Journal of Medical Systems (2019) 43: 193
https://doi.org/10.1007/s10916-019-1265-x

IMAGE & SIGNAL PROCESSING

An Optimized HCC Recurrence Prediction Using APO Algorithm Multiple Time Series Clinical Liver Cancer Dataset

Divya R[1] · Radha P[2]

Received: 12 January 2019 / Accepted: 28 March 2019 / Published online: 22 May 2019

Abstract

The classification of recurrence and non recurrence of Hepato Cellular carcinoma (HCC) outcome after Radio Frequency Ablation therapy is a critical task. Multiple time series clinical liver cancer dataset is collected from different dataset and time interval. A merging algorithm is used to merge all attributes collected from different sources in multiple time periods. In order to preserve the originality of information, statistical measures of each attribute is calculated and considered them as additional attributes for accurate prediction. However the merged dataset is unbalanced, in which, the number of samples from HCC recurrence class is much smaller than from HCC non recurrence. The feature weighting scheme select optimal features and parameter of classifiers are sequentially obtained from multiple iterations which causes higher computation time. In this paper, an efficient sampling approach is proposed using Inverse Random under Sampling (IRUS) to overcome class imbalance issue. IRUS under sample the majority class which creates a number of distinct partitions with a boundary separated minority and majority class samples. Additionally an optimization approach is proposed using Artificial Plant Optimization (APO) algorithm to select optimal features and parameters of classifiers to improve the effectiveness and efficiency of classification. The optimization approach reduces the number of iteration and computation time for feature selection and parameter selection for classifiers which classify the recurrence and non recurrence of HCC. Classify patients with HCC and without HCC based on optimal features and parameters by Support Vector Machine (SVM) and Random Forest(RF) classifiers. Finally, the experimental results are conducted to prove the effectiveness of the proposed method over existing method in terms of accuracy, specificity, sensitivity and balanced accuracy.

Keywords Multiple time series · Hepatocellular carcinoma · Support vector machine · Random Forest · Inverse random under sampling · Multiple measurement data

Introduction

Hepatocellular carcinoma (HCC) is a common malignant tumor with immoderate mortality globally; exceptionally in East Asian nations. Hepatocellular carcinoma (HCC) is a customary malignant tumor with excessive mortality globally, above all in East Asian nations. Forty five million individuals who're suffering from a chronic Hepatitis B virus (HBV) illness and approximately 15 million folks who are stricken with the persistent Hepatitis C virus (HCV) contamination in India [9].HBV and HCV illness is viewed a foremost etiologic element in HCC. In China, 360,000 incident instances and 350,000 deaths secondary to HCC come up yearly [13].

Even though surgical resection and liver transplantation are presently the fine curative options to deal with HCC, recurrence or metastasis is relatively usual in sufferers who have had a resection and the survival expense is 30 % to 40 % at 5 years postoperatively, with almost 600,000 persons die of HCC every 12 months worldwide [1, 5]. However, the contributing reasons for the development of HCC are usually not totally understood, leading to best difficulties within the prediction of survival time and decisions related to therapy. Hence, there's a ought to identify the key attributes that affect the prediction of HCC recurrence after therapy and survival of sufferers with HCC.

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Divya R
r.divyarun@gmail.com

Radha P
radhamuthu.cbe@gmail.com

[1] Research Scholar, PG and Research Department of Computer Science, Govt Arts College(Autonomous), Coimbatore, Tamil Nadu, India

[2] Assistant Professor, PG and Research Department of Computer Science, Govt Arts College(Autonomous), Coimbatore, Tamil Nadu, India

A diversity of data types in databases is a main challengewhile combining attributes for specific data mining goals. SinceMultiple time series clinical liver cancer dataset has been collected from different database an efficient data processing techniques are required to achieve better classification results. Making use of data-processing procedures earlier than data analysis can tremendously enhance the quality of the data, curb the time required for the analysis, and make stronger the quality of analysis [8]. Temporal abstraction was among the data processing methods [16] the place data's are modified from low-level quantitative into high degree qualitative description. This system produces context aware and qualitative interval-based representations from preprocessed data. Temporal classification [3] is time based classificationtechnique where changing behaviors of data over time are considered while temporal mining of medicine dataset. Merging algorithm was proposed [19] to merge medical reports from specific sources at defined interval, and statistical measures were additionally calculated. The merged data were classified using SVM, MMSVM, RF and MMRF classifiers with different sampling rates. The class imbalance problem in the clinical dataset is occurring due to the large difference between the number of samples in recurrence and non recurrence samples. The class imbalance degrades the performance of HCC recurrence detection. So in this paper, the IRUS method is used to overcome the class imbalance problem. The optimal feature selection and optimal parameter selection for classifiers optimized using APO algorithm. The overall organization of our article is given as follows: In the section 2, previous research works are discussed. In section 3, the proposed work is discussed in detailed. In section 4, performance evaluation of the proposed work is discussed with comparison of previous work in the detailed manner. In section 5, overall conclusion of this work is given.

Literature review

In this section, various research works are analyzed for multiple time series data processing, solving class imbalance problem in datasets and about statistical measures utilized in classification process are discussed in the detailed manner.

In [12] proposed a dynamic class imbalance learning (DCIL) approach in incremental LPSVM (IncLPSVM) modeling for solving class imbalance problems. As an alternative of creating balanced data distributions through different sampling methods, weighting methods deal with the imbalanced learning quandary by means of using different weights to stability the contribution of the minority and majority classes. Nevertheless, it is noticed that for IncLPSVM training, even supposing the number of samples is even for each class, bias on decision boundary should happen due to biased data scattering.

