Machine Learning of Patient Similarity

A Case Study on Predicting Survival in Cancer Patient after Locoregional Chemotherapy

LWC Chan
Department of Health Technology
and Informatics
Hong Kong Polytechnic University
Hong Kong SAR, China
wing.chi.chan@inet.polyu.edu.hk

T Chan
Department of Diagnostic Radiology
University of Hong Kong
Hong Kong SAR, China
taochan@hku.hk

LF Cheng, WS Mak
Department of Diagnostic Radiology
Princess Margaret Hospital
Hong Kong SAR, China
rickieclf@yahoo.com.hk
cat.mak@gmail.com

Abstract—**Identifying historical records of patients who are similar to the new patient could help to retrieve similar reference cases for predicting the clinical outcome of the new patient. Amongst different potential applications, this study illustrates use of patient similarity in predicting survival of patients suffering from hepatocellular carcinoma (HCC) treated with locoregional chemotherapy. This study used 14 similarity measures derived from relevant clinical and imaging parameters to classify the HCC patient pairs into two classes, namely the difference between their survival time being longer or no longer than 12 months. Furthermore, this paper proposes and presents a patient similarity algorithm for the classification, named SimSVM. With the 14 similarity measures as input, SimSVM outputs the predicted class and the degree of similarity or dissimilarity. A dataset was collected from 30 patients, forming 300 and 135 patient pairs as training and test datasets respectively. The trained SimSVM with linear kernel gave the best accuracy (66.7%), sensitivity (64.8%) and specificity (67.9%) on the test dataset.**

Keywords-Machine Learning; Patient Similarity; Support Vector Machine; Cancer; Survival

I. INTRODUCTION

Case-based reasoning and patient similarity are well known concepts for mining useful information from database but their applications in clinical decision support have not been deeply investigated. In this paper, the patient similarity method is adopted for prognostication of patients suffering from hepatocellular carcinoma (HCC) who underwent Transarterial chemoembolization (TACE). TACE is a locoregional chemotherapy treatment method, in which chemotherapeutic drugs with embolic material are delivered to the target tumor through the feeding hepatic arteries of liver tumors. Owing to the almost exclusive arterial supply of HCC, the occlusion of the feeding arteries with the embolic material causes ischemia and thus dramatically increases the contact and local concentration of the chemotherapeutic agent with the tumor. TACE is widely used as a treatment option to control symptoms, improve quality of life and extend survival for the inoperable HCCs [1-3].

A number of risk factors potentially affect survival rate after TACE. According to the accepted treatment guidelines, including those published by the American Association for Study of Liver Diseases (AASLD), the European Association for the Study of the Liver (EASL), and the Japan Society of Hepatology, TACE is recommended if some selection criteria are satisfied, e.g. no vascular invasion, no extrahepatic spread and more than 4 lesions [4-6]. Other risk factors include refractory ascites, extrahepatic metastases, hepatofugal blood flow, encephalopathy, active gastrointestinal bleeding, and advanced liver disease [1]. However, these risk factors do not absolutely or directly jeopardize the survival rate as long-term survival was observed in some high risk HCCs. The evaluation of survival benefit of TACE for individual patients becomes a critical issue in the personalized medicine.

Electronic Health Record (EHR) system is patient-centered information resource supported by computer software and hardware infrastructure, providing the archiving and communications of clinical information of each patient throughout the episodes of care [7]. The historical patient records after TACE provide informative evidence for evaluation of survival benefit. It is interesting to hypothesize that the survival time of a patient with HCC, who will undergo TACE, can be predicted from previous records of similar patients whose characteristics were documented in EHR.

According to Trevisani, et al. (1995) [8], degree of liver damage, maximum tumor size, number of lesion(s), portal vein invasion, and alphafetoprotein (AFP) value are five independent predictors of patient prognosis. This study is aimed to establish patient similarity measures using these five predictors and other risk factors, including age, gender, treatments before TACE, complete blood picture, liver function test results, hepatitis-B, renal function test results, locations of lesions and image findings, which are generally crucial for prognosis, to apply machine learning to form the linear or nonlinear combinations of these similarity measures, and to evaluate the accuracy the learnt algorithm in matching patients with similar survival time, which were not investigated or analyzed by other studies..

