International Journal of Computer Applications (0975 - 8887)
Volume 133 - No.13, January 2016

Multiple-Time-Series Clinical Data Processing for Classification: A Review

Priyanka Raj
PG scholar
Department of Computer Science
College of Engineering Perumon,Kerala, India

Surya S. R.
Assistant Professor
Department of Information Technology
College Of Engineering Perumon, Kerala, India

ABSTRACT

Data mining is a multidisciplinary subfield of computer science. It is used in various fields such as medical research, financial, telecommunication, scientific application. Classification is a method used in data mining. Data mining includes wide varieties of data such as clinical, scientific, biological, remote sensing etc. Clinical data can be used for clinical data mining. Clinical data mining helps the clinicians for diagnosis, therapy and prognosis of various diseases. Most popular primary liver cancer is hepatocellular carcinoma (HCC). It is the fifth most common tumour in the world. HCC can be treated by using Radiofrequency ablation (RFA). Recurrence prediction of hepatocellular carcinoma (HCC) after RFA treatment is an important task. This problem can be solved by using a classification technique that classifies persons into two groups: 1) HCC recurrence and 2) no evidence of recurrence of HCC. In this paper a review is being carried out in various techniques used in HCC recurrence prediction are discussed.

General Terms

Data mining, classification

Keywords

Clinical data mining, Hepatocellular Carcinoma (HCC), radiofrequency ablation (RFA)

1. 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. Crosssectional 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 [2] 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 [3] is the process of removing incomplete data. Data integration [4] is used to combine data from disparate sources into meaningful and valuable information. Data transformation converts data into appropriate forms for mining. Data reduction [5] is the process of reducing the size of data. Temporal abstraction (TA) is the process of transforming lower level quantitative into higher level quantitative [6].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.

2. LITERATURE REVIEW

Rainer Schmidt in 2003 [6],developed a method that combines temporal abstractions(TA) with case based reasoning(CBR) for the prediction of temporal courses .Here the method is applied for two applications. One deals with the patients in the intensive care units for kidney prognosis, in order to warn them against threatening kidney failures. The other one deals with providing earlier warnings against infectious diseases (influenza) approach. Temporal abstraction converts temporal sequences of values into a more abstract form. For example for a patient having blood glucose the measured parameters can be abstracted into states (eg: low, medium, high). Case based reasoning is the process of solving new problems based on the solution of similar problems. Prognostic method combines TA and CBR.

One of the main problems is to understand patient's status from large amount of reports. In 2003, Xiao-Ou Ping [7], proposed an efficient method to analyse the condition of patients from different narrative reports. For the case of liver cancer the clinical factors can be extracted from different types of narrative clinical reports, like admission notes, radiology reports, operation notes, ultrasound reports, pathology reports and discharge summaries. This method contains two steps mainly.1) development of information extraction module 2) developing a rule based classifier. The overview of the method is shown in Fig: 1.

Patients may undergo many laboratory and clinical tests within a defined time period. A major problem is that there may be multiple

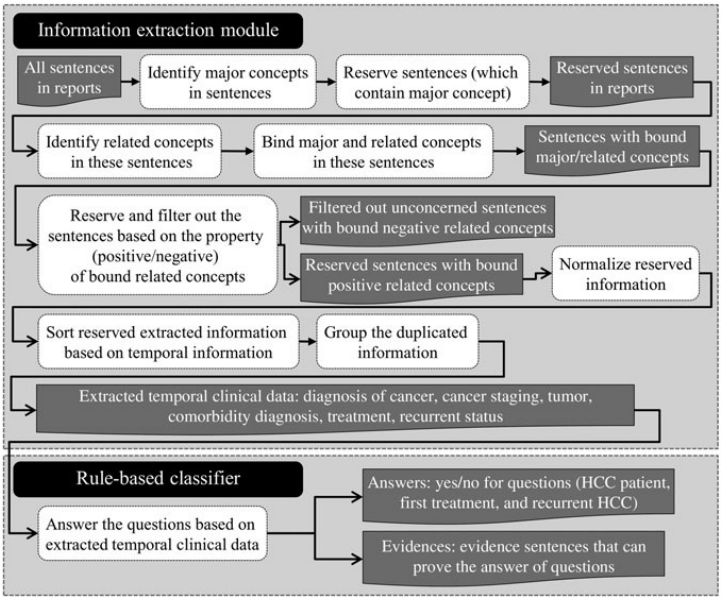

Fig. 1. Overview of the method

values for a feature in one period. In 2014, Wei-Ti Su [8] 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: 2. only one value of a feature which is closest to treatment date is taken when there are many values for that feature.