In [17] proposed a method for class imbalance problems called Inverse Random Under Sampling (IRUS). In this method first inverse the cardinalities of majority class and minority class, then under sample the majority class, and finally construct a composite boundary between the majority class and the minority class by combining multiple designs of classifiers. This method shows better result, even in multi-label classification. In [10] proposed a computationally efficient framework for class imbalance learning ofa concept-drifting data stream called ensemble of subset online sequential extreme learning machine (ESOS-ELM). In this framework the majority classes are classified by 'm' classifiers and minority class are classified by a single classifier based on main ensemble in the current imbalanced environment. Thus the class imbalance problems and concept drift problems with or without concept drift can be tackled by this framework in both the one-by-one and chunk-by-chunk modes. The ELM-Store module is used here to store information about old concepts which is for recurring environment.

In [15] proposed a prognostic model for temporal courses which combines temporal abstractions with case-based reasoning. In temporal abstractions provides tendency about status of patient by describing a temporal sequence. In case based reasoning modify solutions of previous cases that are related to the current case. For temporal abstractions with case-based reasoning includes three steps to case-based reasoning: a state abstraction, a temporal abstraction, and a search for prototypes. In [18] proposed a nonlinear extension of MMSVM. In which kernel methods apply for nonlinear extension of MMSVM where weight vectors are represented as weighted sums of the training data and derived a single objective second-order cone programming problem to obtain a Pareto optimal solution based on eigenvalues of a kernel matrix. The advantages are high geometric margins, classifiers with the high generalization ability.

In [2] proposed a novel temporal approach to classify complex electronic health record data. Temporal pattern mining extracts most of irrelevant data for classification task in order to overcome such extraction minimal predictive temporal pattern framework is presented to discover a small set of predictive and non-spurious patterns. Additionally an efficient mining algorithm is presented which combines pattern selection with frequent pattern mining to directly mine predictive patterns. In [7] proposed a new early detection method called the Multivariate Shapelets Detection (MSD). The presented approach extends the concept of univariateshapelets to multivariate shapelets to improve the prediction accuracy. The approach utilized the information gain-based distance threshold and the weighted information-gain based utility score of a shapelet to incorporate the earliness and assigns a high utility score to the shapelet to improve the early

detection of disease pattern change. Thus the approach can improve the early classification of the multivariate time series data. The drawback in the presented approach is that all the multivariate time series shapelets have the same starting positions which cannot be possible at all situations due to the increasing number of shapelets.

In [14] proposed a unified approach to query and compute various statistical measures. To get high quality affine relationships we used AFCLST algorithm which cluster time series data it can be processed by SYMEX algorithm which compute desire affine relationships. Then the high performance of query processing is attained through SCAPE index, which index all affiine relationships.

In [11] explored a statistical mining model which predicts the scope of thedisease. This model is based on case base features where a bipartite graph is constructed with the help of patient records and features, then the refined HITS algorithm is used to find the weight of the hub and vertices from that we can predict the degree of disease possibility threshold this is used as a scale to predict the possibility of heart diseases.

In [21] presented a novel mental workload (MWL) detection framework based on a combination of unsupervised and supervised learning strategies to improve the prediction accuracy of mental workload. The drawbacks of EEG recordings are high dimensionality of the candidate, less ability to determine the MWL variations and the target class labels. In the presented approach, the locally linear embedding (LLE), support vector clustering (SVC) and support vector data description (SVDD) techniques are combined to overcome the problems of using EEG recording. The LLE technique is employed to find the low-dimensional MWL features in the high-dimensional EEG feature space. Then the SVC-SVDD hybrid framework is employed in which the SVC technique is used to find the data clusters in EEG data space and SVDD technique is used to distinguish the blurred and overlapped cluster into two classes. Thus the presented approach improves the prediction accuracy in the three class MWL temporal data classification.

In [6] two-stage Adaptive Weighted Extreme Learning Machine (AWELM) method in order to address the problems like high false-alarm rate and imbalance problem. In the first stage of AWELM utilize WELM classifier to detect the suspected fall accidents. In the second stage, refine the former detection results by using WELM classifier. It provides good balance, light weight classifiers solution to resources, high detection accuracy and low false-alarm rate. In [4] proposed a Barcelona Clinic Liver Cancer Staging System remains the most widely classification system used for HCC management guidelines. Waller et al. [20] proposed a new debates involve expansion of these criteria to create options for patients with HCC to increase overall survival.

Prediction of HCC occurrence using IRUS and artificial plant optimized classifiers

The multiple series data are collected from different sources at a particular time interval. Then the time related data are merged together by using merging algorithm and calculate the statistical measures. The class imbalance in data is processed by IRUS which constructs the complex boundary for better class separation. After the process of IRUS we obtain balanced data. From the balanced data, optimal features and parameters are selected by using Artificial Plant Optimization algorithm. The optimum features and parameters are fed as input to classifiers to classify the patients with Cirrhosis and patients without Cirrhosis. The overall representation of the proposed work is shown in Fig. 1.

Data sources

Our research collects patient information from the hospitals around TamilNadu at a time of 120 days before the radiofrequency ablation (RFA).therapy for Cirrhosis. In this research work, we included handiest people who had continuously returned for one year or greater. There are different data's are collected from different databases such as from hospitals around TamilNadu it includes the Hospital information System, Laboratory information System and Radiology information System. The Hospital information System contains 152 records with attributes like sex, age, height, weight and status of Cirrhosis. The Laboratory information System contains 152 records with attributes like alkaline phosphatase (ALP), Aspartate Transaminase (AST), Alanine Amino Transferase (ALT), Albumin, Bilirubin, Gamma-glutamyltranspeptidase (GGT) and Creatinine. The Radiology information System contains 152 records with attributes like tumor number and tumor size. There are two classes are considered here one is a recurrence of Cirrhosis and another one is a Non recurrence of Cirrhosis.

Merging algorithm

To merge multiple features of a data first we have to define the length of the period. In this research, we take time interval of 7, 14, 21, 60, 90, and 120 days. In the target event of 120 days the number of periods is either 3 with 60-day periods or 2 with 90-day periods. From the length of this period there might be more than one value for a particular feature which means there may be more than one AST test has been conducted within 60 days. In order to choose only one value for a feature we use merging algorithm [19]. Thus, in this algorithm the most recent values for a feature is considered and some important information in the data may be omitted by the merging algorithm.

