EXPLORATIONS ON ENGINEERING LETTERS (EEL)
VOLUME 1, ISSUE 1 (2016):PP.17-26
SANA ACADEMIC PRESS

A Data Analysis Model Using Variable Time-Series Data for Health Care

P.Saravanakumar[1], S.Lina[2], N.Venkatesan[3]
[1]Research Scholar, St.Peter's University, Chennai.
[2,3]Asst.Prof. Dept. of Computer Applications, St.Peter's University, Chennai.
djsaravanaa@gmail.com

Abstract - *A massive amount of time stamped data are generated in various domains such as health care, engineering, intrusion detection, stock market analysis and many more. This increasing availability, demands the analysis of temporal data for the purpose of prediction and classification. Time series classification is a challenging task due to the presence of high dimensional heterogeneous data; hampered with missing values and sampled at irregular time intervals. The project thus focuses on designing a classification model for multivariate time serious data that could be used for decision support system. In order to preserve the temporal characteristics of the data, the proposed methodology includes three phase. First, the missing values are imputed using Particle Swarm Optimization (PSO) based Inverse Distance Weighted Interpolation method (IDW). Secondly, the temporal data are transformed into high level symbolic representation (states and trends), through temporal abstraction based on Piecewise aggregation using dynamic window sizing. Finally, the abstracted states and trends are used as features to induce the Support Vector Machine (SVM) classifiers. The work is demonstrated using real dataset containing clinical trials of thrombosis patients and the results shows that the SVM classifiers has outperformed other classifiers with 95% classification accuracy.*

Keywords: *PSO, IDW, SVM, multivariate, temporal data.*

1. INTRODUCTION

Large volume of information's (data) are generated in many fields like medicine, stock markets geosciences etc., these data may have temporal characteristics. Existing Knowledge from temporal data is a cumbersome process as traditional data mining techniques cannot be applied directly. Time series data are also hampered with missing values which may produce inaccurate results. Moreover these data are challenging and requires immense pre-processing before applying mining technique to get accurate results. The project focuses on finding solutions to these challenges by designing a classification model suitable for decision support system employing appropriate pre-processing and abstraction methods. In this project Temporal Data Mining technology is used, which mines or discovers knowledge and patterns from temporal databases, with capability to include time series data.

2. CHALLENGES IN MINING TIME SERIES DATA

The challenges in mining time series data are: Irregular or uneven intervals, Presence of Outliers, Missing data and Complexity in handling high dimensional data. In clinical trials a patient's state of health may be observed only at irregular time intervals, and different patients are usually observed at different points of time. Through re-sampling of the data could be a solution it does have some drawbacks. It may produce biased result, affect the casual relationships in a multivariate time series, reduces and dilutes the information content of a data set causing statistical inference to be less efficient. It is common for longitudinal clinical trials to face problems with missing data that occur when patients do not complete the study or lose some visit. In most clinical trials, some patients do not complete their intended follow-up according to protocol, for a variety of reasons, and are often described as having 'dropped out' before the conclusion of the trial. Their subsequent measurements are missing, and this makes the analysis of the trial's repeated measures data more difficult.

3. TEMPORAL CLASSIFICATION

The task of temporal classification is defined as "Given an unlabeled sequence or time series T, assign it to one of predefined classes". On the other hand, the task of time series forecasting is defined as follows: "Given a time series T that contains n data points, predict its future values at time n+1, n+2,….n+z". In time series forecasting, the goal is to learn a model that can predict future values of a time series based on its past values. This area has been extensively studied in statistics [Shumway and Stoffer, 2006]. One of the most popular techniques is Auto –Regressive Integrated Moving Average (ARIMA). Generalized Auto-Regressive Conditional Heteroscedastic (GARCH) is another popular method that is used to model changes of variance along time. In temporal classification, each sequence (time series) is assumed to belong to one of the many predefined classes and the goal is to learn a model that can classify future sequences. In the following, the main temporal classification approaches are described.

4. ARCHITECTURE OF THE PROPOSED SYSTEM

The overall architecture of the proposed system is shown in Figure 1. Data pre-processing is done as an initial step of data analysis, the proposed system imputes the time series data using PSO based Inverse Distance Weighing Interpolation method. Temporal abstraction is applied to the high dimensional time series data. The discretized output (states and trends) is used to train the SVM Classifiers thus resulting in a classification model.

