International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 19 (2017) pp. 8443-8449

An Enhanced HCC Recurrence Prediction with Common Interest Features in Multiple Measurement Time Series Clinical Dataset

Dr. P. Radha
Assistant Professor, Department of Computer Science,
Government Arts College (Autonomous), Coimbatore, Tamilnadu, India.

R. Divya
Research Scholar, Department of Computer Science,
Government Arts College (Autonomous), Coimbatore, Tamilnadu, India.

Orcid: 0000-0002-7051-7449

Abstract

Hepatocellular Carcinoma (HCC) is the most common type of primary liver cancer in adults and it is the most common cause of death in people with cirrhosis. Hence, the HCC prediction is more important to provide treatment for patients. The data mining techniques is widely used for prediction of HCC patients. The accuracy of clinical outcome prediction has been increased by utilizing multiple measurements which are collected from different time period and dataset. It data is merged using merging algorithm and distribution of data is calculated by statistical measurement and it feed as input to classifiers which predict the recurrence and non recurrence of HCC. But sensitivity and Positive Predictive Value(PPV) of this method is low. So in this proposed work, additional features such as frequency based measurement features and common interest features are considered to improve the sensitivity and PPV value. The frequency based measurement feature is computed using wavelet transform and curvelet transform function. Common interest features are gender, habits, heredity, lifestyle, work, region and customs are obtained by using Latent Dirichlet Allocation(LDA). Then optimal features are selected using firefly algorithm. Finally, the selected optimal features are learned by using classifier called Support Vector Machine(SVM) to predict the patients with HCC and without HCC. The experimental results are conducted to prove effectiveness of proposed method over existing method.

Keywords: Hepatocellular Carcinoma, Multiple time series, clinical data mining, common interest features, Latent Dirichlet Allocation

INTRODUCTION

Generally, data mining refers the process of exploring and learning the most significant characteristics from the huge amount of database and performed based on the different processes. During data processing of data mining technique, the major issue is the different varieties of data characteristics like time-series data characteristics and cross-sectional data characteristics [1]. A sequence of data features ordered in a specific time is referred as time-series data characteristics and a collection of numerous features at the equivalent time is referred as the cross-sectional data characteristics. Nowadays, such data processing design is mostly required for clinical data analytics in order to handle both cross-sectional and time-series data characteristics simultaneously at the equivalent time [2].

Over the past decades, different data mining techniques have been developed for predicting the outcomes of patients. The prediction of HCC patients was developed and the patients were diagnosed based on the Radio Frequency Ablation (RFA) [3]. At first, the multiple time-series data characteristics were gathered and features were extracted and cleaned. After that, the time-related data characteristics from the certain time duration were merged by using data merging algorithm [4] and then the statistical measures were determined. Then, the merged multiple measurements data characteristics including with or without statistical measures were generalized for classification process. For classification process, Multiple Measurements SVM (MMSVM) [5] and Multiple Measurements Random Forest (MMRF) classifiers [6] were used. At last, the classification model was optimized by using grid search and cross validation methods. However, the sensitivity and Positive Predictive Value (PPV) of this method is low.

Hence in this paper, additional features such as frequency based measurement features and common interest features such as gender, habits, heredity, lifestyle, work, region and customs are considered along with the multiple measurement data for prediction of HCC. It improves the sensitivity and PPV values of HCC prediction method. The frequency based

measurement features are obtained based on the curvelet and wavelet transforms and the common interest features are acquired by using Latent Dirichlet Allocation (LDA) method. Then, the most optimal features are selected for classification process by using firefly optimization algorithm [7]. Finally, SVM based classifier is applied for predicting the HCC disease recurrences accurately.

RELATED WORKS

Harrington, P. L., et al. [8] proposed the classification of multiple time series by using the boosting algorithm. However, an efficient classification depends on the most significant processes like feature selection and sub-space selection. Su, W. T. et al. [9] proposed clinical data classification for multiple time series data processing by using the period merging algorithm. However, the performance values were not obtained to provide the significant impact effectively. Priyanka, M., et al. [10] proposed the multiple time series clinical data processing by suing the modified artificial bee colony and artificial neural network algorithm. In this approach, different classification algorithms were proposed such as Multiple Measurements Support Vector Machine (MMSVM), Multiple Measurements Random Forest Regression (MMRF), Improved Particle Swarm Optimization (IPSO), and Modified Artificial Bee Colony Algorithm (MABCA) [11] for solving the multiple time series issues by maximizing the optimal feature information. However, the accuracy requires further improvement. Batal, I., et al. [12] proposed the pattern mining approach for multivariate temporal data classification. In this approach, the minimal predictive temporal patterns method was presented for generating the smallest set of predictive and non-spurious patterns. However, the complexity of the approach was high.

