Journal of Engineering and Applied Sciences 12 (21): 5473-5477, 2017
ISSN: 1816-949X

A HCC Recurrence Prediction in Multiple Time Series Clinical Data with Merging Statistical Measures of Advanced Frequency Spectrum of Time Series Features

[1]P. Radha and [2]R. Divya
[1]Department of Information Technology,
[2]Department of Computer Science, Government Arts College (Autonomous), Coimbatore,
Tamil Nadu, India

Abstract: Now a days clinical data mining is used for clinicians in order to provide diagnosis, therapy and prognosis of different diseases. The accuracy of clinical-outcome prediction has been increased by using multiple measurements which are gathered from different time period and dataset. The multiple measurements are merged by using merging algorithm and the distribution of data is determined by statistical measurement. Then those data are given to the classifier for predicting the recurrence and non-recurrence of Hepatocellular Carcinoma (HCC) patients. In this study, an improved multiple time series clinical data processing is proposed. In the proposed approach, an additional measurement feature according to the frequency interval of features is included for reducing the error rate of classifier and increasing the prediction rate. The frequency based measurement feature is computed based on curvelet transform. Then, the optimal features are selected based on the Firefly optimization algorithm for reducing the classification overhead. The selected optimal features are learned by using the Support Vector Machine (SVM) classifier for predicting the patients with HCC disease and patients without HCC effectively. Finally, the experimental results prove that the proposed method has better performance than other classification methods.

Key words: Clinical data mining, Hepatocellular carcinoma, Curvelet transform, Firefly algorithm, support vector machine, optimization

INTRODUCTION

The major issue in data processing is different data characteristics since, there are two types of data are such as time-series data and cross-sectional data. The time-series data is referred as the sequence of observations of the certain feature which are ordered in time whereas the cross-sectional data is referred as the collection of many features at the equivalent time. For clinical data analytics, designing of data processing techniques such as data cleaning, data integration, data transformation and data reduction are the most essential for handling both cross-sectional data and time series data at the identical time and improving the quality of the analysis.

The prediction of the patients who had HCC was done by Radio Frequency Ablation (RFA) (Tseng *et al*., 2015). Initially, the multiple-time-series data from different datasets were extracted and cleaned. Then, the time-related data from the defined time period were combined by the data merging algorithm and the statistical measures were measured for classification which is performed by using Multiple Measurements SVM (MMSVM) (Tatsumi *et al*., 2010) and Multiple Measurements Random Forest (MMRF) classifiers (Khalilia *et al*., 2011). The classification performance was optimized based on the grid search and cross validation. However, the optimal time period for generating the time series data was required to be automatically determined according to the characteristics of the dataset used. Hence, in this study, the optimal time period is selected by using the additional frequency based measurement feature which is calculated by curvelet transform. The Firefly algorithm is introduced for selecting the optimal features and the selected optimal features are given to the SVM classifier for classification process which is used to predict the HCC disease accurately.

Literature review: Feature selection in clinical data processing (Seethal *et al*., 2016) was proposed for predicting the HCC recurrence from multiple measurements data. Initially, each feature in a definite time

Corresponding Author: P. Radha, Department of Information Technology, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

period was combined by the period merging algorithm and divided into training set and testing set. In training set, the irrelevant features were eliminated and correlation was measured by the relevant features and minimum spanning tree was constructed. By removing the edges from tree, the most representative features were only selected for classification. However, the construction of tree was difficult and time delay was high. The novel hierarchical technique (Liu and Hauskrecht, 2015) was proposed by combining the linear dynamical system and the Gaussian process for modeling the clinical time series data with varied data length and irregularity sampled observations. However, the limitation of the system was modeling of univariate time series and its analysis.