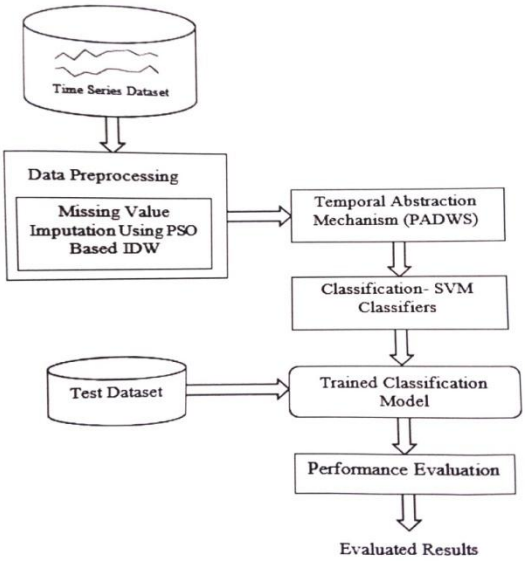

Figure1. System Architecture

5. IMPUTATION OF MISSING VALUES

Imputation provides a useful strategy for dealing with data sets with missing values. Instead of filling in a single value for each missing value, use multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. In the case of mean imputation the precision tends to be higher and bias nearly zero for MAR and MCAR but accuracy decreases for non-ignorable missing data. Hot deck imputation with Euclidean distance and z -score standardization outperforms list wise deletion and mean imputation as it shows less bias and higher precision even for non-ignorable missingness.

5.1 PARAMETER OPTIMIZATION

IDW assumes that each measured point has a local influence that diminishes with distance. It gives greater weights to points closest to the prediction location, and the weights diminish as a function of distance, hence the name inverse distance weighted. Power function k, weights are proportional to the inverse of the distance (between the data point and the prediction location) raised to the power value k. As a result, as the distance increases, the weights decrease rapidly. The rate at which the weights decrease is dependent on the value of k. As it increases, the weights for distant points decrease rapidly. If the k value is very high, only the immediate surrounding points will influence the prediction. To estimate the value of Z(x)

$$Z(x) = \frac{\sum_{i=1}^{n} Z(i)/d(x,i)k}{\sum_{i=1}^{n} 1/d(x,i)k} \quad \text{[Equation 1]}$$

Where, Z(i) is the measured value at point i, d(x,i) is the distance between the point x and the point i, N is the total number of known neighbour points for point i, and k is a positive power parameter. An optimal power value can be determined by minimizing the root mean square error. The RMSE is obtained for several different power values (using the same dataset), the power that provides the smallest RMSE is determined as the optimal power. Particle Swarm Optimization is employed to obtain the optimal value for k.

5.2 PARTICLE SWARM OPTIMIZATION

In 1995, Kennedy and Eberhart [35] firstly introduced particle swarm optimization (PSO), which is a robust stochastic evolutionary computation technique based on the movement and intelligence of swarms. The PSO algorithm is based on a social -psychological principle. Compared with other stochastic optimization techniques like genetic algorithms (GAs) and simulated annealing (SA), PSO has fewer complicated operations and fewer defining parameters, and can be implemented easily. Because of these advantages, the PSO method has attracted a wide spread attention and has been widely used. The PSO algorithm randomly initializes the locations and velocities of particles in the D -dimensional search space. Each particle flies to the target area and adjusts its position according to its own experience and the best experience of its topological neighbour. The position and velocity of the particles are represented as X=(x11,x12,—,xiD) and Vi =(vi1,vi2,vi3,....viD) respectively. The personal previous position and global best is expressed as pbest and gbest. The positions and velocities of particles are updated according to the following equations:

vid = w * V + C_1* rand() * (pbest — xid) + C_2 * rand() * (g best —xid)
Xid = X id + vid,1 d <D
Velocity Clamping limits the velocity to the range (-vmax, vmax) to keep the particles from moving too far beyond the search space.