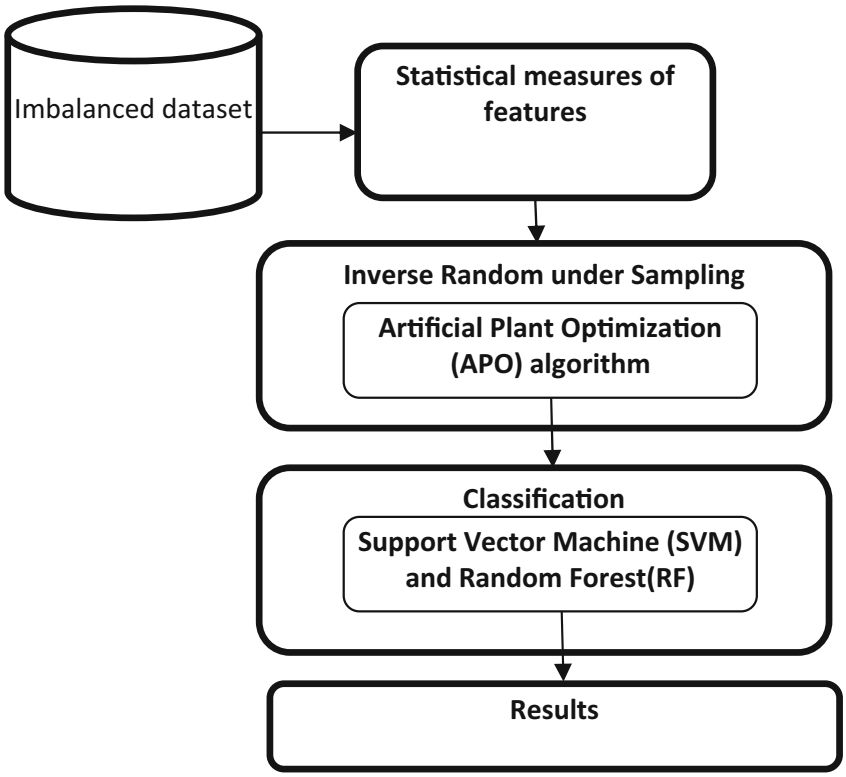

Fig. 1 Overall architecture of the proposed work

Statistical measures

Statistical measures may retain some original information that may lose after merging algorithm and it defines the data distribution in a specific period. The statistical measure was explained briefly in [19]. The maximum and minimum measurement of time related features were used to illustrate the distribution of data. Average is to find the central tendency of sample space. The variability and diversity are measured in standard deviation of time related feature. Pearson's correlation coefficient is to find the strongly pairs of features and expressed their relationship between the range 1 to −1. This is used to describe the data that have been increased or decreased over a time period and it represents the long-term movement in time-series data.

Inverse random under sampling

During classification the number of samples in one class is more than the other class in training dataset may lead to a class imbalance problem in the dataset. In our dataset, data with the occurrence of Cirrhosis are far lesser than data with the non-occurrence of Cirrhosis. It is impossible to equal the number of samples in majority class and the number of samples in the minority class by adding large number samples with the minority class, so IRUS sampling method is used to solve this problem.

For an individual training set, decision boundary is used to separate majority class from minority class and it yields one classifier design. Here we construct complex boundary by combining multiple designs through fusion which achieves better class separation. The class with huge number of samples called majority class (negative) and the class with the least number of samples called minority class (positive). In IRUS first we have to inverse the cardinalities of majority and minority class and it leads the majority class with positive samples and the minority class with negative samples. Because the positive class is more important than the negative in order to reduce the error rate in the results. To inverse the classes we would take few sample sets from the negative class with the probability of negative class is $R^2$and the samples of positive class with the probability of R. From this we may attain high true positive rate for positive class. This called under sampling, where the number samples in the negative class can be omitted. Aim of IRUS is to control the false positive rate by using the classifier bagging concept. Bagging concept is used for huge data sets it will be partitioned into anumber of subsets and each subset contains all samples of positive class and few samples of negative class and it trained with different classifiers and makes the result from the majority voted on those results. Thus the sensitivity and positive predictive value (PPV) values can be increased by IRUS technique.

Inverse Random under Sampling Algorithm

Input:
D_d= the merged records
d = number of days
$T_{s_{mino}}$: Minority patterns training set with cardinality s_{mino} in D_d
$T_{s_{majo}}$: Majority patterns training set with cardinality s_{majo}in D_d
N : Number of samples from $T_{s_{majo}}$ for each Model, N< s_{mino}
Sets : Number of classifiers, Default: 1.5× ceil(s_{majo}/N)
Test : Test sample
Ensure : Confidence score of Test, confidence(test)
confidence(test)=0
for i=1 to Sets do
$T'_{s_{majo}}$ = Randomly pick N samples without replacement from $T_{s_{majo}}$
$X_n = T'_{s_{majo}} \cup s_{mino}$
Train base classifier $B_i$using X_n samples
P = Probability of positive class assigned by $B_i$to the test sample Test
P_{NORM} = z-score normalization of P
confidence(test) = confidence(test) +P_{NORM}
end for
confidence(test) = confidence(test)/ Sets

In the above algorithm s_{mino} indicates number of minority samples and s_{majo} defines the number of majority samples. N is the number of majority samples that draws randomly for each model which controls the number of majority samples now the positive class becomes majority class so N<s_{mino}. The output of this IRUS technique is confidence score which is used to identify how much that the base classifier can be trusted or not. In this algorithm confidence score for each test sample has been tested to the majority class. The z-score normalization is to normalize classifier score with the same mean and variance. It increases the effectiveness of the classifier aggregation which is done by simple score averaging. A score y is normalized to y_{norm} by using following equation

$$y_{norm} = \frac{y - \mu(Y_i)}{\sigma(Y_i)}$$

Where $Y_i$is set of scores obtained from the training data model of i. $\mu(Y_i)$is the mean of Y_i and $\sigma(Y_i)$ is the standard deviation of Y_i.

