International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

An Efficient Feature Selection Method for Multiple Time Series Clinical Data Classification

[1] Priyanka Raj [2], Surya S. R
[1] PG Scholar, [2] Assistant Professor,
[1] Department of Computer Science [2] Department of Information Technology
College of Engineering, Perumon
[1] rajplr12@gmail.com, [2] sooryasr07@gmail.com

Abstract- **Patient's condition description consists of combination and changes of clinical measures. Conventional data processing methods and classification algorithms may reduce the prediction performance of clinical data.Inorder to improve the accuracy of clinical data outcome prediction by using feature selection method with multiple measurement support vector machine(MMSVM) classification algorithm is proposed. Most popular primary liver cancer is hepatocellular carcinoma (HCC). It stands in the fifth position in the world considering the tumour ranking. HCC can be treated by using Radiofrequency ablation (RFA). Recurrence prediction of hepatocellular carcinoma (HCC) after RFA treatment is an important task. The proposed method uses Binary krill herd method as the feature selection method for classification of clinical data.This method can be used for prediction of Hepatocellular Carcinoma (HCC) recurrence. After data processing, multiple measurement support vector machine(MMSVM) is used as classification method to predict HCC recurrence.The method classify data into two classes-1) HCC recurrence and 2) no evidence of recurrence of HCC.The performance accuracy of HCC recurrence prediction was significantly improved by using the feature selection method.**

Index Terms— **Classification, Hepatocellular Carcinoma (HCC), Multiple measurement support vector machine (MMSVM), Radiofrequency ablation (RFA)**

I. INTRODUCTION

Data processing [1] depends on the type of data used. Two varieties of data:1) time series data 2) cross-sectional data. Time series data are data from a unit (or a group of units) observed in several successive periods. Cross–sectional data are data from units observed at the same time or in the same time period. Example for time series data includes time sequence of blood pressure and blood glucose. For cross sectional data, consider a routine examination of health which includes a number of physical examinations, such as weight, vision, height, breathing rate. Here the clinical data processing method includes both time series and cross sectional data.

Data pre-processing [1] techniques can be used before data analysis which decreases the analysis time and increases prediction performance. Data pre-processing techniques includes the following: 1) data cleaning 2) data integration 3) data transformation 4) data reduction. Data cleaning [2] is the process of removing incomplete data. Data integration [3] is used to combine data from disparate sources into meaningful and valuable information. Data transformation converts data into appropriate forms for mining. Data reduction [4] is a method for reducing the size of data. Temporal abstraction (TA) is the process of transforming lower level quantitative into higher level quantitative [5].Multiple measurement clinical data are merged using various time periods and they are transformed based on TA.

For many years, RFA has been the widely used method of treatment for HCC. It has many advantages over other therapy techniques, such as:1) more effective destruction of cancer cells 2) fewer complication 3) reduced risk of complications such as infection.

II. LITERATURE REVIEW

Patients may undergo many laboratory and clinical tests within a defined time period. A major problem is that there may be multiple values for a feature in one period. In 2014, Wei-Ti Su [6] introduced a method for multiple feature time period merging. It takes every 30 days multiple days' features for merging. Time periods used are 7, 14, 21,30,60,90,120 days. An example for time period merging is shown in Fig: 1. only one value of a feature which is closest to treatment date is taken when there are many values for that feature.

Selecting relevant features for classification process is a major task. In 2001, Leo Breiman [7], proposed a feature selection method. Random forest is a combination of tree predictors. It develops decision tree based on random selection of data and variables and also provides the class of dependent variables. These trees combine to form random forest. It selects features based on random with replacement method and groups them to form random space. A scoring function is used for assigning accuracy for features in

random space and search method is used to obtain top ranked features.