Support Vector Machine (SVM), derived from the statistical learning theory, is a linear weighted combination of symmetric kernels satisfying the Mercer's condition [9-11]. The input of the kernels consists of the features of interest, while the output is a dichotomous or scalar outcome estimate with respect to the features of interest. Because of

978-1-4244-8302-0/10/$26.00 ©2010 IEEE

strong theoretical foundations and excellent empirical success, SVM can be used to combine the abovementioned similarity measures with statistical learning and then form a model for classifying the patient pairs into "similar" or "dissimilar" class. The direct classification of patients into long or short survival time using SVM is not considered in this study because the model becomes unreliable and inaccurate to estimate the survival time of the current patient if the SVM is trained based on the empirical data from TACE records of patients who may be dissimilar to the current patient. To obtain a more accurate estimate of survival time, the records of the similar patients are retrieved and sorted with the similarity first and then the decision support system picks a number of patient records according to the similarity and predict the survival time of the current patient from that of the selected similar patients. Such case-based reasoning approach makes the solution more specific to individuals when compared with the global model identification approach. The success of the patient similarity in identifying HCC patients with similar survival time after TACE represents a patient-oriented clinical decision support system for the personalized medicine.

II. METHODS

A. Dataset

Historical data of 30 HCC patients who had undergone TACE were retrospectively collected from the EHR system from a regional hospital in Hong Kong. The data were de-identified during the collection. The fields of the collected data and their corresponding categories and variables used for the computer algorithm are shown in Table I.

TABLE I. DATA FIELDS AND THE CORRESPONDING CATEGORIES AND VARIABLES IN THE COLLECTED DATASET

Category	*Data field*	*Variable*
Age	Age	age
Gender	Male Female	gender1 gender2
Treatment before TACE	Operation Radiofrequency ablation Alcohol Embolization	ptreat1 ptreat2 ptreat3 ptreat4
Complete blood picture	Hemoglobin Platelet International normalized ratio	cbp1 cbp2 cbp3
Liver function test	Albumin Bilirubin Alkaline phosphatase Alanine aminotransferase	lft1 lft2 lft3 lft4
Serology –Hepatitis	Hepatitis-B Not hepatitis-B	hepb1 hepb2
Serology –AFP	Alphafetoprotein value	afp
Renal Function Test	Serum urea Serum creatinine	rft1 rft2
Number of lesions	Number of lesions	lesion
Locations of lesions	Left side Right side	side1 side2
Tumor size	Tumor size	tsize
Portal vein invasion	Portal vein invasion No portal vein invasion	pvein1 pvein2
Degree of liver damage	Cirrhosis No cirrhosis	damage1 damage2
Other image findings	Tumor enhancement Splenomegaly Ascites Varices Metastasis	image1 image2 image3 image4 image5

B. Patient Similarity Measure

The data fields, representing TACE risk factors, can be classified into 14 independent categories, including the five predictors mentioned in the previous section. A similarity measure is used to generate a similarity score for each category. The scores quantify the similarity between two patients with respect to the corresponding categories.

For those categories consisting of a single scalar data field (age, AFP value, number of lesions and tumor size), the similarity measure between the ith and jth patients is given by the following expression (we use AFP value as an example).

$$SimAFP(i,j)=\frac{1}{1+\left|afp(i)-afp(j)\right|}. \qquad (1)$$

For those categories consisting of two mutually exclusive binary data fields (gender, hepatitis, portal vein invasion, degree of liver damage), the patient similarity is given by the following expression (we use degree of liver damage as an example).

$$SimDamage(i,j)=damage(i)\bullet damage(j). \qquad (2)$$

where damage(k)=[damage$_1$(k),damage$_2$(k)] and • is the dot product.