Seethal, C. R., et al. [13] proposed the feature selection approach in multiple time series clinical data processing for predicting the HCC recurrence. However, the time delay and the complexity of tree construction were high.

Spiegel, S., et al. [14] proposed the pattern recognition and classification for multivariate time series. In this approach, the time series were separated into the segments and the recognized segments were clustered into the similar group context. This approach was evaluated by using the real-life sensor data from different vehicles and not evaluated using the clinical data processing. Ghassemi, M., et al. [15] proposed the multivariate time series approach to severity illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In this approach, Multi-Task Gaussian Process (MTGP) [16] was proposed for modeling the multiple correlate multivariate physiological time-series simultaneously. The computational cost of this method was high.

PROPOSED METHODOLOGY

In this section, the proposed methodology for prediction of HCC patients is explained in detail. Common interest features such as Gender, heredity, lifestyle, habits, work, region and customs are considered along with multiple measurement data and frequency based measurement feature. Common interests features are obtained by introducing LDA [17]. Frequency based measurement feature is calculated using wavelet and curvelet transform function. Then the most relevant features are selected using firefly algorithm. The selected optimal features are learned by using Support Vector Machine (SVM) [18] to predict the patients with HCC disease and patients without HCC disease.

Dataset Collection

The liver patient database is collected around Tirupur area at the time interval of 7, 21, 60, 90 and 120 days which includes Hospital Information System (HIS), Laboratory Information System (LIS) and Radiology Information System (RIS) database. The multiple features from the various datasets at similar time period are combined together using merging algorithm. In addition to multiple time series data, frequency measurement data and common interest data are considered for prediction of patients with HCC. The frequency measurement data represents the frequency of each data in every feature which is measured by using wavelet transform and curvelet transform. Common interest features includes gender, heredity, lifestyle, habits, work, region and customs of patients are presented in the dataset. These features are used to predict the patients with HCC.

Merging algorithm for Multiple Time

The multiple time series data defined at different time period are merged by using merging algorithm. The main aim of the merging algorithm is selecting the most recent values to represent a feature. In the merging algorithm initially define the length of time period and select only one value for a feature in one time period.

Statistical measurement and Frequency Measurement

The merged data are taken into statistical measurement and frequency measurement. The statistical measurement computes the distribution of data in each time period. While merging multiple time series data, some valuable information might be loss. It might be partly retained by calculating statistical measurement which is calculated using Pearson correlation coefficient and average. In frequency measurement, the data are transformed into frequency domain.

The data are decomposed into multiple components by using wavelet transform [19] and curvelet transform [20] without the loss of original information in the data. From the wavelets and curvelet transform the frequency of data is calculated and it is added as additional features with the merged dataset.

Common interests feature with LDA

Gender, heredity, lifestyle, habits, work, region, customs are the common interests feature are considered for prediction of patients with HCC. Along with multiple time series data and frequency measurement data, common interest features related with HCC are obtained by using LDA. LDA is a probabilistic, generative model which is used to obtain common interest features. In the LDA technique, documents refer the data, topics are referred as common interests and word represents the features of common interests. The process of LDA is defined as follows:

For each data indexed by $m \in \{1,2, \dots M\}$ in a dataset:

1. Choose a K-dimensional common interests weight vector θ_m from the distribution $p(\theta|\alpha) = Dirichlet\ (\alpha)$
2. For each features indexed by $n \in \{1, \dots, N\}$ in the data
    a. Choose a common interest from $z_n \in \{1, \dots, K\}$ from the multinomial distribution, $p(z_n = k|\theta_m) = \theta_m^k$.
    b. Given the chosen common interest z_n, draw a feature x_n from the probability $p(x_n = i|z_n = j, \beta) = \beta_{ij}$

In the LDA technique, α denotes the parameter of the Dirichlet prior on the per-data common interest distributions, β denotes the parameter of the Dirichlet prior on the per-common interest on the feature distributions. Thus the common interest features related with HCC through LDA.