The novel method (Durichen *et al*., 2015) was proposed by using the Multi-Task Gaussian Process (MTGP) for modeling the multiple correlated multivariate physiological time-series simultaneously. The flexible MTGP Model was used to learn the correlation within the multiple signals with different frequencies and different time periods. However, the computational cost was high. An imputation approach named FLk-NN (Rahman *et al*., 2015) was proposed for incorporating the time lagged correlations both within and across variables by combining k-NN and fourier transform methods. The proposed approach was used for enabling the imputation of missing values even when all data at the time point was missing and when there were different types of missingness both within and across variables. However, the time complexity was high. The Multivariate Shapelets Detection (MSD) method (Ghalwash and Obradovic, 2012) was proposed to allow early and patient-specific classification of multivariate time series. The extracted time series from all dimensions of the time series were called as multivariate shapelets which distinctly manifest the target class locally. After that, the time series were classified by searching for the earliest closest patterns. However, the limitation of the proposed method was run time of the MSD.

MATERIALS AND METHODS

In the proposed research, the liver patient database collected around Tirupur at the time of 7, 14, 21, 60, 90 and 120 days which includes Hospital Information System (HIS), Laboratory Information System (LIS) and Radiology Information System (RIS) database are given as input and the multiple features from different datasets at similar time period are combined using merging algorithm. Then, the statistical measure and frequency of each data in every feature is measured by using curvelet transform and the firefly algorithm is applied for selecting the optimal features and the selected features are given to the MMSVM for classifying the data as patients with HCC disease and patients without HCC disease.

Merging algorithm for multiple time series data: The multiple time series data based on the defined time period are merged by the merging algorithm with the aim of selecting the most recent values for representing the feature. Initially, the length of time period is defined and only one value is selected for feature in one period.

Statistical measurement and frequency measurement: The merged data are considered into statistical measurement and frequency measurement. In statistical measurement, the values are measured for acquiring the distribution of the data in each time period. The distribution of data is measured by using Pearson's correlation coefficient range between -1 and 1. Due to statistical measure, the valuable information are retained which are lost during merging process. In frequency measurement, the time-series data are transformed into the frequency domain and curvelet functions are used for decomposing the data into multiple components. Then, it is reconstructed without losing of the original information in the database. From curvelets, the frequencies of data are calculated and it is included as additional features with the merged database. The basic curvelet ϕ is constructed based on the two major principles (Ma and Plonka, 2010) such as consider the polar coordinates in frequency domain and construct the curvelet elements being locally supported near wedges. The number of wedges is $N_j = 4.2^{j/2}$ at the scale 2^{-j} such that it doubles in each second circular ring. Consider the variable in frequency domain is: $\xi = (\xi_1, \xi_2)^T$ and the polar coordinates in the frequency domain are $r=\sqrt{\xi_1^2+\xi_2^2}$ and $\omega = \arctan \xi_1/\xi_2$. For the dilated basic curvelets in polar coordinates, the ansatz is utilized:

$$\hat{\phi}_{j,0,0}(r,\omega) := 2^{\frac{-3j}{4}} W\left(2^{-j_r}\right)\tilde{V}_{N_j}(\omega),\ r \geq 0,\quad \omega \in (0,2\pi),\ j \in N_0 \tag{1}$$

The basic curvelet along with the compact support near a basic wedge is constructed in which the two windows W and $\tilde{V}_{N_j}$ are required to have compact support. The principle is taking W(r) to cover (0, ∞) with dilated curvelets and $\tilde{V}_{N_j}$ such that each circular ring is covered by the translations of $\tilde{V}_{N_j}$. Then, the admissibility yields:

$$\sum_{j=-\infty}^{\infty}\left|W\left(2^{-j_r}\right)\right|^2 = 1, r \in (0,\infty) \tag{2}$$

For tiling a circular ring into N wedges, 2π periodic nonnegative window $\tilde{V}_{N_j}$ is required with support inside $\left[\frac{-2\pi}{N}, \frac{2\pi}{N}\right]$ such that:

$$\sum_{l=0}^{N-1} \tilde{V}_N^2\left(\omega - \frac{2\pi l}{N}\right) = 1, \text{for all } \omega \in (0, 2\pi) \quad (3)$$