For those categories consisting of multiple binary data fields (treatment before TACE and other image findings), the patient similarity is given by the following expression (we use treatment after TACE as an example).

$$SimLtreat(i,j)=\frac{ltreat(i)\bullet ltreat(j)}{\left|ltreat(i)\right|\cdot\left|ltreat(j)\right|}. \qquad (3)$$

where ltreat(k)=[ltreat$_1$(k), ltreat$_2$(k), ltreat$_3$(k), 1] and |.| is the modulus of a vector.

For those categories consisting of two independent binary data fields (locations of lesions), the patient similarity is given by the following expression.

$$SimLocLesion(i,j)=\frac{side(i)\bullet side(j)}{\left|side(i)\right|\cdot\left|side(j)\right|}. \qquad (4)$$

where side(k)=[side$_1$(k), side$_2$(k)].

For those categories consisting of multiple scalar data fields (Complete blood picture, liver function test and renal function test), the patient similarity is given by the following expression (we use liver function test as an example).

$$SimLFT(i,j)=\frac{lft(i)\bullet lft(j)}{|lft(i)|\cdot|lft(j)|}. \qquad (5)$$

where lft(k)=[lft$_1$(k), …, lft$_4$(k), 1].

C. *Performance analysis*

The dataset was used in the generation of 435 patient pairs. The survival time of the 30 patients are known. Those patient pairs are labeled with 1 as similar if the difference in the survival time is not greater than 12 months; otherwise, -1 as dissimilar. All the patient pairs are divided into training dataset of 300 patient pairs and test dataset of 135 patient pairs. The ratios of similar pairs to dissimilar pairs in training and test datasets are 130:170 and 54:81 respectively.

Regarding similar pairs as positives and dissimilar pairs as negatives, the accuracy of matching is defined as the sum of the true positives and negatives divided by the total number of pairs in the test dataset. The sensitivity is defined as the true positive rate and specificity, 1- false positive rate.

D. *Support Vector Machine*

SVM in the binary classification setting is considered in this study. For the kth patient pair, the input of SVM is the feature vector x(k) consisting of the 14 similarity measures as its elements. For n patient pairs, the training dataset comprises feature vectors {x(1), …, x(n)} and the given labels {y(1) . . . y(n)} where y(i) $\in$ {−1, 1}. Using two-dimensional feature space as an example, SVM learns a hyperplane that separates the training data by a maximal margin as illustrated in Fig. 1. Each point (circle or cross) in this plot represent the feature vector of a patient pair. Feature vectors lying on one side of the hyperplane are ideally the similar patient pairs, while those on the other side are dissimilar patient pairs. The feature vectors that lie closest to the hyperplane are called support vectors. SVM allow one to transform the feature vector x to a higher dimensional space through a kernel K. The classifier f(x) is represented in the following form.

$$f(x)=\sum_{i=1}^{n}\alpha_i K(x(i),x)+b. \qquad (6)$$

where α_i is the Lagrange multiplier and b is the bias. The kernel K(u,v) can be represented by a dot product of the nonlinear regressors, ϕ(u) and ϕ(v), when the Mercer's condition is satisfied [9]. Commonly used kernels include linear kernel (u•v), polynomial kernel ($(u\bullet v+1)^r$) and radial basis function kernel ($e^{-\gamma|u-v||u-v|}$). The classifier can be rewritten by the following form.

$$f(x)=w\bullet\phi(x)+b. \qquad (7)$$

where w is the weight vector normal to the hyperplane and defined by the following equation.

$$w=\sum_{i=1}^{n}\alpha_i\phi(x(i)). \qquad (8)$$

As the available information from the 35 data fields may not be sufficient to identify the similarity of some patient pairs, there exist pairs misclassified by the SVM.