To find the majority vote of results here we use soft voting (MEAN rule) which utilize the aggregation of generated classifiers which gives the confidence score of each test sample in the positive class. Then calculate the confidence score of other samples. Further it is used to evaluate the performance measures. Thus the class imbalance problem in multiple time series clinical data is solved by proposed IRUS method. The optimal features of balanced data are selected and optimize the cost and gamma valuesfor SVM classifiers using Artificial Plant Optimization Algorithm.

Feature selection and parameter optimization based on artificial plant optimization algorithm

To extract valuable information from the dataset the balanced data are split into 5 equal parts, then used artificial plant optimization algorithm, which is a feature selection tragedy. In this proposed work, Artificial Plant Optimization is used for feature selection and parameter optimization. By this tragedy the classification performance can be increased by population-based evolutionary algorithm by simulating the plant growing process.

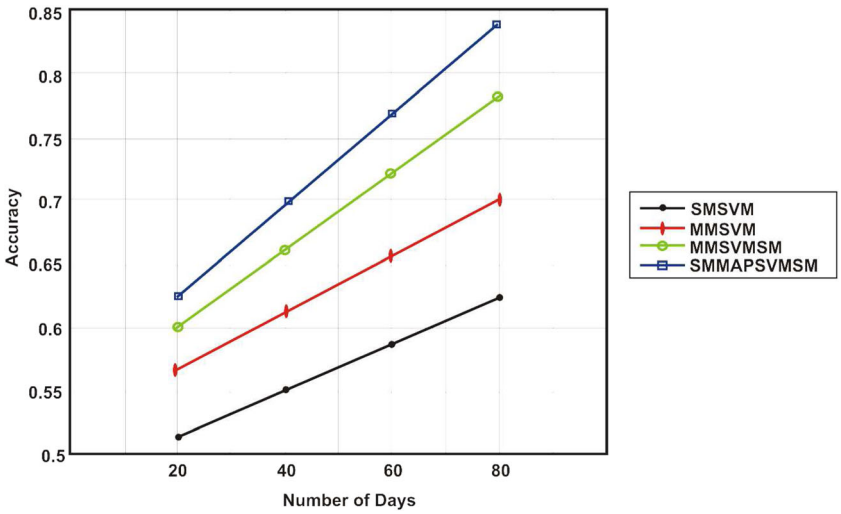

Fig. 2 Comparison of Accuracy

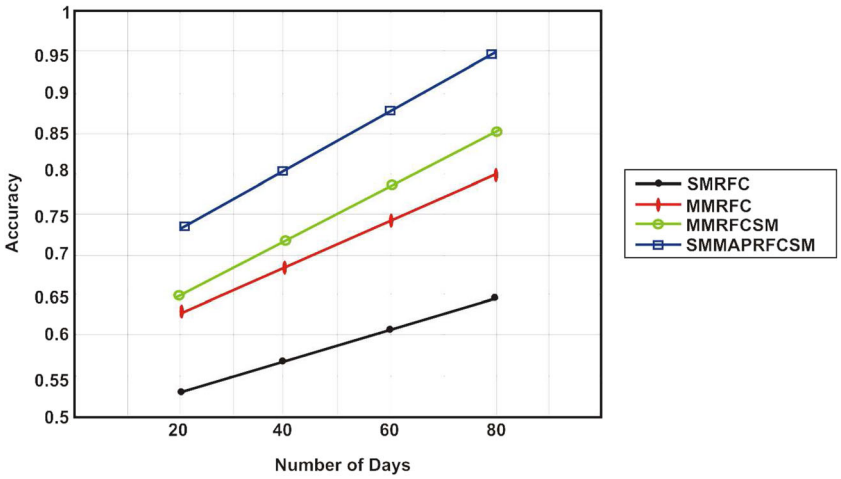

Fig. 3 Comparison of Accuracy

The light intensity (accuracy) guides the plant growing direction, and the photosynthesis provides necessary energy, the light intensity can be viewed as the fitness value which guides the search direction in the problem space. Furthermore, one point can be viewed as a branch, and the search strategy can be regarded as the growing trajectory. By considering these conditions select the optimal features and parameters for classification. These features and parameters are used intheclassification algorithm to predict HCC recurrences. There are four techniques based on SVM and random forest regression (RF). The data from majority class and minority class can be separated by hyperplane with a nonlinear mapping which converts data into higher dimension. For single measurement, the classification methods utilized LIBSVM library and RBF kernel function for the implementation and establishment of SVM classification model is implemented via the use of the MATLAB function. To increase the evaluation additionally we use RF regression. In multiple measurements, the data are collected at a particular time interval. The classification results of Random Forest can be finalized by voting mechanism. In voting mechanism the most voted classes are considered as final results. In MMSVM tool conduct cross validation and prediction from the voting mechanism results. Like as in single measurement the evaluation can be increased through MMRF in multiple measurement data. Similar to voting mechanism the averaging mechanism takes the final results as with the most voted regression results.

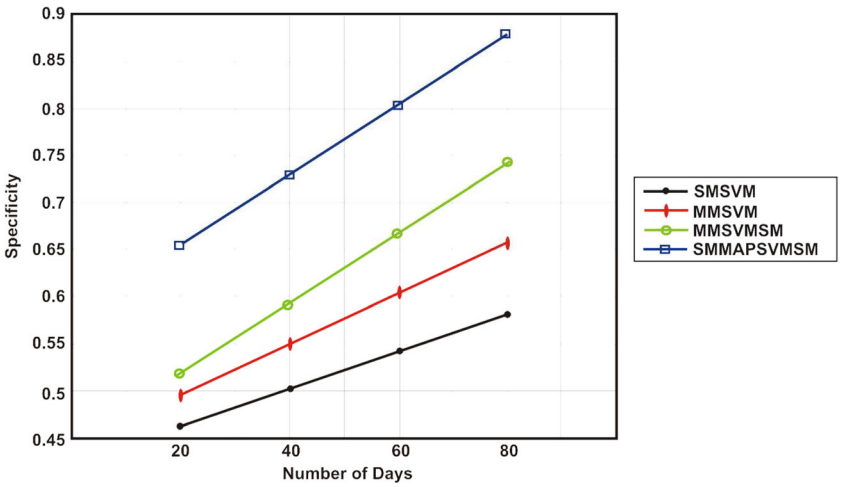

Fig. 4 Comparison of Specificity

Feature selection and parameter optimization using artificial plant optimization algorithm