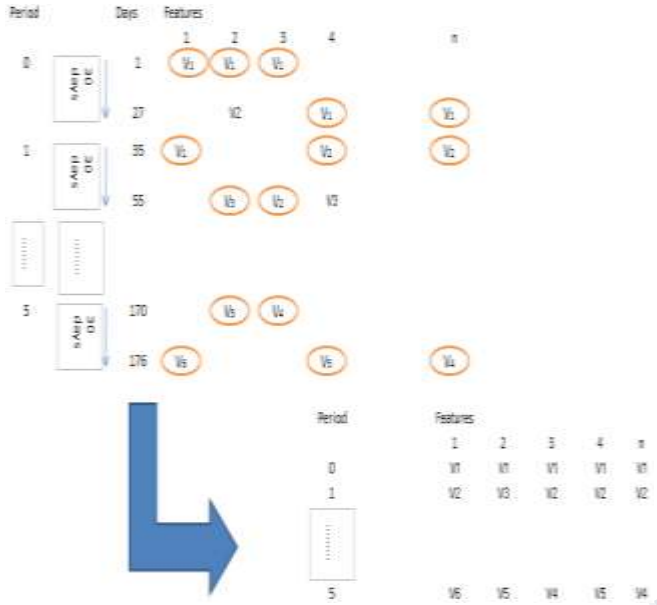

Fig:1.Time period merging example

In 1995,C.Cortes [8],proposed a method for classification. Support vector machine (SVM) is the classification method. Nonlinear mapping of data into a higher dimension is done in this method. It works by finding an optimal margin hyper plane for making the data set into two different classes the optimal margin is obtained by using the support vectors.SVM model is represented in Fig: 2.

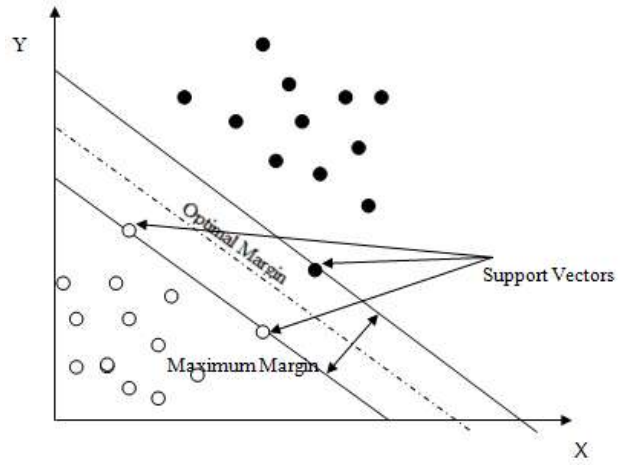

Fig 2: Support Vector Machine

LIBSVM is the library used by SVM for classification. Advantages of using the SVM are avoids over-fitting, can model complex linear decision boundaries, highly efficient and accurate.

III. PROPOSED METHOD

The proposed system aims to predict the hepatocellular carcinoma (HCC) reccurence from multiple time-series data collected. The basic idea is to merge all features that occur within a defined time period. They are then split into training set and test set.The next major step is the feature selection. The improved method for feature selection is the binary krill herd method, which chooses the most representative features. It then undergoes cross validation with the classification algorithm[9] resulting in a predictive model. The SVM(Support Vector Machine) is a data-mining method that constructs a prediction model. The system architecture is shown in fig:3

Data Set: Mainly 16 clinical features are taken from laboratory information system (LIS)database, radiology information system (RIS)database and hospital information system(HIS) database for the model establishment stage. The ten laboratory tests selected were aspartate transaminase (AST),AlphaFetoprotein,alanine aminotransferase (ALT), hepatitisCvirus(HCV) total bilirubin, platelet,albumin, creatinine, prothrombin and hepatitisBvirus (HBV). The six other features were tumor size and tumor number, extracted from radiology reports or ultrasound reports in RIS; stage of Barcelona- Clinic Liver Cancer (BCLC) classification and cirrhosis status, extracted from narrative clinical reports in HIS and age and gender, collected from the demographic database within HIS.Data set is divided into single measurements and multiple measurements data.Single measurement data combines all the values of a feature to produce a single value whereas multiple measurement data takes multiple values of a feature separately.