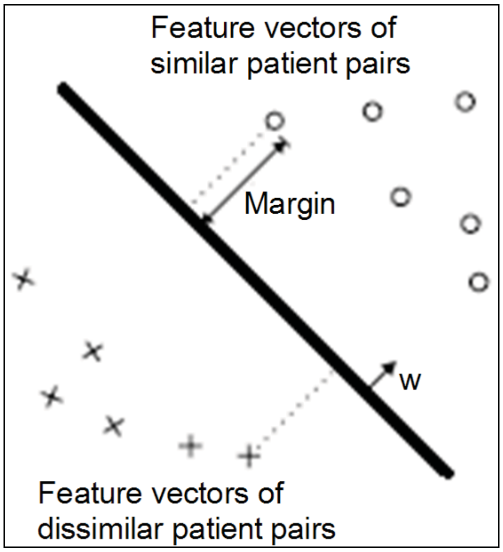
Feature vectors of similar patient pairs
Margin
w
Feature vectors of dissimilar patient pairs

Figure 1. The ideal hyperplane learnt by SVM separates the feature vectors with a maximal margin. The weight vector w and the bias specify uniquely the hyperplane.

Through the constrained optimization of a linear combination of the model complexity and an error penalty function, the maximal margin is "soft" so that the learnt hyperplane can split the feature vectors as cleanly as possible. A regulating constant is used to achieve a balance between model complexity and error penalty. After the constrained optimization, some Lagrange multipliers become zeros and those feature vectors of training dataset corresponding to the non-zero Lagrange multipliers represent the support vectors of the SVM. The resulting SVM is expressed in the following form.

$$z=\sum_{j=1}^{p}\alpha_j K(SV_j,x)+b. \qquad (9)$$

where z is the output of the SVM, SV$_j$ is the jth support vectors and p is the number of support vectors. After the training, the SVM given by (9) can be used to determine whether, by logic sign(z), and how much, by logic |z|, a new patient pair in the test dataset is similar or dissimilar with each other. The computation of the similarity measures, the trained SVM and the determining logics can be combined to form a patient similarity algorithm, which is called SimSVM in this paper. Fig. 2 illustrates the architecture of the SimSVM where the input comprises of two patient records to be compared and the output includes the predicted class of the patient pair (similar or dissimilar) and the corresponding degree of similarity or dissimilarity.

III. Results

SimSVMs with linear, radial basis function (RBF) and polynomial kernels were trained using the training dataset. The values of γ for RBF and r for polynomial were 1 and 3 respectively. The soft margin coefficient was set at infinity because the accuracy of the prediction is more important than the model complexity in this application. These three SimSVMs were trained using the training dataset and then applied to the test dataset.

It is shown in Table II that the SimSVM with linear kernel outperformed SimSVMs with RBF kernel and polynomial kernel as it gave the best accuracy, sensitivity and specificity on the test dataset.

TABLE II. Accuracy, sensitivity and specificity of SimSVMs with radial basis function (RBF), polynomial and linear kernels.

SimSVM	**Performance**		
	Accuracy	***Sensitivity***	***Specificity***
RBF	0.600	0.500	0.667
Polynomial	0.578	0.482	0.642
Linear	0.667	0.648	0.679

IV. Discussion

This study investigated the use of 14 similarity measures in determining the similarity between patients with HCC with respect to the survival time after TACE.

A patient similarity algorithm, called SimSVM, is presented in this paper. SimSVMs were learnt using the training dataset and applied to the test dataset. The accuracy and sensitivity of SimSVM with linear kernel were superior to that with polynomial or RBF kernel.

SimSVM with linear kernel can be regarded as a weighted combination of the 14 similarity measures and the weights were estimated using the training dataset. The use of linear kernel outperforming the use of RBF and polynomial indicates that the patient similarity in survival time is just linearly related to the 14 similarity measures.

Further research work would focus on the derivation and choice of similarity measures that can yield better performance in classification of similar and dissimilar patient pairs.