Where N is an arbitrary positive integer. $\tilde{V}_N$ is simply constructed as 2π periodizations of the scaled window $V\left(\frac{N\omega}{2\pi}\right)$. Then it follows that:

$$\sum_{l=0}^{N_{j-1}} \left|2^{\frac{-3j}{4}} \hat{\phi}_{j,0,0}\left(r, \omega - \frac{2\pi l}{N_j}\right)\right|^2 = \left|W\left(2^{-j_r}\right)\right|^2 \sum_{l=0}^{N_{j-1}} \tilde{V}_N^2 \quad (4)$$

$$\left(\omega - \frac{2\pi l}{N}\right) = \left|W\left(2^{-j_r}\right)\right|^2$$

For the complete covering of the frequency plane including the region around zero, the low pass element is required to be defined:

$$\hat{\phi}_{-1} = W_0(|\xi|), \text{ with } W_0^2(r)^2 = 1 - \sum_{j=0}^{\infty} W\left(2^{-j_r}\right)^2 \quad (5)$$

Equation 5 is supported on the unit circle where any rotation is not considered. Thus, the curvelets and statistical measurement features are obtained.

Feature selection and classification: The feature selection process has high complexity due to the including additional features in the dataset (Reyhaneh and Mahdi, 2016). Hence, firefly optimization algorithm is applied for selecting the optimal features. The population of fireflies are initialized in which each firefly has two significant characteristics like variation in light intensity and formulation of the attractiveness. Consider the objective function f (a) and the intensity of each firefly is defined as:

$$I(a) = \max f(a) \quad (6)$$

The attractiveness function of each firefly is measured as follows:

$$\xi(r) = \xi_0 \times e^{-\gamma . r^2} \quad (7)$$

Here, ξ_0 is the attractiveness at distance r = 0. The light absorption coefficient γ is computed as $\gamma = 1/\Gamma^m$ where Γ is called as the characteristic length scale in an optimization problem. The distance between two fireflies such as x and y at position p_x and p_y is computed based on the cartesian distance:

$$r_{xy} = \left\|p_x - p_y\right\| = \sqrt{\sum_{k=1}^{d}\left(S_{x,k} - S_{y,k}\right)^2} \quad (8)$$

Here, $S_{x,k}$ is the kth component of the spatial coordinate p_x of xth firefly. The movement of xth firefly to yth firefly which has more attractiveness is defined as:

$$p_x = p_x + \xi_0 e^{-\gamma^2 xy}\left(p_y - p_x\right) + a \operatorname{sign}\left[\text{rand} - \frac{1}{2}\right] \oplus \text{Levy} \quad (9)$$

Where,

$$\text{Levy} \sim u = t^{-\lambda}, 1 < \lambda \leq 3 \quad (10)$$

Here the first term denotes the current position of xth firefly, the second term refers the attractiveness of the firefly x and y. The third term represents the randomization through the Levy flights where α is assumed as randomization parameter. The sign [rand-1/2], rand ϵ[0, 1] is used for providing the random sign or direction in which the random step length is obtained from the Levy distribution with infinite variance and mean. Thus, the optimal features which has high accuracy are selected based on the attractiveness of the fireflies. Then, the optimized features are given to the input of SVM classifier. Initially, the training set of instance-label pairs are considered as (x_i, y_i) whereas $x_i \epsilon R^d$, $y_i \epsilon \{1, -1\}$, I = 1, 2, ..., N. The kernel function is given as:

$$k\left(x_i, x_j\right) = \exp\left(\frac{1}{\sigma^2}\left|x_i - x_j\right|\right)^2 \quad (11)$$

SVM is used for finding an optimal hyper plane by solving the following optimization problem:

$$H(x) = < w.x > + b \quad (12)$$

$$\text{Minimize}: \frac{1}{2} < w.w > + \sum_{ci=1}^{n} \eta_i$$
$$\text{Subject to}: y_i\left(< w.y_i > + b\right) + \eta_i - 1 \geq 0 \quad (13)$$

In Eq. 13, $\eta_i > 0$ and c refers the penalty parameter and η_i refers the slack variables. Based on this optimization problem, SVM detects the hyperplane which provides the minimum number of training errors.