For a search space bounded by the range (—xmax, xmax).
vmax = k * xmax
For a search space bounded by the range (xmin, xmax)
vmax = k * (xmin — xmax)/2
Objective Function of PSO
The power factor and the RMSE are taken into consideration in framing the objective function of PSO.

Symbol	Description	Values
w	Weight	$0.4 \le w \le 0.9$
C_1 , C_2	acceleration constants	$0 \le C_l, C_2 \le 2$
rand ()	random number	[0, 1]
k	velocity clamp factor	$0.1 \le k \le 2.0$
xmin, xmax	Search space range	(0,25)
vmax	Velocity range	computed

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Z(i) - Zi(k,n)\right)^2}{n}} \quad \text{[Equation 2]}$$

Z, is the actual value measured at location i. Z,(k,n) is the estimated or the predicted value at location i with power k and the neighbour n.
Objective function is min, k[1,25] RMSE
From this the optimized value of k is obtained and is used as the power factor for imputing missing data using the IDW interpolation method.

6. TEMPORAL ABSTRACTION

Temporal abstraction [Shahar, 1997] transforms numeric time series from a low-level quantitative form into a high level qualitative form. More specifically, temporal abstraction segments the numeric time series into a sequence of states, where each state represents a property that holds during a time interval. These states become the building blocks to construct more complex time interval patterns. That is, temporal abstraction can be seen as a pre-processing step for time interval pattern mining. Usually, we would likethese abstractions to be meaningful and easy to interpret by humans so that the resulting patterns would be useful for knowledge discovery. In the following, we describe the three main temporal abstraction approaches: abstraction by clustering, trend abstractions and value abstractions.

6.1 ABSTRACTION BY CLUSTERING

Abstraction by clustering is the process of inductively deriving a set of the time series states for a numeric time series by clustering similar parts of [Das et al., 1998, Kadous, 19991 This approach, first extract all subsequence's of length w using a sliding window. It then Clusters the set of all subsequence's to obtain clusters Cl ,...,Ck.,based on certain distance metric for measuring similarity between two subsequence's of length w (e.g., the Euclidean distance or dynamic time wraping. Next it defines the abstracted version of T by assigning every subsequence to its corresponding cluster symbol: TO=(sj(1),sj(2),...,sj(n—w+1)),where Sj is the symbol for cluster Ci.

7. SVM CLASSIFICATION

SVM is primarily a classier that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. Each instance in the training set contains one target value and several attributes. The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes. If data is linear, a separating hyper plane may be used to divide the data. However it is often the case that the data is are non-linear and inseparable. To allow for this, kernels are used to non -linearly map the input data to a high-dimensional space. The new mapping is then linearly separable as shown in Figure 2

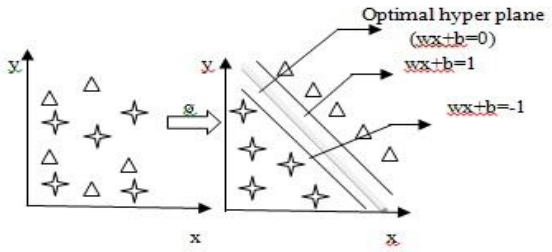

Figure 2. SVM CLASSIFIERS USING KERNEL

The project uses L1BSVM model to perform multiclass classification. To construct an optimal hyperplane. LIBSVM employs an iterative training algorithm, which is used to minimize an error function. According to the form of the error function LIBSVM models can be classified into four distinct groups: C-SVM, nu-SVM, epsilon-SVM, nu-SVM.

The model applied in the project is of type C-SVM. For this type of SVM, training involves the minimization of the error function:

$\frac{1}{2} w^t w + C \sum_{i=1}^{n} Ei$

Subject to the constraint

$yi(w^t 0(x_i) + b) \geq 1 — E_i$ and $E_{i \geq} 0, i = 1..N$

Where C is the capacity constant, w is the vector of coefficients, b is a constant, and EL represents parameters for handling non-separable data (inputs). The kernel 0 is used to transform data from the input (independent) to the feature space.