Feature selection and classification

The optimal multiple time series data, frequency measurement data by wavelet and curvelet transform and common interest features are selected by using firefly algorithm. The population of fireflies is initialized in which each firefly randomly chosen the features of multiple measurement data, frequency measurement data by wavelet and curvelet transform and common interest features. Each firefly has two important characteristics like variation in light intensity and formulation of the attractiveness. Consider the classification accuracy as its objective function which is denoted as f (a) and the intensity of each firefly is defined as,

$$I(a) = maxf(a) \quad (1)$$

The attractiveness function of each firefly is calculated by using equation is given as follows:

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

where, ξ_0 represents the attractiveness at distance $r = 0$. The light absorption coefficient γ is computed as $\gamma = \frac{1}{\Gamma^m}$ where Γ is termed 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.

$$r_{xy} = \|p_x - p_y\| = \sqrt{\sum_{k=1}^{d} (s_{x,k} - s_{y,k})^2} \quad (3)$$

where, $s_{x,k}$ is the k^{th} component of the spatial coordinate p_x of x^{th} firefly. The movement of x^{th} firefly to y^{th} firefly which has more attractiveness is defined as,

$$p_x = p_x + \xi_0 e^{-\gamma r_{xy}^2} (p_y - p_x) + \alpha\ sign\left[rand - \frac{1}{2}\right] \oplus Levy \quad (4)$$

where,

$$Levy \sim u = t^{-\lambda},\ 1 < \lambda \leq 3 \quad (5)$$

where, the first term represents the current position of x^{th} firefly, the second term denotes the attractiveness of the firefly x and y. The third term refers the randomization through the Levy flights where α is assumed as randomization parameter. The $sign\left[rand - \frac{1}{2}\right]$, $rand \in [0,1]$ is utilized to provide the random direction or sign where the random step is obtained from the Levy distribution with infinite variance and mean. Hence the optimal features with high classification accuracy are selected based on the attractiveness of fireflies.

After the selection of optimal features, the SVM classifier is processed with the optimal features. Initially, the training set of instance-label pairs are considered as (x_i, y_i) whereas $x_i \epsilon R^d, y_i \epsilon \{1, -1\}, i = 1,2, \dots, N$. The kernel function is given as,

$$k(x_{i,}, x_j) = \exp(\frac{1}{\sigma^2} ||x_{i,} - x_j||)^2 \quad (6)$$

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

$$H(x) =< w.x > +b \quad (7)$$

Minimize: $\frac{1}{2}$ <w.w> +c$\sum_{i=1}^{n} \eta_i$

Subject to: $y_i(< w.y_i > +b) + \eta_i - 1 \geq 0 \quad (8)$

In equation (8), $\eta_i > 0$ and c denotes the penalty parameter and η_i represents the slack variables. Based on this optimization problem, SVM detects the hyperplane which provides the minimum number of training errors. Based on the training data, the test data are processed to predict the patients with HCC and patients without HCC.

Algorithm:

Input: Liver cancer dataset, HIS, LIS, RIS dataset, number of fireflies f_i, i =1,2,… n, light absorption coefficient γ, maxgen, t=1

Output: Patients with HCC and Patients without HCC

Step 1: Obtain the records from the dataset and time for specific event

Sep 2: Arrange the records in descending order of date

Step 3: Initialize the merged records array based on features and time

Step 4: For each record in the dataset

Step 5:

$$i = \frac{\text{time of specific event} - \text{time of all records before time specific event}}{days\ period}$$

Step 6: Get the value of each feature nearest to the specific event time period i

Step 7: Assign the value for each record in the dataset as the most recent values

Step 8: End for

Step 9: Determine the frequency measurement using wavelet and curvelet function for each data in the merged data

Step 10: Determine the statistical measure for each feature from the merged record

Step 10: Determine the common interest features in the merged data using LDA

Step 11: Include all measured features as additional features in the merged data

Step 12: Each firefly randomly choose the features

Step 13: Calculate objective function $f(a_i), i = 1,2, \dots, d$

Step 14: for each firefly

Step 15: Compute the light intensity

Step 16: while (t<maxgen)