***Feature selection*:** Feature selection, also called as variable selection, attribute selection or variable subset selection, is the process of choosing a subset of relevant features (variables) for use in model construction.Feature selection method is used for three reasons:

- To simplify the models to make them easier to interpret
- To Reduce training times
- Enhanced generalisation by reducing overfitting

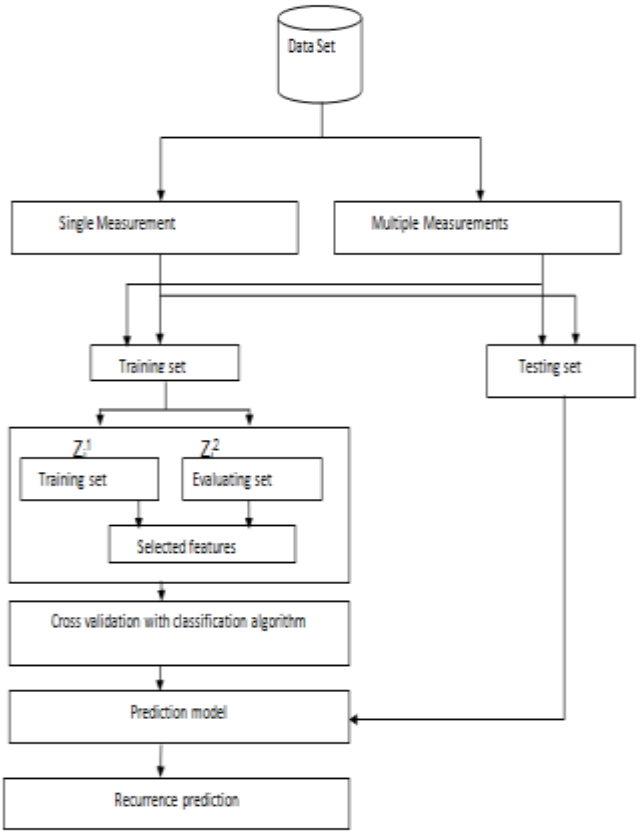

Fig 3: System Architecture

The central premise when using a feature selection technique is that the data may contains many features which are either redundant or irrelevant, and can thus be removed without causing too much information loss. Redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of some other relevant feature with which it is strongly correlated.The feature selection method used in this paper is binary krill herd [10] method.

The problem is to select or not a given feature, a solution binary vector is employed, where 1 represents whether a feature will be choosed to compose the new dataset, and 0 otherwise.

A. BKH - Binary Krill Herd Algorithm

Input: Training set Z_1,Testing set Z_2, krills n, dimension d, iterations T.
Output: Selected feature set.
Auxiliaries: Fitness vector f with size m and variables acc, maxfit, globalfit and maxindex.

1. For each krill n_i,where i=(1,..,m) do
2. For each dimension j, where j=(1,...,d)do
3. a_i^j is randomly selected feature
4. $f_i \leftarrow -\infty$
5. globalfit $\leftarrow -\infty$
6. For each iteration t(t=1,...,T)do
   For each krill n_i,where i=(1,..,m) do
   create Z_1^* and Z_2^* from Z_1 and Z_2 respectively
   Train SVM over Z_1^*,evaluate its over Z_2^*andstores the accuracy in *acc*
   If *acc* > f_1
   $f_i \leftarrow acc$
   [maxfit, maxindex] ←max(f)
   If (maxfit>globalfit)then
   globalfit←maxfit
   For each krill n_i,where i=(1,..,m) do
   For each dimension j, where j=(1,...,d)do
   If(threshold < 1/1 + e $^{a_i^{j(t)}}$) then
   $a_i^j \leftarrow 1$;
   print globalfit
   else
   $a_i^j \leftarrow 0$;

Partition the data set randomly into m folds $Z=Z_1 \upsilon Z_2 \upsilon,... \upsilon Z_m$.For each Z_i ,train a given instance over SVM classifier over Z_i^1 subset of Z_i and an evaluation set and an evaluation set $Z_i^2 \leftarrow Z_i / Z_i^1$ is then classified in order to compute a fitness function which will guide the stochastic optimization algorithm to select the most representative set of features.

Classification method: SVM is the technique used for classification process. Support vector machines are models for supervised learning which are associated with learning algorithms that recognize data and patterns, that can be used for classification .It uses nonlinear mapping to transform the data into a higher dimension. It is a nonlinear mapping to a higher dimension, data from two classes are separated by a hyperplane.The LIBSVM library was used for implementing the SVM classification method. The most popular kernel function, RBF, was used for SVM model establishment.

***Reccurence Prediction*:**The SVM method builds a predictive model for a binary class.It can predict whether a patient is HCC recurrent or no evidence of HCC reurrence.This helps the patient to take adequate treatment against HCC.