Algorithm:

Input: Liver cancer dataset, HIS, LIS, RIS dataset
Output: Patients with HCC and Patients without HCC

1 Obtain the records from the dataset and time for specific event and sort in descending order of date
2 Initialize the merged records array based on time and features
3 For each record in the dataset
4 $i=\frac{\text{Time of specific event} - \text{Time of all records before time specific event}}{\text{Daysperiod}}$
5 Obtain the value of each feature nearest to the specific event time period i
6 Assign the value for each record as the most recent values
7 End for
8 Determine the frequency measurement for each data in the merged record
9 For each and every data in the record
10 Convert the data into frequency domain and calculate the frequency of data in each feature by curvelet function
11 Determine the statistical measure for each feature from the merged record
12 For each time period and each time related features
13 Determine the maximum, minimum, average, standard deviation, Pearson's correlation coefficient, slope of trend line of each feature in the record before specific event
14 Include all measured features as additional features into the merged record
15 Initialize the number of fireflies f_i, i = 1, 2, ..., n
16 Assign the objective function f (a_i), i = 1, 2, ..., d
17 For each firefly
18 Compute the light intensity I (a_i), i = max f (a_i)
19 Define light absorption coefficient γ
20 While (t<MaxGeneration)
21 For i = 1: n
22 For j = 1: i
23 If (I (a_y)>I (a_x)
24 Change attractiveness with distance r through $e^{\gamma r}$
25 Move xth firefly to yth firefly through Levy flights
26 Compute new solutions and update the light intensity
27 End if
28 End for j
29 End for i
30 Sort the fireflies and find the current best
31 End while
32 Obtain the optimal features and provide tuned features as input to SVM
33 Identify the right hyperplane by using the Eq. 12
34 Compute minimum training errors using the Eq. 13
35 Classify the data and predict the patients with HCC and patients without HCC

RESULTS AND DISCUSSION

In this study, the performance of proposed multiple time series classification method is analyzed in terms of accuracy and balanced accuracy. For experimental analysis, the data are collected around Tirupur location at a time of 120 days which consisting of HIS, LIS and RIS. The HIS consists of 152 records with attributes such as sex, age, height, weight and status of Cirrhosis. The LIS consists of 152 records with attributes such as Alkaline Phosphatase (ALP), Aspartate Transaminase (AST), Alanine Amino Transferase (ALT), Albumin, Bilirubin, Gamma-Glutamyl Transpeptidase (GGT) and Creatinine. The RIS includes 152 records with attributes such as tumor number and tumor size.

Accuracy: Accuracy is referred as the fraction of true outcomes such as both true positives and true negatives among the total number of cases examined. It is computed as follows:

$$\text{Accuracy} = \frac{\text{True positive+True negative}}{\text{True positive+True negative+ False positive+False negative}}$$

Balanced accuracy: Balanced accuracy is defined as the average of specificity and sensitivity values and is computed as follows:

$$\text{Balanced accuracy} = \frac{\text{Specificity+Sensitivity}}{2}$$

Figure 1 and 2 shows that the comparison of accuracy and balanced accuracy with different time interval and different methods such as MMSVM with statistical measure (MMSVMSM), MMRF classifier with statistical measure (MMRFCSM) and MMSVM with curvelet transform and statistical measure (MMSVMCTSM). From the graph, it is proved that the proposed MMSVMCTSM has better accuracy and balanced accuracy than other techniques.