Step 17: for i=1:n

Step 18: for j=1:i

Step 19: $If\left(I(a_y) > I(a_x)\right)$

Step 20: Change the attractiveness with distance r through $e^{-\gamma r}$

Step 21: Move x^{th} firefly to y^{th} firefly through Levy flights

Step 22: Compute new solutions and update the light intensity

Step 23: End if

Step 24: End for j

Step 25: End for i

Step 26: Sort the fireflies and find the current best which represents theoptimal features

Step 27: End While

Step 28: Provide the optimal features as input to SVM

Step 29: Determine the right hyperplane by using equation 7

Step 30: Calculate minimum training errors using equation 8

Step 31: Classify the data and predict the patients with HCC and patients without HCC

RESULTS AND DISCUSSION

In this section, 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:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Balanced Accuracy

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

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

Fig.1 and Fig.2, shows the comparison of accuracy and balanced accuracy between comparison of accuracy with different time interval and different methods such as MMRF Classifier with Statistical Measure (MMRFC-SM), MMSVM with Statistical Measure (MMSVM-SM), MMSVM with Curvelet Transform and Statistical Measure (MMSVM-CT-SM) and MMSVM with Curvelet Transform, Wavelet Transform, Common Interest and Statistical Measure (MMSVM-CT-WT-CI-SM). From the graph, it is proved that the proposed MMSVM-CT-WT-CI-SM has better accuracy and balanced accuracy than other techniques.

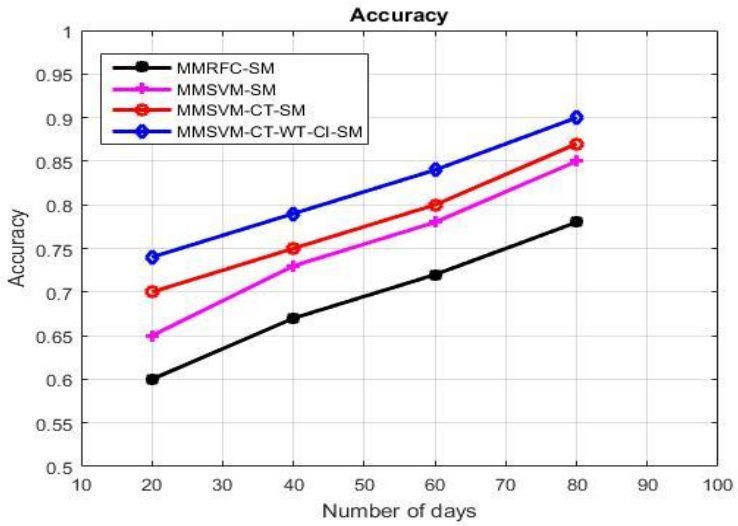

Figure 1: Comparison of Accuracy

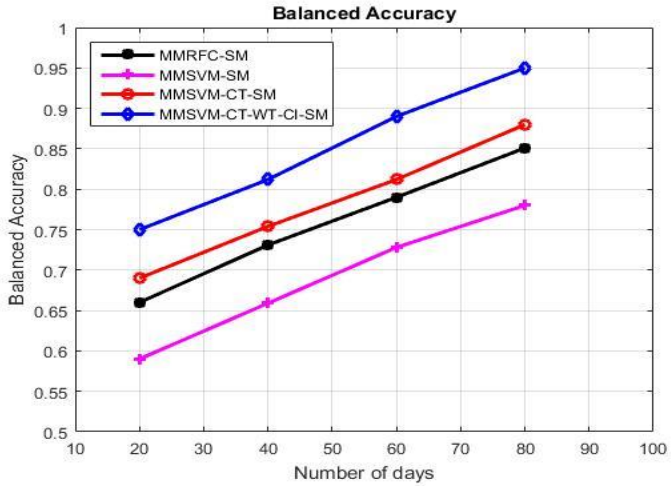

Figure 2: Comparison of Balanced Accuracy

CONCLUSION

The proposed system improves the classification of multiple time series clinical data by introducing common interest features are gender, heredity, lifestyle, habits, work, region, customs and frequency based measurement features. It improves the sensitivity and PPV value of HCC prediction method. By adding additional features in the dataset, the classification process becomes complex. So, the optimal features are selected using firefly algorithm. Thus it reduces the complexity of the SVM classifier. The optimal features are given as input to SVM that classifies the patients as patients with HCC and patients without HCC. The experiments are conducted in the liver patient dataset which is collected from Tirupr area and the performance of proposed system is measured in terms of accuracy and balanced accuracy. It proves that the proposed system has high accuracy and balanced accuracy than the other methods.