Input: balanced dataset, num_branches, maxiter, iter=1, rand1, rand2
Output: Patients with HCC or Patients without HCC

Step 1: Split balanced dataset into 5 equal parts
Step 2:for each part of dataset initialize number of branches
Step 3: Each branch randomly choose features and value of Cost C and gamma as its position
Step 4:while(iter<maxiter)
Step 5:for i=1:num_branches
do
Step 6: Calculate light intensity $Uf_i(x_i)$ is given as follows for both features $Uf_{if}(x_i)$ and parameters$Uf_{ip}(x_i)$

$$Uf_{if}(x_i) = \frac{f_{worstf}(t) - f_{if}(t)}{f_{worstf}(t) - f_{bestf}(t)}$$
$$Uf_{ip}(x_i) = \frac{f_{worstp}(t) - f_{if}(t)}{f_{worstp}(t) - f_{bestp}(t)}$$

Step 7: Compute photosynthetic rate p_i for both features $p_{if}(t)$and parameters$p_{ip}(t)$

$$p_{if}(t) = \frac{\alpha Uf_{if}(x_i) P_{max}}{\alpha Uf_{if}(x_i) + P_{max}} - R_d$$
$$p_{ip}(t) = \frac{\alpha Uf_{ip}(x_i) P_{max}}{\alpha Uf_{ip}(x_i) + P_{max}} - R_d$$

end do
Step 8:fori=1:num_branches
do
Step 9: Sorting the accuracy values of all branches, the better half of the best populations are taken as growing motion branches, as well as other branches are maturing motion branches
Step 10: Update feature and parameter for growing motion branch using following equation

$$\mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}) + \left(\mathrm{x}_{\mathrm{bestf}}^{\mathrm{k}}(\mathrm{t}) - \mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t})\right).growth.r + C.rand$$
$$\mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}) + \left(\mathrm{x}_{\mathrm{bestp}}^{\mathrm{k}}(\mathrm{t}) - \mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t})\right).growth.r + C.rand$$

Step 11: Update feature and parameter for maturing motion branch using following equation

$$\mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}) + growth.r.D_i^k$$
$$\mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}) + growth.r.D_i^k$$

Step 12: end do
Step 13: If rand1<rate

$$\mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}) + \left(\mathrm{x}_{\mathrm{bestf}}^{\mathrm{k}}(\mathrm{t}) - \mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t})\right).growth.r$$
$$\mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}) + \left(\mathrm{x}_{\mathrm{bestp}}^{\mathrm{k}}(\mathrm{t}) - \mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t})\right).growth.r$$

Else if rand2<1/n

$$\mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{if}}^{\mathrm{k}}(\mathrm{t}) + growth.r$$
$$\mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{ip}}^{\mathrm{k}}(\mathrm{t}) + growth.r$$

Else

$$\mathrm{x}_{\mathrm{bestf}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{bestf}}^{\mathrm{k}}(\mathrm{t})$$
$$\mathrm{x}_{\mathrm{bestp}}^{\mathrm{k}}(\mathrm{t}+1) = \mathrm{x}_{\mathrm{bestp}}^{\mathrm{k}}(\mathrm{t})$$

Step 14: iter=iter+1
Step15: End while
Step 16: Classify patients with HCC and without HCC based on optimal features and parameters by SVM and RF classifiers

In the above feature selection, parameter optimization and classification algorithm num_branches denotes the number of branches in artificial plant, maxiter represents the maximum iteration, $Uf_{if}(x_i)$ and $Uf_{ip}(x_i)$ represents the light intensity (accuracy) of feature selection and parameter optimization respectively, $f_{worstf}(t)$ and $f_{bestf}(t)$represents the worst and best light intensities at time t of feature selection respectively, $f_{worstp}(t)$ and $f_{bestp}(t)$represents the worst and best light intensities at time t of parameter optimization respectively $f_{if}(t)$ and $f_{ip}(t)$refers the light intensity of branch i for feature selection and parameter optimization respectively, $p_{if}(t)$ and $p_{ip}(t)$ represents the photosynthetic rate of both feature selection and parameter optimization respectively, α denotes the initial quantum efficiency, P_{max} denotes the maximum net photosynthetic rate, R_d denotes the dark respiratory rate, growth is one parameter, r is one random number sampled with uniformly distribution, $x^k_{bestf}(t)$ and $x^k_{bestp}(t)$ Denotes the highest accuracy for feature selection and parameter optimization respectively, and rand1 and rand2 are the random numbers.

Experimental results

In this section, the data's are tested with IRUS technique. The performance of multiple time series classification can be tested in terms of specificity, accuracy, balanced accuracy and sensitivity. For experimental purposesdata are collected around TamilNadulocationat a time of 120 days, which includestheHospital information System, Laboratory information System and Radiology information System. The Hospital information System contains 152 records with attributes like sex, age, height, weight and status of Cirrhosis. The Laboratory information System contains 152 records with attributes like alkaline phosphatase (ALP), Aspartate Transaminase (AST), Alanine Amino Transferase (ALT), Albumin, Bilirubin, Gamma-glutamyltranspeptidase (GGT) and Creatinine. The Radiology information System contains 152 records with attributes like tumor number and tumor size.

Accuracy

Accuracy is defined as the proportion of addition of true positive and true negative among the addition of true positive, true negative, false positive and false negative.

