation for the attribute A1 with OID (Object ID) 3, taken on 23/4/1981 is found missing. The attributes dependent for this missing attribute A1 is obtained from the attribute dependency set. Here, in this example {A2, A3} is dependent attributes for A1. The tolerance class from the non-missing records for the attributes {A2, A3} are generated. The maximum set which includes the missing OID 3 is identified as significant set for imputing its attribute A1 using interpolation.

7.2.2 Influence Factor Optimization and Imputation

The traditional IDW (Shepard 1968) works with an assumption that the distance between the measured values and the prediction location has direct influence on the predicted value. The general mathematical formulation for IDW is given in the Equation (7.6) to estimate the value at unknown point x,

$$M_E(x) = \frac{\sum_{i=1}^{n} M_A(i)/d(x,i)^k}{\sum_{i=1}^{n} 1/d(x,i)^k} \quad (7.6)$$

where $M_E(x)$ is the estimated value at point x, $M_A(i)$ is the actual value at point i, n is the total number of known neighbour points for point i, d(x,i) is the distance between the point x and the point i, and k is the influence factor parameter. This parameter value controls the weightage for each considered known points in interpolation. However, choosing this value needs to be done carefully since the wrong choice of this value affects the accuracy of the interpolation results. The Influence factor optimization (IFO) process uses PSO technique to overcome the limitation of choosing optimized value for the influence factor (k). The chosen value for 'k' fixes the weights for each considered known data points in the significant set formed from the SDS process. The optimal value minimizes the error rate and improves the accuracy of the imputed results. The concept of PSO was introduced by Kennedy & Eberhart (1995), which is a robust evolutionary computation technique based on the swarm intelligence. In PSO, particles represent solutions in a search space. A fitness function is used to evaluate the particles, which is represented by the objective function that has to be optimized. Each particle is guided by a measure called velocity to follow the best particles to search optimum point in problem space. The initial particles position, velocity and its updated position is defined using the Equations (7.7), (7.8) and (7.9).

$$P_i = range\_min + (range\_max - range\_min).* rand(numInd, n\_var) \quad (7.7)$$

$$V_i = V_i + C1 * r1 * (pbest\,i - P_i) + C2 * r2 * (gbest\,i - P_i) \quad (7.8)$$

$$P_i = P_i + V_i \quad (7.9)$$

where rand() refers to the function that generates random numbers, $range\_min$ and $range\_max$ is the minimum and maximum range of the search space, $pbest_i$ is the local best position of particle i , P_i is the current position of particle iand $gbest_i$ is the global best position of particle i, V_i is the velocity of the i^{th} particle, C1 is the cognitive coefficeint and C2 is the social component. r1, r2 are the stochastic influence component.

The fitness function of the proposed TRiBS is shown in the Equation (7.10).

$$\text{Minimize } f_{RMSE_i} = \sqrt{\frac{(M_A(i) - M_E(i))^2}{n}}. \quad (7.10)$$

The algorithm Influence_Factor_Optimization given below describes the overall steps involved in IFO process.

Algorithm 2: Influence_Factor_Optimization (n, l)

Input : n be the number of particles, p be the number of iterations

Output: 'k'is the Influence factor value

1. For each particle i =1 to n do
2. Initialize particle position (P_i) using the equation (7.7) // represents the values for influence factor (k).
3. End for
4. $pbest_i = P_i$ // Initialize the pbest for each particle as its initial value.