REFERENCES

[1] Nagavelli, R., & Rao, C. G. (2014, May). Degree of Disease possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining. In Recent Advances and Innovations in Engineering (ICRAIE), 2014 (pp. 1-6). IEEE.

[2] Scheidt-Nave, C., Kamtsiuris, P., Gößwald, A., Hölling, H., Lange, M., Busch, M. A., & Hapke, U. (2012). German health interview and examination survey for adults (DEGS)-design, objectives and implementation of the first data collection wave. BMC Public health, 12(1), 730.

[3] Tseng, Y. J., Ping, X. O., Liang, J. D., Yang, P. M., Huang, G. T., & Lai, F. (2015). Multiple-Time-Series Clinical Data Processing for Classification with Merging Algorithm and Statistical Measures. IEEE journal of biomedical and health informatics, 19(3), 1036-1043.

[4] Raj, P., & Surya, S. R., (2016). An Efficient Feature Selection Method for Multiple Time Series Clinical Data Classification. In International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 3(2), 15-18, 2016.

[5] Tatsumi, K., Kawachi, R., & Tanino, T. (2010, October). Nonlinear extension of multiobjective multiclass support vector machine. In Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on (pp. 1338-1343). IEEE.

[6] Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1), 51.

[7] Emary, E., Zawbaa, H. M., Ghany, K. K. A., Hassanien, A. E., & Parv, B. (2015, September). Firefly optimization algorithm for feature selection. In Proceedings of the 7th Balkan Conference on Informatics Conference (p. 26). ACM.

[8] Harrington, P. L., Rao, A., & Alfred, O. (2009, January). Classification of Multiple Time-Series via Boosting. In Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th (pp. 410-415). IEEE.

[9] Su, W. T., Ping, X. O., Tseng, Y. J., & Lai, F. (2014). Multiple Time Series Data Processing for Classification with Period Merging Algorithm. Procedia Computer Science, 37, 301-308.

[10] Priyanka, M., Rani, K. S. K., Pavithra, M., & Yamunadevi, S. (2016). The multiple time series clinical data processing with modified artificial bee colony algorithm and artificial neural network. Indian Journal of Innovations and Developments, 5(5), 1-12.

[11] Akay, B., & Karaboga, D. (2012). A modified artificial bee colony algorithm for real-parameter optimization. Information Sciences, 192, 120-142.

[12] Batal, I., Valizadegan, H., Cooper, G. F., & Hauskrecht, M. (2011, November). A pattern mining approach for classifying multivariate temporal data. In Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on (pp. 358-365). IEEE.

[13] Seethal, C. R., Panicker, J. R., & Vasudevan, V. (2016, August). Feature selection in clinical data processing for classification. In Information Science (ICIS), International Conference on (pp. 172-175). IEEE.

[14] Spiegel, S., Gaebler, J., Lommatzsch, A., De Luca, E., & Albayrak, S. (2011, August). Pattern recognition and classification for multivariate time series. In Proceedings of the fifth international workshop on knowledge discovery from sensor data (pp. 34-42). ACM.

[15] Ghassemi, M., Pimentel, M. A., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., & Feng, M. (2015, January). A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In AAAI (pp. 446-453).

[16] Bonilla, E. V., Chai, K. M., & Williams, C. (2008). Multi-task Gaussian process prediction. In Advances in neural information processing systems, 153-160.

[17] Hu, D. J. (2009). Latent dirichlet allocation for text, images, and music. University of California, San Diego. Retrieved April, 26, 2013.

[18] Zhang, D., Zuo, W., Zhang, D., & Zhang, H. (2010, August). Time series classification using support vector machine with gaussian elastic metric kernel. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 29-32). IEEE.

[19] Kanarachos, S., Mathew, J., Chroneos, A., & Fitzpatrick, M. (2015, July). Anomaly detection in time series data using a combination of wavelets, neural networks and Hilbert transform. In Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on (pp. 1-6). IEEE.

[20] Ma, J., & Plonka, G. (2010). The curvelet transform. IEEE signal processing magazine, 27(2), 118-133.