Accuracy can be calculated by the formula given below:

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

Figure2, shows the comparison of accuracy with different time interval for SVM classifier. X axis represents the number of days and Y axis represents the accuracy value. There are different methods like Single Measurement Support Vector Machine (SMSVM), Multiple Measurement Support Vector Machine (MMSVM), Multiple Measurement Support Vector Machine with Statistical Measure (MMSVMSM) and Sampling based Multiple MeasurementArtificial Plant optimized Support Vector Machine with Statistical Measure (SMMAPSVMSM) are compared. From the graph it is proved that the SMMAPSVMSM has high accuracy than the other methods.

Figure 3, shows the comparison of accuracy with different time interval for Random Forest Classifier (RFC). X axis represents the number of days and Y axis represents the accuracy value. There are different methods like Single Measurement Random Forest Classifier (SMRFC), Multiple Measurement Random Forest Classifier (MMRFC), Multiple Measurement Random Forest Classifier with Statistical Measure (MMRFCSM) and Sampling based Multiple Measurement Artificial Plant optimized Random Forest Classifier with Statistical Measure (SMMAPRFCSM) are compared. From the graph it is proved that the SMMAPRFCSM has high accuracy than the other methods.

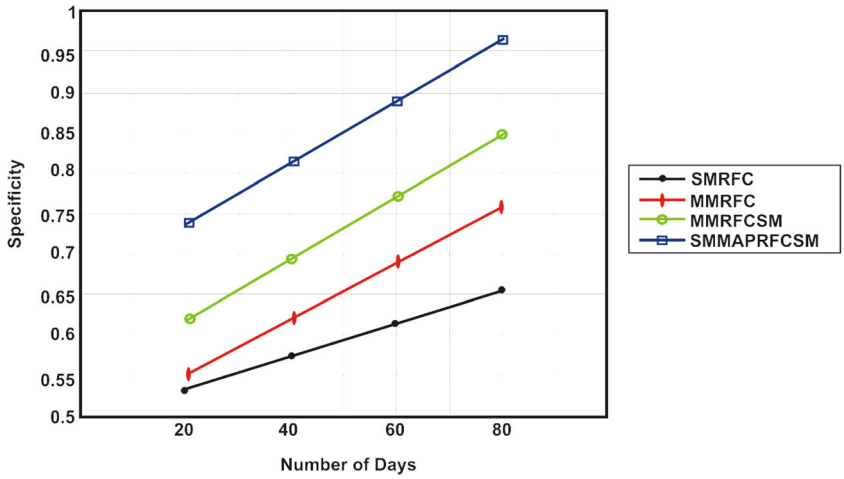

Fig. 5 Comparison of Specificity

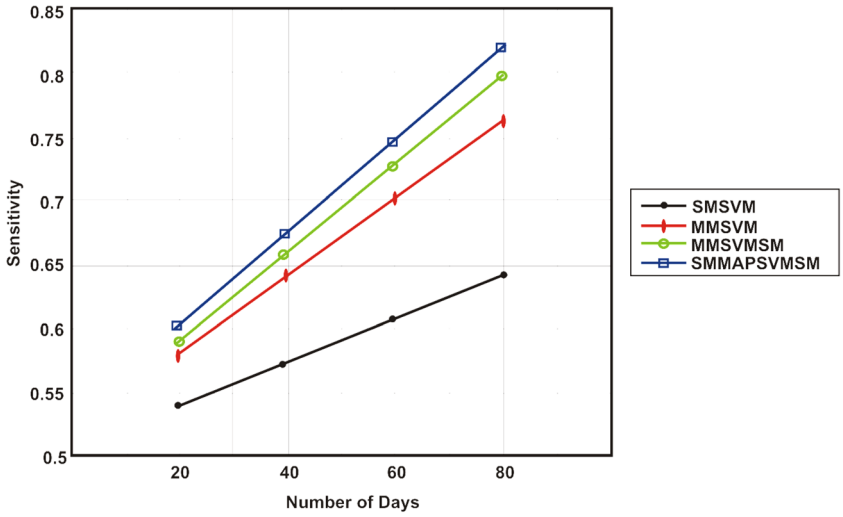

Fig. 6 Comparison of Sensitivity

Specificity

Specificity is calculated by True negative is divided by addition of True negative and false positive. It has the ability to identify the patients without diseases. Specificity is calculated as follows:

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

Fig. 4, shows the comparison of specificity with different time interval for SVM classifier. X axis represents the number of days and Y axis represents the specificity value. There are different methods like SMSVM,MMSVM, MMSVMSM andSMMAPSVMSM are compared. From the graph it is proved that the SMMAPSVMSM has high specificity than the other methods.

Fig 5, shows the comparison of specificity with different time interval for Random Forest Classifier (RFC). X axis represents the number of days and Y axis represents the specificity value. There are different methods like SMSVM, MMSVM, MMSVMSM and SMMAPSVMSM are compared. From the graph it is proved that the SMMAPRFCSM has high specificity than the other methods.

Sensitivity

In classification task, sensitivity of a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class. It measures the exactness or quality of results and it has the ability to identify the patient with diseases. The data's with more relevant results from high sensitivity values. It is evaluated by using following formula:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