8. IMPLEMENTATION DETAILS

TABLES USED

Table 2: Description of tables contained in the dataset

S. No	Table Name	Description	No. of Attributes	No. Of Rows
1	TSUM_A (Table 2.1)	Contains general information's about the patients	8	1243
2	TSUM_B (Table 2.2)	Contains results of special laboratory tests (specific to thrombosis).	13	807
3	TSUM_C (Table 2.3)	Contains time stamped results of various lab tests	44	57453

Table 2.1 TSUM A

Attribute Name	Description	Type
ID	Patient unique identification	Numeric
Sex	Patient sex	Categorical
DOB	Date of Birth	YYYY/M/D
Description Date	The first date when a patient was recorded	YY.MM.DD
First Date	The date when a patient came to the hospital	YY.MM.DD
Admission	patient was admitted to the hospital (+) or followed at the outpatient clinic (-)	Categorical
Diagnosis	Disease names	Categorical

Table 2.2 TSUM_B

Attribute Name	Description	Type
ID	Identification of the patient	Numeric
Test Date	date of the test	YYYY/MM/DD
aCL IgG	anti-Cardiolipin antibody (IgG) concentration	Numerical
aCL IgM	anti-Cardiolipin antibody (IgM) concentration	Numerical
ANA	anti-nucleus antibody concentration	Numerical
ANA Pattern	Pattern observed in the sheet of ANA examination	Categorical
aCL IgA	anti-Cardiolipin antibody (IgA) concentration	Numerical
Diagnosis	disease names	Multi-valued attribute
KCT	measure of degree of coagulation	Categorical
RVVT	measure of degree of coagulation	Categorical
LAC	measure of degree of coagulation	Categorical
Symptoms	other symptoms observed	Multi-valued attribute
Thrombosis	degree of thrombosis	0: negative (no thrombosis) I: positive (most severe) 2:positive (severe) 3:positive (mild)

Table 2.3 TSUM_C

Attribute Name	Description	Type
Date	Date of the laboratory tests	YY.MM.DD
GOT	AST glutamic oxaloacetic transaminase	Numerical
GPT	ALT glutamic pylvic transaminase	Numerical
LDH	Lacate dehydrogenase	Numerical
ALP	Alkaliphophatase	Numerical
TPN	Total protein	Numerical
ALB	Albumin	Numerical
UA	Uric acid	Numerical
UN	Urea nitrogen	Numerical
CRE	Creatinine	Numerical
T-BIL	Total bilirubin	Numerical
T-CHO	Total cholesterol	Numerical
TG	Triglyceride	Numerical
CPK	Creatinine phosphokinase	Numerical
GLU	Blood glucose	Numerical
WBC	White blood cell	Numerical
RBC	Red blood cell	Numerical
HGB	Haemoglobin	Numerical
HCT	Hematoclit	Numerical
PLT	Platelet	Numerical
PT	prothrombin time	Numerical
APTT	Activated partial prothrombin time	Numerical
FG	Fibrinuria	Numerical
AT3	marker of DIC, one of the most important complications of collagen diseases	Numerical
A2PI	marker of DIC	Numerical
U -PRO	Proteinuria	Numerical
CRP	C-reactive protein	Categorical or Numerical
RA	Rhuematoid Factor	Categorical
RF	RAHA	Numerical
RNP	anti-ribonuclear protein	Numerical
CENTROMEA	anti-centromere	Numerical
DNA	anti -DNA	Numerical
DNA II	anti -DNA	Numerical
SM	anti-SM	Categorical
SCI70	anti-sc170	Categorical
SSA	anti -SSA	Categorical
C3	complement 3	Numerical
C4	complement 4	Numerical

9. ALOGORITHMS

Imputation of Missing Values

A PSO based IDW method is used to impute the missing values for the training dataset temporal in nature. The algorithms for PSO and IDW are given below. PSO is used to generate the optimal power function 'le which is later applied in the IDW method for imputing the missing values.

Input: The Objective function and the initial values for various parameters used in PSO,
Output: Optimized power function 'k'.