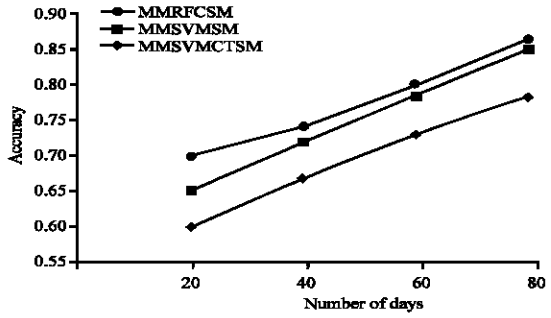

Fig. 1: Comparison of accuracy

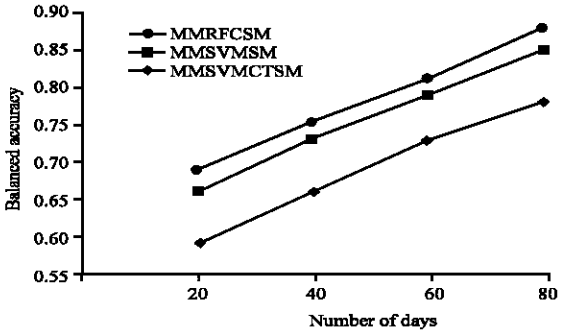

Fig. 2: Comparison of balanced accuracy

CONCLUSION

In this study, the multiple time series clinical data processing is enhanced by including the new frequency measurement feature and feature selection process. The proposed technique reduces the time consumption during data classification and improves the classification performance by considering the additional features. Then, firefly optimization algorithm is applied for selecting the optimal features and SVM classifier is used for classification. Thus, the proposed system predicts the patients with HCC and patients without HCC accurately. Finally, the experimental results are proved that the proposed MMSVMCTSM approach has better performance than the other multiple time series classification approaches.

REFERENCES

Durichen, R., M.A. Pimentel, L. Clifton, A. Schweikard and D.A. Clifton, 2015. Multitask Gaussian processes for multivariate physiological time-series analysis. IEEE. Trans. Biomed. Eng., 62: 314-322.

Ghalwash, M.F. and Z. Obradovic, 2012. Early classification of multivariate temporal observations by extraction of interpretable shapelets. BMC. Bioinf., 13: 195-195.

Khalilia, M., S. Chakraborty and M. Popescu, 2011. Predicting disease risks from highly imbalanced data using random forest. BMC. Med. Inf. Decis. Making, 11: 51-51.

Liu, Z. and M. Hauskrecht, 2015. Clinical time series prediction: Toward a hierarchical dynamical system framework. Artif. Intell. Med., 65: 5-18.

Ma, J. and G. Plonka, 2010. The curvelet transform. IEEE Signal Process. Mag., 27: 118-133.

Rahman, S.A., Y. Huang, J. Claassen, N. Heintzman and S. Kleinberg, 2015. Combining fourier and lagged K-nearest neighbor imputation for biomedical time series data. J. Biomed. Inf., 58: 198-207.

Reyhaneh, K. and A. Mahdi, 2016. Features selection from data in order to improve classification methods performance. J. Eng. Appl. Sci., 11: 1859-1865.

Seethal, C.R., J.R. Panicker and V. Vasudevan, 2016. Feature selection in clinical data processing for classification. Proceedings of the International Conference on Information Science (ICIS), August 12-13, 2016, IEEE, Kerala, India, ISBN:978-1-5090-1988-5, pp: 172-175.

Tatsumi, K., R. Kawachi and T. Tanino, 2010. Nonlinear extension of multiobjective multiclass support vector machine. Proceedings of the 2010 IEEE International Conference on Systems Man and Cybernetics (SMC), October 10-13, 2010, IEEE, Suita, Japan, ISBN:978-1-4244-6586-6, pp: 1338-1343.

Tseng, Y.J., X.O. Ping, J.D. Liang, P.M. Yang and G.T. Huang *et al*., 2015. Multiple time series clinical data processing for classification with merging algorithm and statistical measures. IEEE. J. Biomed. Health Inf., 19: 1036-1043.