5. do
6. For each particle i =1 to n do
7. Compute the fitness value f_{RMSE_i} using the Equation (7.10)
8. If the fitness value $f_{RMSE_i} > \forall$ pbest $_i$ then
9. pbest $_i = P_i$ // set current particles position as local best
10. end if
11. If the fitness value f_{RMSE_i} > gbest

    gbest= P_i ;gbestval= f_{RMSE_i} // select gbest(the particle with the least fitness value)
12. end if
13. update the particle velocity using the equation (7.8)
14. update the particle position using the equation (7.9)
15. End for
16. while (maximum iteration or minimum error criteria)

7.3 EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed framework has been experimented with two clinical time series data of Hepatitis and Thrombosis patients. A detailed explanation about the experimental settings and the results obtained during the imputation process using the proposed framework is discussed in this section. The proposed framework performs two major functionalities namely significant data selection and influence factor optimization. The Tolerance Rough Set (TR) concept is used in significant data selection process to select the relevant known data points which forms the nearest significant set. The initial step of TR starts with finding the dependent attribute set for every attribute using the

non-missing records of Hepatitis and Thrombosis dataset. Since the data considered for experimentation is a time series data, the TR procedure computes the tolerance class, lower approximation, positive region and degree of dependency with respect to the temporal duration among the clinical observations.

The clinical attribute that will be considered for Hepatitis and Thrombosis diagnosis is taken along the X-axis. The corresponding number of dependent attributes is taken along the Y-axis. To identify and group the dependent attributes a tolerance threshold level of 0.48 is considered with expert guidance. An attribute is said to be dependent on the other only when its tolerance degree of dependency is less than or equal to 0.48. This dependency attribute set acts like a dictionary in the imputation process. To impute a missing data first its dependent attributes are identified and a similarity search is performed for the non-missing records in the data for those dependent attributes. This is done by generating interval-based tolerance class. The maximum set in the tolerance class, which includes the missing record, is considered as nearest significant set. The Figure 7.3 and 7.4 summarizes the number of dependent attributes identified for each clinical attributes in Hepatitis and Thrombosis patients.