Begin
For each patient and each attribute
Initialize the particle position and velocity
Initialize the pbest to be its initial position
Repeat for each particle
Obtain the fitness value (rmse)

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(Z(i)-Zi(k,n)\right)^{2}}{n}} \quad \text{[Equation 3.]}$$

If the fitness value is better (minimum)
than all the previous values
set current pbest= particle position of the new
fitness value.
Choose the particle position with best fitness value
as the gbest.
For each particle
Update particle velocity vid
vid = w * vid + cl * rand( ) * (pbest —
xid) + c2 * rand() = (gbest — xid)
Update particle position xid
xid = xid + vid
End for.
Until maximum iteration or minimum error criteria is
not attained.
End for
End

Algorithm for Particle Swarm Optimization

Input: k value generated from PSO and the incomplete training dataset.
Output: Imputed complete dataset.
Begin
For each patient and attribute
Obtain the optimal power function 'k'
Identify the known and unknown points from
the training dataset.
For each unknown point (observed value)
Estimate the distance between each

data
points and its
neighbors using the sampled intervals
Compute the predicted value as

$$z(\mathrm{x}) = \frac{\sum_{i=1}^{n} z(i) / d(x,i)^k}{\sum_{i=1}^{n} 1 / d(x,i)^k}$$

[Equation 4.]

Impute the predicted value in the dataset.
End For
End For
End

Algorithm for Imputing Missing Values using IDW

10. RESULTS

IMPUTING OF MISSING VALUES

The dataset considered for the project is a real time temporal dataset of thrombosis patients. The dataset contains missing values as shown in the figure 3.

The original dataset is given as input to the pre-processing phase. In order to impute the missing values the optimal power function 'k used by the IDW interpolation is generated by the PSO technique. Figure 4 shows the optimal value generated for a particular patient and specific variable.

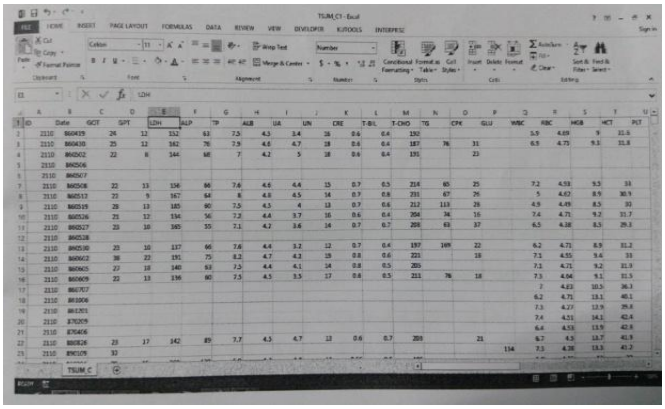

Figure 3: Original Dataset (Incomplete)

The imputed dataset using PSO based IDW. It does not impute for attributes with more than 50% missingness. Hence temporal attributes like TAT, APTT, U -PRO with high percentage of missingness are not imputed.

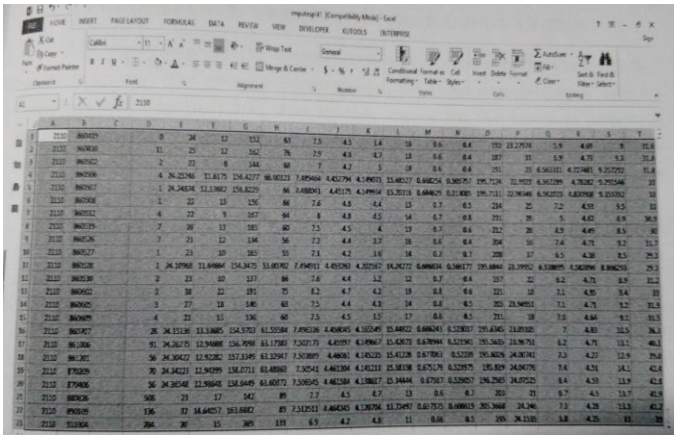

Figure 4: Optimized power function 'k' using PSO.

At the end of pre-processing phase, the incomplete dataset taken as input is transformed into a complete dataset by applying PSO based IDW interpolation method. This method is applied to TSUM_C.xls table which contains temporal data. During this phase the attributes with more than 50% missingness are not imputed and hence not considered for further data analysis. Thus out of 40 attributes 16 attributes are not considered for further analysis.