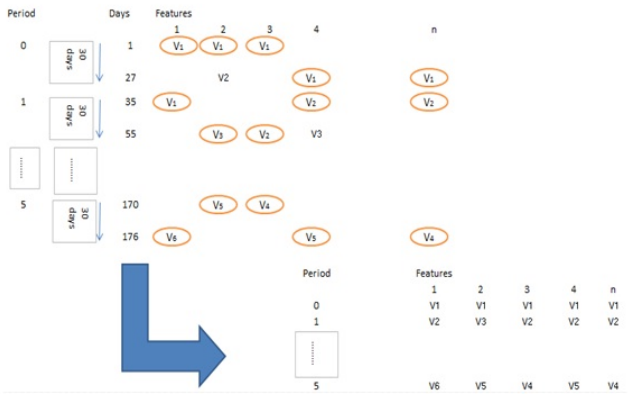

Fig. 2. Time period merging example

Selecting relevant features for classification process is a major task. In 2001, Leo Breiman [9], 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.

In 1995,C.Cortes [10],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: 3

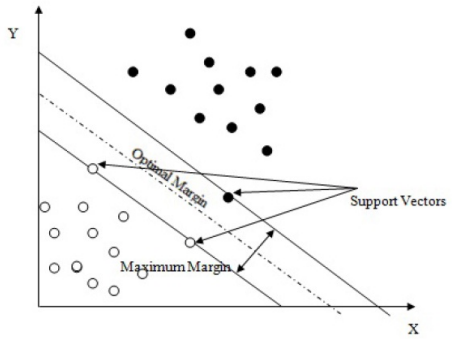

Fig. 3. Support Vector Machine

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.

3. TABULAR REPRESENTATION OF METHODS AND FUNCTIONS USED

Functions, advantages and disadvantages of different types of methods used are represented in table 1.

Table 1. Tabular representation of methods and functions used

Method	Function	Advantages	Disadvantages
Merging algorithm	For merging multiple time series data	• Obtains single value for a feature over a defined period • Different time period multiple measurements are merged	Time taken for computation is slightly high
Random forest	Feature Selection	• Relevant features are extracted • High accuracy in feature selection	They produce less accuracy in prediction
Support vector Machine	Classification	• Avoids over-fitting • Highly efficient • Used as classical linear classification technique	• High algorithmic complexity • Extensive requirements for memory

4. EXPERIMENTAL ANALYSIS

Dataset contains 83 patients details,where 18 were recurrent patients and 65 were nonrecurrent.the classification performance for HCC recurrence prediction using single measurement,multiple measurements were analysed.Optimal model is estabhlished with MMSVM with RBF using multiple measurements and a period of 120 days(accuracy 0.771,BAC 0.603)(Table 1).Performance model of MMSVM with RBF using multiple measurements were higher

than for those estabhlised by SVM with RBF using single measurements (accuracy 0.626,BAC 0.459).

Table 2. Performance of HCC Recurrence Prediction Based on Single Measurement,Multiple Mesurements

Classification algorithm	Measurement type	Periods(days)	ACC[a]	BAC[b]
SVM[c] with RBF[d] Kernel	Single measurement	-	0.626	0,459
MMSVM[e] with RBF Kernel	Multiple measurements	120	0.771	0.603

[a] Accuracy=(TP+TN)/(TP+TN+FP+FN),[b] Balancedaccuracy=(Sensitivity+Specificity)/2, Sensitivity=TP/(TP+FN), Specificity=TN/(TN+FP),[c] SVM:support vector machine,[d] RBF:radial basis function,[e] MMSVM:multiple measurements support vector machine,True positive (TP): a recurrent patient correctly identified as a recurrent patient. False positive (FP): a non-recurrent patient wrongly identified as a recurrent patient. True negative (TN): a non-recurrent patient correctly identified as a non-recurrent patient. False negative (FN): a recurrent patient wrongly identified as a nonrecurrent patient.

5. CONCLUSION

Various methods used in 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 random forest method improves the accuracy of feature selection. 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.In future,an efficient feature selection algorithm can be used which enhances the accuracy and also reduces the time complexity of the system.

6. REFERENCES

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

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

[3] M.A.Hernandez 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. 937, 1998.

[4] M. Lenzerini, Data integration: A theoretical perspective, in Proc. 21st ACMSIGMOD-SIGACT-SIGART Symp. Principles Database Syst., Madison, WI, USA, 2002, pp. 233246.

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

[6] M. Stacey and C. McGregor, Temporal abstraction in intelligent clinical data analysis: A survey, Artif. Intell. Med., vol. 39, no. 1, pp. 124, 2007.

[7] R. Schmidt and L. Gierl, A prognostic model for temporal courses that Combines temporal abstraction and case-based reasoning, Int. J. Med.Informat., vol. 74, nos. 24, pp. 307315, 2005.

[8] 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

[9] L. Breiman, Random forests, Mach. Learning, vol. 45, no. 1, pp. 532, 2001

[10] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learning, vol. 20, no. 3, pp. 273297, 1995.