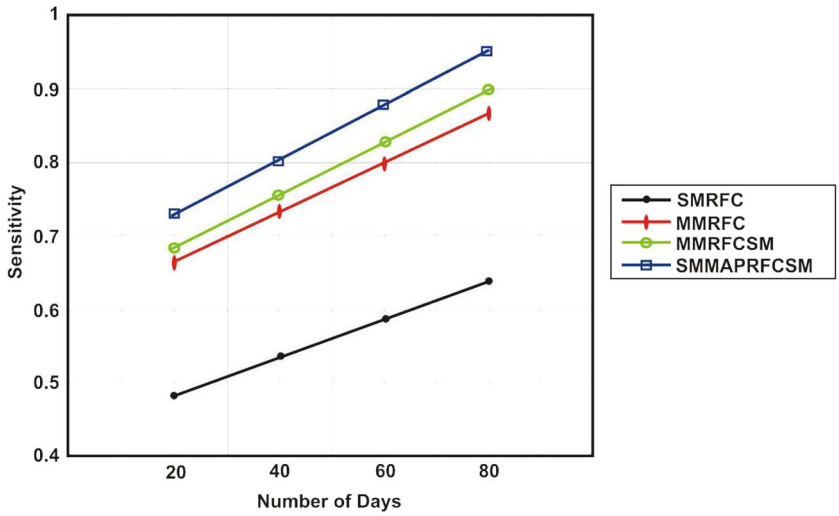

Fig. 7 Comparison of Sensitivity

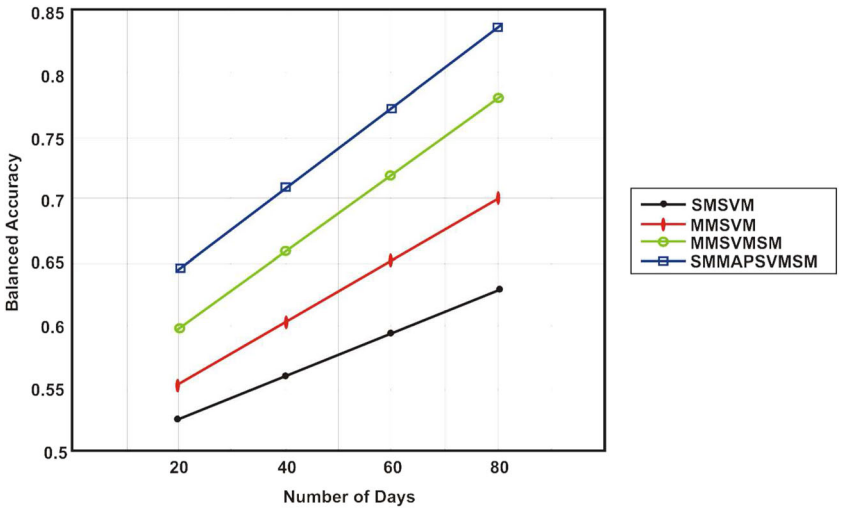

Fig. 8 Comparison of Balanced accuracy

Figure 6, shows the comparison of sensitivity with different time interval for SVM classifier. X axis represents the number of days and Y axis represents the sensitivity value. There are different methods like SMSVM, MMSVM, MMSVMSM and SMMAPSVMSM are compared. From the graph it is proved that the SMMAPSVMSM has high sensitivity than the other methods.

Figure 7, shows the comparison of sensitivity with different time interval for Random Forest Classifier (RFC). X axis represents the number of days and Y axis represents the sensitivity value. There are different methods like SMSVM, MMSVM, MMSVMSM and SMMAPSVMSM are compared. From the graph it is proved that the SMMAPRFCSM has high sensitivity than the other methods.

Balanced accuracy

Balanced Accuracy is defined as average of specificity and sensitivity values. It is calculated by using following formula:

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

Figure 8, shows the comparison of balanced accuracy with a different time interval for the SVM classifier. X axis represents the number of days and Y axis represents the balanced accuracy value. There are different methods like SMSVM, MMSVM, MMSVMSM and SMMAPSVMSM are compared. From the graph it is proved that the SMMAPSVMSM has high balanced accuracy than the other methods.

Figure 9, shows the comparison of balanced accuracy with different time interval for Random Forest Classifier (RFC). X axis represents the number of days and Y axis represents the balanced accuracy value. There are different methods like SMSVM, MMSVM, MMSVMSM and SMMAPSVMSM are compared. From the graph it is proved that the SMMAPRFCSM has high balanced accuracy than the other methods.

The overall results of the existing and proposed methods are tabulated in Table 1 with different measure are true positive, true negative, false positive, false negative, accuracy, specificity, sensitivity and balanced accuracy

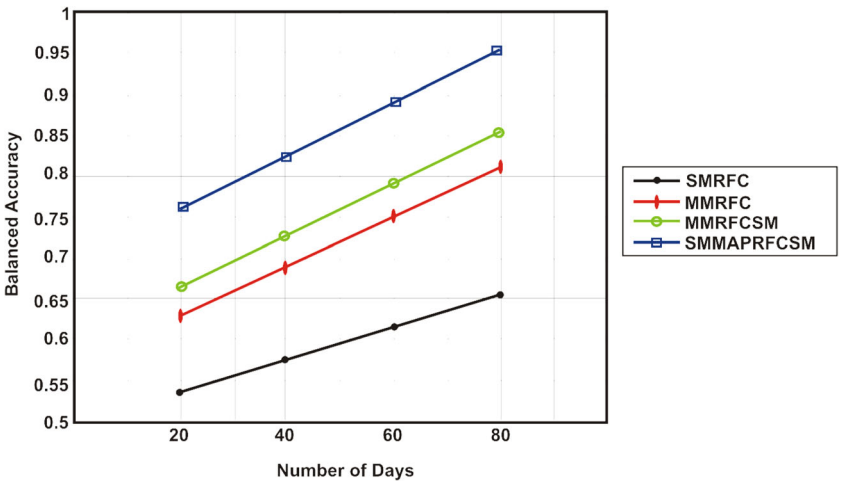

Fig. 9 Comparison of Balanced accuracy

Table 1 Overall Results

Methods	True Positive	False Positive	True Negative	False Negative	Accuracy	Specificity	Sensitivity	Balanced Accuracy
SMSVM	43	33	52	24	0.625	0.612	0.642	0.627
MMSVM	61	26	46	19	0.703	0.639	0.763	0.701
MMSVMSM	64	19	55	14	0.783	0.743	0.821	0.782
SMMAPSVMSM	55	10	73	14	0.842	0.880	0.797	0.839
SMRFC	56	21	43	32	0.651	0.672	0.637	0.655
MMRFC	64	22	58	8	0.802	0.725	0.899	0.812
MMRFCSM	72	11	58	11	0.856	0.841	0.867	0.854
SMMAPRFCSM	78	3	67	4	0.954	0.958	0.951	0.955

From the experimental results it proved that the proposed SMMAPRFCSM method has high accuracy, sensitivity, specificity and balanced accuracy than the other methods.