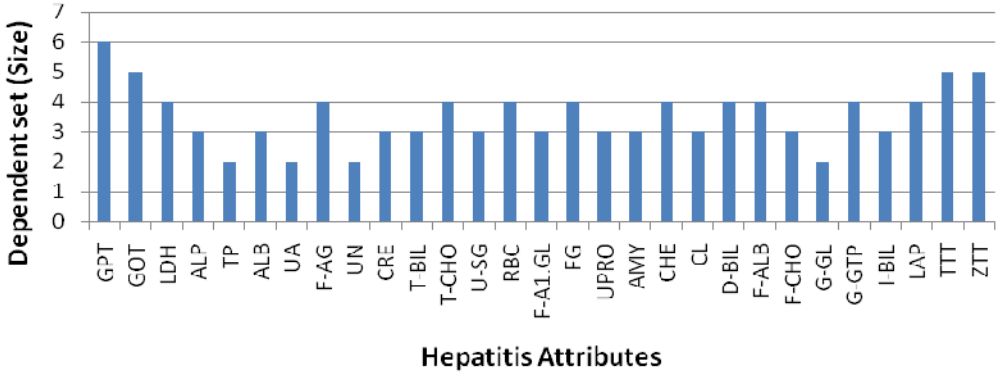

Figure 7.3 Dependency Set for Hepatitis attributes

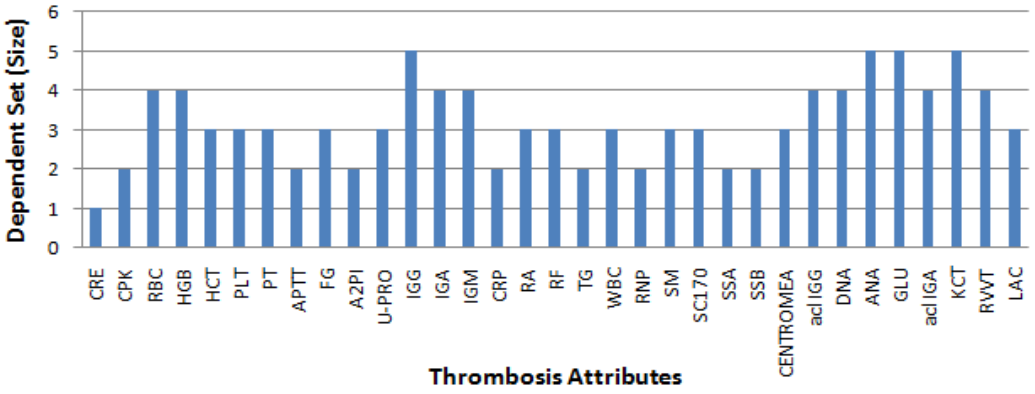

Figure 7.4 Dependency Set for Thrombosis attributes

The characteristic of the missing attributes namely its temporal interval and its degree of dependency with other attributes has higher impact in choosing the nearest significant set boundaries. The missing attribute that are highly dependent on the other attributes and has lower temporal interval compared to that of average interval spacing among the clinical observations has high nearest significant set. The PSO technique now fixes weights for each element in nearest significant set by finding the optimal value for the influence factor parameter. The search space to form the solution set is fixed to be in the range of 0 to 15. Table 7.1 summarizes the best fitness position on an average obtained for each particle at the iterations 100, 150, 200, 300, 500 and 600 of a when imputing a missing data in the attribute LAP for Hepatitis patients.

Table 7.1 Fitness Position (Missing Attribute-LAP)

Number of Particles	Fitness Position					
	100 Iterations	150 Iterations	200 Iterations	300 Iterations	500 Iterations	600 Iterations
5	4.22	4.25	4.11	3.98	4.01	3.87
10	3.145	3.241	3.334	3.51	3.11	3.07
15	3.411	3.417	3.356	2.882	2.545	2.541
20	2.449	2.401	2.417	1.503	1.503	1.502

The selection of the number of particles and iterations is based on minimal RMSE value. For Hepatitis and Thrombosis data the number of particles was chosen to be 20 and the iterations to be 300. Figure 7.5 shows the plot to illustrate the best fitness position taken on average at different iterations when imputing a missing data in the attribute LAP for Hepatitis patients. The X-axis represents the 300 iterations; Y-axis represents the best fitness value of nearest significant neighbour set formed for the corresponding missing attribute. It was observed that on an average best fitness value is found to be at the position 1.5; this value is taken as optimal value for the influence factor. This process is repeated for each generated nearest significant set in significant data selection process.

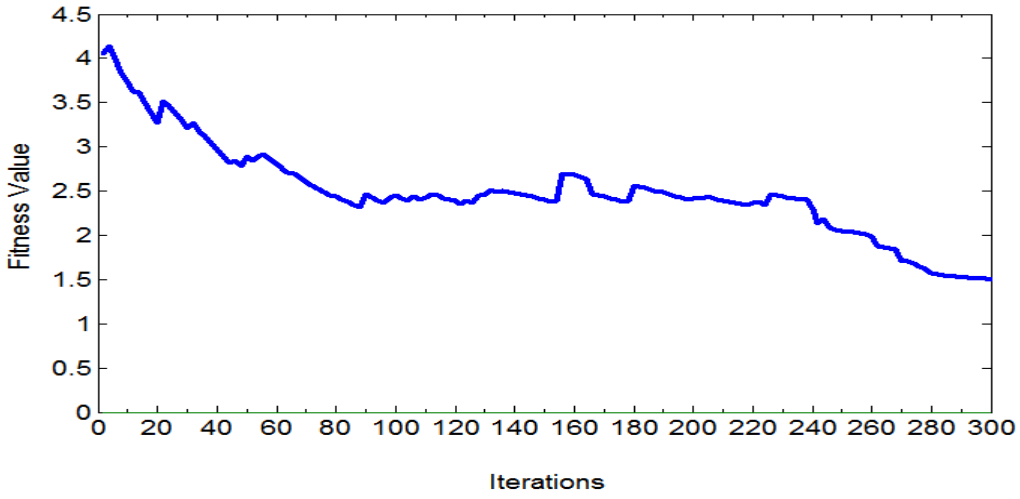

Figure 7.5 Plot on Fitness Value Vs Iterations (Missing Attribute-LAP)