IV. EXPERIMENTAL RESULTS

For our experiment,we used a clinical data set of HCC patients containing 16 features.Here we perform the feature selection using two methods:1)random forest 2)Krill herd algorithm.A graph is plotted between number of features aganist accuarcy.It is observed that feature slection performed with krill herd algorithm produces better accuarcy against the random forest algorithm. The graph is shown in fig 4.

V. CONCLUSION

The method is used for predicting recurrence of hepatocellular carcinoma for patients, those who have returned within one year after Radiofrequency ablation (RFA) are mentioned. Multiple time series data are merged using merging algorithm. By using binary krill herd method improves the accuracy of feature selection. A binary version of the continuous-valued Krill Herd to handle with feature selection problems is discussed here.It decides whether to select a feature or not.It is a fastest technique and the one that has selected the fewer number of features.The technique can be used with any other supervised classification technique. Support vector machine method was not applied in the past for recurrence prediction. The system uses multiple measurement support vector machine for the classification process. Method can also used for classification of meteorological data and financial data .

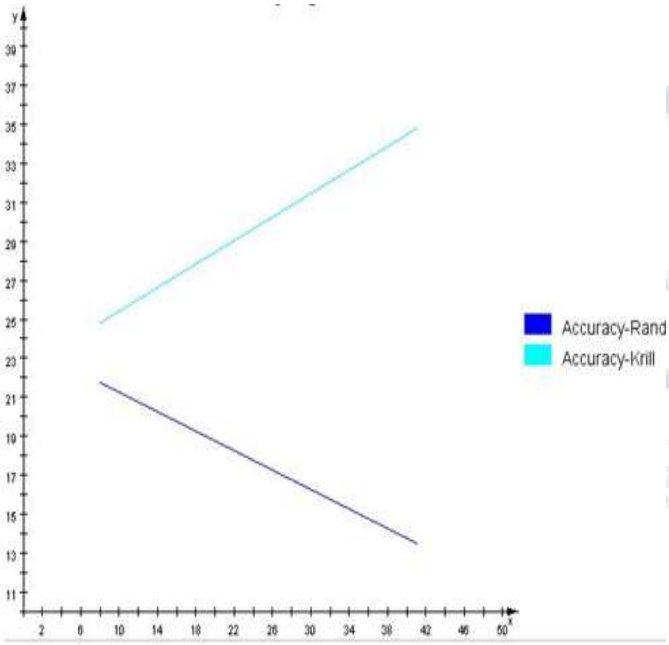

Fig 4:Graph representing accuracy against feature selection between random forest and krill algorithm. x-axis represents number of features and y-axis represents accuracy

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. SanFrancisco, CA, USA: Morgan Kaufmann, 2006.

[2] M.A.Hern´andez and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," Data Mining Knowl. Discovery, vol. 2, no. 1, pp. 9–37, 1998K.

[3] M. Lenzerini, "Data integration: A theoretical perspective," in Proc. 21st ACMSIGMOD-SIGACT-SIGART Symp. Principles Database Syst., Madison, WI, USA, 2002, pp. 233–246.

[4] A. S. C. Ehrenberg, Data Reduction: Analysing and Interpreting Statistical Data. New York, NY, USA: Wiley, 1975.

[5] M. Stacey and C. McGregor, "Temporal abstraction in intelligent clinical data analysis: A survey," Artif. Intell. Med., vol. 39, no. 1, pp. 1–24, 2007.

[6] Wei-Ti Su ,Xiao-Ou Ping, Yi-Ju Tseng, Feipei Lai," Multiple Time Series Data Processing for Classification with Period Merging Algorithm", Procedia Computer Science 37 (2014) 301 – 308.

[7] L. Breiman, "Random forests," Mach. Learning, vol. 45, no. 1, pp. 5–32,2001.

[8] Cortes and V. Vapnik, "Support-vector networks," Mach. Learning vol. 20, no. 3, pp. 273–297, 1995.

[9] Yi-Ju Tseng, Xiao-Ou Ping, Ja-Der Liang "Multiple-Time-Series Clinical Data Processing for Classification With Merging Algorithm and Statistical Measures" Ieee Journal Of Biomedical And Health Informatics, Vol. 19, No. 3, May 2015.

[10] Douglas Rodrigues, Luis A. M. Pereira, Jo'ao P. Papa,"A Binary Krill Herd Approach for Feature Selection",2014 22nd International Conference on Pattern Recognition.