11. CONCLUSION

The project aims on designing a classification model for multivariate time series using SVM classifiers. Real dataset that includes clinical trials of thrombosis patients is used for experimental purpose. As pre-processing step imputation is carried out using PSO based IDW interpolation method. The experimental results show that PSO based IDW produces a minimal RMSE rate and thus outperforms various other existing imputation methods like MI. KNN and Simple IDW. The imputed temporal values are discretized through temporal abstraction. TA requires guidelines from domain experts, however, the technique used is not oriented to a particular domain it can be applied in other domains which share the same problem characteristics. Abstracted states and trends are used as features to train the SVM classifiers. From the experimental results it is found that SVM classifiers shows 95% classification accuracy and performs much better than other classifiers like Simple logistics, Naïve Bayes, MLP, Logistics and SMO. Thus a classification model is designed using SVM classifiers and can be used for decision support system.

12. FUTURE WORK

As a future work temporal patterns could be extracted from the time series data using Allen's temporal relation. These patterns can be used as features for classification. Moreover, the rules which divide the hypothetical planes in SVM classification can be obtained to identify the temporal patterns that are specific to each target attribute.

REFERENCES

[1]Alireza, F Lukasz Kurgan, A & Witold, P 2007, 'A novel framework for imputation of missing values in databases', IEEE transactions on systems, man, and cybernetics -part a: systems and humans, vol. 37, no. 5, pp. 692- 709.

[2] Dempster, AP Land, RJ & Rubin, DB 1977, 'Maximum likelihood from incomplete data via the EM algorithm', Journal of Royal Statistical Society, vol. 39, pp. 1-38.

[3] Marion, V Lucia, S Niels, P Riccardo, B Evert, J & Bas, M 2007, 'Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data', Elsevier, Artificial Intelligence in Medicine, vol. 41, pp. 1-12.

[4] Batal, I Valizadegan, H Cooper, G & Hauskrecht, M 2013 'A Temporal pattern mining approach for classifying electronic health record data', ACM Transaction on Intelligent Systems and Technology (ACM TIST), Special Issue on Health Informatics.

[5] Strike, K Khaled El, E & Nazim, M 2001, 'Software cost estimation with incomplete data', IEEE Transactions on Software Engineering, vol. 27, no. 10, pp. 890-908.

[6]Tanja, Genevieve, C Alexander, S & Carrie, R 2012, 'A global Water Quality Index and hot -deck imputation of missing data', Elsevier, Ecological Indicators, vol.17, pp. 108-119.

[7] Danielle, S & Rebecca, A 2015, 'A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern –mixture hot deck', Elsevier, Computational Statistics and Data Analysis, vol. 82, pp. 173-185.

[8] lngunn, M Erik, S & Ulf Olsson, H 2001, 'Analyzing datasets with missing data: an empirical evaluation of imputation methods and likelihood –based methods', IEEE Transactions on Software Engineering, vol. 27, no. 11, pp. 999-1013.

[9] Alan, 0 Shaw, C & Lisa, H 2003, 'The comparative efficacy of imputation methods for missing data in structural equation modeling', Elsevier, European Journal of Operational Research, pp. 53-79.

[10] Federico, C Andre Fialhoa, S Susana Vieirab, M Shane, R Joa Sousab, MC & Stan Finkelsteina, N 2013, 'Missing data in medical databases: Impute, delete or classify?', Elsevier, Artificial Intelligence in Medicine, pp. 63-72.

[11] Jose Jerez, M Ignacio, M Pedro Garci Laencina, J Emilio, A Nuria, R Miguel, M & Leonardo, F 2010, `Missing data imputation using statistical and machine learning methods in a real breast cancer problem', Elsevier, Artificial Intelligence in Medicine, pp. 105-115.

[12] Yaohui, D & Arun, R 2012, 'A comparison of imputation methods for handling missing scores in biometric fusion', Elsevier, Pattern Recognition, vol. 45, pp. 919-933.

[13] I ffat, A Gheyas, N & Leslie Smith, S 2010, 'A neural network –based framework for the reconstruction of incomplete data sets', Elsevier, Neuro Computing, vol. 73, pp. 3039-3065.