A 10-fold cross validation with two independent runs evaluation scheme is adopted for generating the training and test subset. For evaluating the experimental results the performance metrics considered are MAD (Mean Absolute Deviation), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), FB (Fractional bias error), Index of Agreement (IA). These metrics were defined in the Equations (3.6) to (3.10) in Chapter 3. Table 7.2 and 7.3 summarizes the imputation results over the considered

performance measures for Hepatitis and Thrombosis dataset. The imputation results for the nine attributes (F-A/G, LAP, F-A1_GL, U-SG, AMY, CHE, CL, F-CHO and G-GTP) in the Hepatitis dataset and seven attributes (TG, CPK, WBC, RBC, PLT CRE and HGB) in the Thrombosis dataset are shown in the Table 7.2 and 7.3. The proposed TRiBS method is compared with the imputation techniques such as KNN, EM and IDW.

The error rates MAD, RMSE and MAPE value for the proposed TRiBS has reduced compared to the KNN, EM and IDW. The IA value of TRiBS is closer to 1 which proves the effectiveness of the imputed value. Similary the FB value of TRiBS for the imputed attributes in Hepatitis and Thrombosis dataset is closer to 0 which shows the reduction in the bias error of the imputed results.

From the obtained imputation results it can be observed that on an average the IA for the proposed improved IDW method is closer to 1 thereby ensuring that there is strong correlation between the actual observed and imputed value. The FB values prove that the proposed methodology highly limits underestimation and overestimation. A statistical paired t-test was carried out with significance value of 0.05 to assess the error rates of TRiBS over KNN, EM and IDW. The p-value obtained (p-vale<0.05) shows that there is a significant difference in the error rates for TRiBS over KNN, EM and IDW.

The impact of the proposed missing data imputation is also evaluated by extracting temporal features from the imputed data and applying classification techniques SVM, Neural Networks (NN), Decision Trees (DT) using R-Data mining tool. Figure 7.6 shows the comparison of classification accuracies of the classifiers NN, DT and SVM in combination with the proposed TRiBS, KNN, EM and IDW imputation techniques.

Table 7.2 Comparison of Imputation Results TRiBS, KNN, EM, IDW for Hepatitis Datasets

Performance Measures	Imputation Techniques	Hepatitis Attribute								
		F-A/G	LAP	F-A1_GL	U-SG	AMY	CHE	CL	F-CHO	G-GTP
MAD	TRiBS	0.22	14.2	0.32	0.73	4.3	0.41	6.9	2.8	2
	KNN	0.72	36.8	0.88	1.65	6.9	1.17	22.2	7	5.1
	EM	1.04	43.2	0.75	2.33	22.9	1.61	23.7	7.7	8.2
	IDW	1.25	61.9	1.29	2.81	27.1	2.37	31.5	10.2	9.3
RMSE	TRiBS	0.26	16.04	0.33	0.82	4.83	0.5	7.4	2.93	2.1
	KNN	0.77	44.59	0.96	1.77	17.93	1.24	25.1	7.55	5.92
	EM	1.05	50.63	0.9	2.37	24.77	1.76	26.17	8.61	9.1
	IDW	1.26	69.43	1.84	2.87	28.74	2.46	31.68	11.33	10.44
MAPE	TRiBS	10.37	3.94	9.56	9.62	5.41	4.39	6.61	5.03	7.76
	KNN	33.29	10.38	26.25	21.55	21.07	12.62	21.26	12.52	18.58
	EM	48.36	12.12	22.01	30.39	28.15	17.16	22.71	13.69	30.68
	IDW	58.74	17.41	40.77	36.7	33.37	25.19	30.16	18	34.33
IA	TRiBS	0.63	0.93	0.86	0.73	0.84	0.77	0.35	0.63	0.98
	KNN	0.24	0.7	0.39	0.44	0.37	0.46	0.08	0.08	0.88
	EM	0.2	0.63	0.37	0.34	0.26	0.38	0.09	0.17	0.76
	IDW	0.18	0.53	0.24	0.29	0.23	0.21	0.09	0.15	0.7
FB	TRiBS	0.11	0.02	0.09	0.1	0.02	0.01	0.07	0.03	0.02
	KNN	0.41	0.11	0.31	0.25	0.18	0.03	0.13	0.04	0.09
	EM	0.64	0.13	0.26	0.36	0.27	-0.04	0.15	0.06	0.14
	IDW	0.84	0.2	-0.18	0.45	0.31	0.05	0.35	0.09	0.17