References

[1] T.J. Vogl, N.N.N. Naguib1, N.E.A. Nour-Eldin, et al., "Review on transarterial chemoembolization in hepatocellular carcinoma: Palliative, combined, neoadjuvant, bridging, and symptomatic indications", European Journal of Radiology, 72, 2009, pp. 505–516.

[2] L.F. Cheng, K.F. Ma, W.C. Fan, et al., "Hepatocellular carcinoma with extrahepatic collateral arterial supply", Journal of Medical Imaging and Radiation Oncology, 54, 2010, pp. 26–34.

[3] F. Trevisani, S.D. Notariis, C. Rossi, and M. Bernardi, "Randomized Control Trials on Chemoembolization for Hepatocellular Carcinoma: Is There Room for New Studies?", J Clin Gastroenterol, 32(5), 2001, pp. 383–389.

[4] J. Bruix, and M. Sherman, "Management of hepatocellular carcinoma", Hepatology, 42(5), 2005, pp. 1208–36.

[5] M. Kudo, and T. Okanoue, "Management of hepatocellular carcinoma in Japan: consensus-based clinical practice manual proposed by the Japan Society of Hepatology", Oncology, 72(Suppl 1), 2007, pp. 2–15.

[6] J. Bruix, M. Sherman, J.M. Llovet, et al., "Clinical management of hepatocellular carcinoma, Conclusions of the Barcelona-2000 EASL conference", J Hepatol, 35(3), 2001, pp. 421–30.

[7] N.T. Cheung, V. Fung, W.N. Wong, et al., "Principles-based medical informatics for success--how Hong Kong built one of the world's largest integrated longitudinal electronic patient records", Studies in health technology and informatics, 129(1), 2007, pp. 307-310.

[8] F. Trevisani, P.E. D'Intino, P. Caraceni, et al., "Etiologic factors and clinical presentation of hepatocellular carcinoma. Differences between cirrhotic and noncirrhotic Italian patients", Cancer, 75, 1995, pp. 2220–32.

[9] C.J.C. Burges, "Geometry and invariance in kernel based methods", In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), Advanced in Kernel Methods – Support Vector Learning, MIT Press, Cambridge, USA, 1998, pp. 89-116.

[10] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[11] W.C. Chan, C.W. Chan, K.C. Cheung, and C.J. Harris, "On the modeling of nonlinear dynamic systems using support vector neural networks", Engineering Applications of Artificial Intelligence, 14, 2001, pp. 105-13.

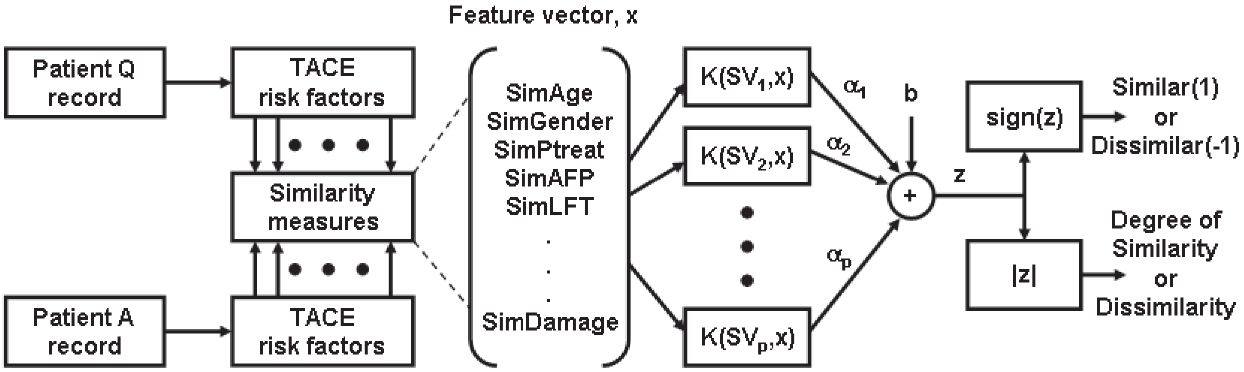

Figure 2. Architecture of SimSVM.