Conclusion

In this section, we conclude that the proposed system has better performance by overcoming the class imbalance problem using IRUS technique and this system defines the detecting ability of HCC uses AIC values. In IRUS the cardinalities of classes are inversed and under sample the majority class, thus the class imbalance problem can be avoided and it increase the sensitivity and PPV values by using bagging. In addition to the feature selection and parameter optimization improves the accuracy of HCC prediction. Thus, the experimental result proved that the proposed system is better than the existing system in terms of accuracy, balanced accuracy, specificity and sensitivity. The proposed work is difficult to apply for real time applications which is kept as scope of the future work.

Compliance with ethical standards

Disclosure of potential conflicts of interest The authors have no conflict of interests. We approave there is no conflicts of interest among all the authors and Co-Authors.

Research involving human participants and/or animals This Manuscriptstrictly does not involve any human or animals participants performed by any of the authors.

Informed consent The dataset is taken only from online databases. So, it is not required.

References

1. Aravalli, R. N., Steer, C. J., and Cressman, E. N., Molecular mechanisms of hepatocellular carcinoma. *Hepatol* 48(6):2047–2063, 2008. https://doi.org/10.1002/hep.22580.
2. Batal, I., Valizadegan, H., Cooper, G. F., and Hauskrecht, M., A pattern mining approach for classifying multivariate temporal data. *Bioinform Biomed*:358–365, 2011. https://doi.org/10.1109/BIBM.2011.39.
3. Campos, M., Palma, J., and Marín, R., Temporal data mining with temporal constraints. *Artifintell Med*:67–76, 2007.
4. Dimitroulis, D., Damaskos, C., Valsami, S., Davakis, S., Garmpis, N., Spartalis, E., Athanasiou, A., Moris, D., Sakellariou, S., Kykalos, S., and Tsourouflis, G., From diagnosis to treatment of hepatocellular carcinoma: An epidemic problem for both developed and developing world. *World J. Gastroenterol.* 23(29):5282, 2017.
5. El-Serag, H. B., and Rudolph, K. L., Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastronterol* 132(7): 2557–2576, 2007. https://doi.org/10.1053/j.gastro.2007.04.061.
6. Gao, X., Chen, Z., Tang, S., Zhang, Y., and Li, J., Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing* 173:1927–1935, 2016. https://doi.org/10.1016/j.neucom.2015.09.064.
7. Ghalwash, M. F., and Obradovic, Z., Early classification of multivariate temporal observations by extraction of interpretable shapelets. *BMC Bioinform* 13(1):195, 2012. https://doi.org/10.1186/1471-2105-13-195.
8. Han, J., Pei, J., Kamber, M. (2011) Data mining: concepts and techniques. Elsevier.
9. Kar, P., Risk factors for hepatocellular carcinoma in India. *J. Clin. Densitom.* 4:S34–S42, 2014. https://doi.org/10.1016/j.jceh.2014.02.155.
10. Mirza, B., Lin, Z., and Liu, N., Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift. *Neurocomputing* 149:316–329, 2015. https://doi.org/10.1016/j.neucom.2014.03.075.
11. Nagavelli, R., and Rao, C. G., Degree of disease possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining. *In Recent AdvInnovEng*:1–6, 2014. https://doi.org/10.1109/ICRAIE.2014.6909265.
12. Pang, S., Zhu, L., Chen, G., Sarrafzadeh, A., Ban, T., and Inoue, D., Dynamic class imbalance learning for incremental LPSVM. *Neural Netw.* 44:87–100, 2013. https://doi.org/10.1016/j.neunet.2013.02.007.
13. Poon, D., Anderson, B. O., Chen, L. T., Tanaka, K., Lau, W. Y., Van Cutsem, E., and Khin, M. W., Management of hepatocellular carcinoma in Asia: Consensus statement from the Asian oncology summit 2009. *Lancetoncol* 10(11):1111–1118, 2009. https://doi.org/10.1016/S1470-2045(09)70241-4.
14. Sathe, S., Aberer, K. (2013) AFFINITY: Efficiently querying statistical measures on time-series data. Data Eng IEEE 29th IntConf: 841–852. doi: https://doi.org/10.1109/ICDE.2013.6544879.
15. Schmidt, R., and Gierl, L., A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning. *Int. J.*

Med. Inform. 74(2):307–315, 2005. https://doi.org/10.3233/978-1-60750-939-4-571.

16. Stacey, M., and McGregor, C., Temporal abstraction in intelligent clinical data analysis: A survey. *Artifintell Med* 39(1):1–24, 2007. https://doi.org/10.1016/j.artmed.2006.08.002.
17. Tahir, M. A., Kittler, J., and Yan, F., Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recogn.* 45(10):3738–3750, 2012.
18. Tatsumi, K., Kawachi, R., and Tanino, T., Nonlinear extension of multiobjective multiclass support vector machine. *Syst Man Cyber*: 1338–1343, 2010. https://doi.org/10.1109/IJCNN.2011.6033411.
19. Tseng, Y. J., Ping, X. O., Liang, J. D., Yang, P. M., Huang, G. T., and Lai, F., Multiple-time-series clinical data processing for classification with merging algorithm and statistical measures. *IEEE J Biomed Health Inform* 19(3):1036–1043, 2015. https://doi.org/10.1109/JBHI.2014.2357719.
20. Waller, L. P., Deshpande, V., and Pyrsopoulos, N., Hepatocellular carcinoma: A comprehensive review. *World J. Hepatol.* 7(26):2648, 2015.
21. Yin, Z., and Zhang, J., Identification of temporal variations in mental workload using locally-linear-embedding-based EEG feature reduction and support-vector-machine-based clustering and classification techniques. *Comput. Methods Prog. Biomed.* 115(3):119–134, 2014. https://doi.org/10.1016/j.cmpb.2014.04.011.