Table 7.3 Comparison of Imputation Results TRiBS, KNN, EM, IDW for Thrombosis datasets

Performance Measures	Imputation Techniques	Thrombosis Attribute						
		TG	CPK	WBC	RBC	PLT	CRE	HGB
MAD	TRiBS	8.7	1.9	0.58	0.39	13.7	0.08	0.59
	KNN	19.7	6.2	1.038	0.79	51.3	0.13	1.78
	EM	16.8	6	1.34	0.75	56.1	0.12	2.12
	IDW	27.4	8	1.76	1.267	62.4	0.22	2.84
RMSE	TRiBS	10.38	1.97	0.65	0.49	16.54	0.1	0.68
	KNN	23.83	7.5	1.15	0.94	56.81	0.15	1.88
	EM	21.61	6.84	1.41	0.91	62.04	0.13	2.22
	IDW	29.25	8.99	1.85	1.52	64.55	0.26	2.9
MAPE	TRiBS	8.86	8.2	8.41	8.38	4.78	8.97	5.06
	KNN	20.04	24.34	14.74	17.16	16.93	14.53	15.67
	EM	16.4	24.69	19.21	16.2	18.21	13.02	18.61
	IDW	29.44	31.85	25.11	27.17	21.11	26.39	24.85
IA	TRiBS	0.95	0.98	0.75	0.39	0.97	0.92	0.97
	KNN	0.73	0.74	0.44	0.15	0.67	0.81	0.82
	EM	0.74	0.62	0.38	0.17	0.63	0.87	0.76
	IDW	0.66	0.67	0.3	0.11	0.62	0.69	0.68
FB	TRiBS	0.15	0.08	0.1	0.11	0.05	0.05	0.08
	KNN	0.34	0.1	0.29	0.17	0.26	0.07	0.24
	EM	0.27	0.16	0.38	0.19	0.32	0.07	0.28
	IDW	0.47	0.25	0.5	0.31	0.32	0.14	0.38

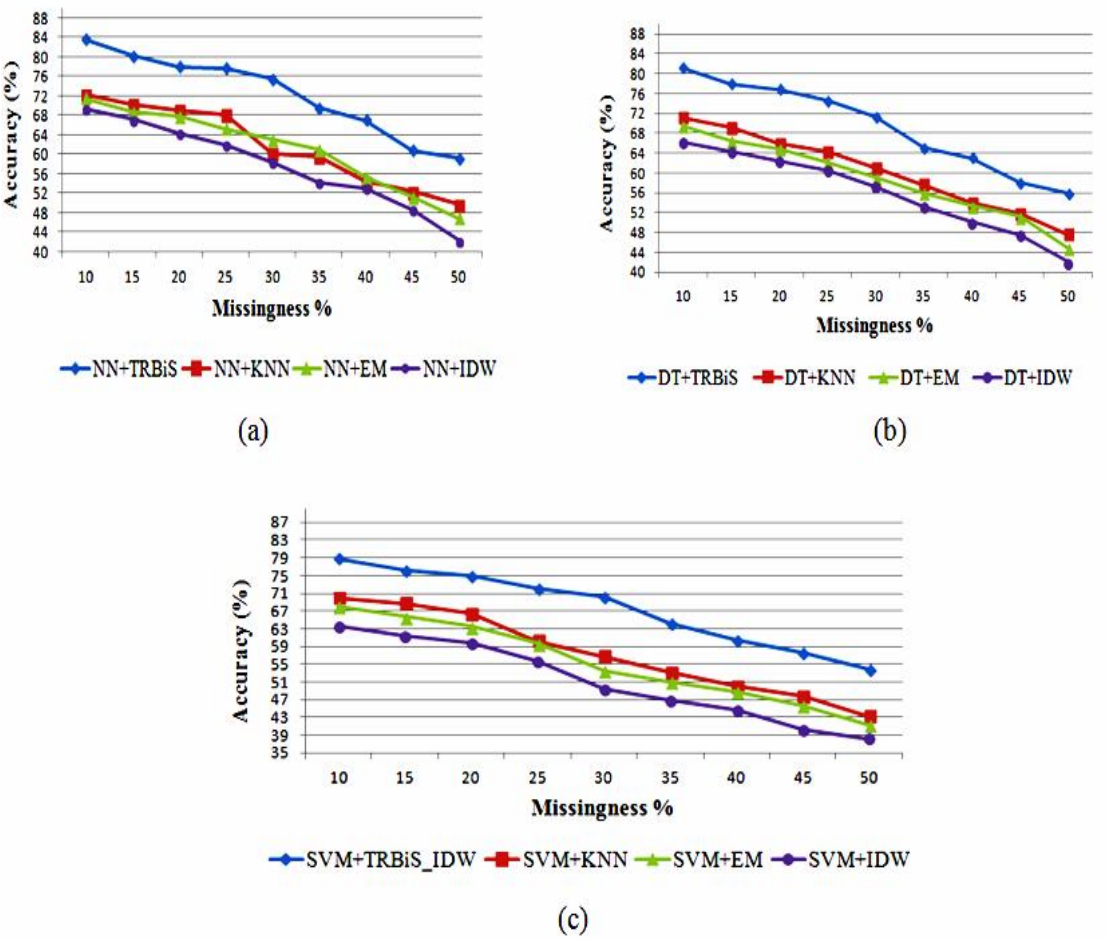

Figure 7.6 Comparison of classification accuracies: (a) NN with TRiBS, KNN, EM and IDW (b) DT with TRiBS, KNN, EM and IDW and (c) SVM with TRiBS, KNN, EM and IDW

The obtained classification results show that the combination of NN with TRiBS attains the classification accuracy of 83.57 % and 80.15 % for the missing rate of 10% and 15 % respectively. This was found to be higher when compared to NN with KNN, EM and IDW imputations. The combination of DT with TRiBS attains the classification accuracy 81.14 % and 77.91 % for the missing rate of 10% and 15 % respectively which was found to be higher when compared to DT with KNN, EM and IDW imputations. The combination of SVM with TRiBS attains the classification accuracy 78.89 % and 76.19 % for the missing rate of 10% and 15 % respectively which was found to be higher when compared to DT with KNN, EM and IDW

imputations. It can be inferred from the Figure 7.6 that there is a significant improvement in the classification accuracy for the data pre-processed data using proposed TRiBS imputation technique even when the percentage of missingness increases.

7.4 CONCLUSION

Clinical time series data are often observed at irregular intervals and hence they are referred to as unevenly spaced data. The presence of missing data makes it difficult to be used in knowledge discovery process. The proposed tolerance rough set induced bio-statistical framework handles the missing value by improving the traditional IDW techniques using the concept of tolerance rough set (TR) and particle swarm optimization (PSO). A rough set concept was used in TR to select the neighbourhood set for each unknown data points. The PSO technique has been used to find the optimal value for the influence factor to fix the weightage of the each known data points in the neighbourhood set. IDW interpolation process is carried out using the obtained neighbourhood set and influence factor for the corresponding unknown data locations. To demonstrate the proposed work two clinical time series data of Hepatitis and Thrombosis patients were considered. The experimental results shows that the proposed system has reduced the error rate and improved the accuracy of imputed results compared to the other imputation techniques like hot-deck, KNN and traditional IDW.