<table>
<thead>
<tr>
<th>Title</th>
<th>Machine learning of patient similarity: a case study on predicting survival in cancer patient after locoregional chemotherapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Chan, LWC; Chan, T; Cheng, LF; Mak, WS</td>
</tr>
<tr>
<td>Citation</td>
<td>The 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW 2010), Hong Kong, China, 18-21 December 2010. In Conference Proceedings, 2010, p. 467-470</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2010</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/197308">http://hdl.handle.net/10722/197308</a></td>
</tr>
<tr>
<td>Rights</td>
<td>IEEE International Conference on Bioinformatics &amp; Biomedicine Workshops. Copyright © IEEE.; ©2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
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
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>
<thead>
<tr>
<th>Category</th>
<th>Data field</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
<td>age</td>
</tr>
<tr>
<td>Gender</td>
<td>Male, Female</td>
<td>gender1, gender2</td>
</tr>
<tr>
<td>Treatment before TACE</td>
<td>Operation, Radiofrequency ablation, Alcohol, Embolization</td>
<td>ptre1, ptre2, ptre3, ptre4</td>
</tr>
<tr>
<td>Complete blood picture</td>
<td>Hemoglobin, Platelet, International normalized ratio</td>
<td>cbp1, cbp2, cbp3</td>
</tr>
<tr>
<td>Liver function test</td>
<td>Albumin, Bilirubin, Alkaline phosphatase, Alanine aminotransferase</td>
<td>il1, il2, il3, il4</td>
</tr>
<tr>
<td>Serology –Hepatitis</td>
<td>Hepatitis-B, Not hepatitis-B</td>
<td>hepb1, hepb2</td>
</tr>
<tr>
<td>Serology –AFP</td>
<td>Alphabetoprotein value</td>
<td>afp</td>
</tr>
<tr>
<td>Renal Function Test</td>
<td>Serum urea, Serum creatinine</td>
<td>rft1, rft2</td>
</tr>
<tr>
<td>Number of lesions</td>
<td>Number of lesions</td>
<td>lesion</td>
</tr>
<tr>
<td>Locations of lesions</td>
<td>Left side, Right side</td>
<td>side1, side2</td>
</tr>
</tbody>
</table>

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 $i$th and $j$th patients is given by the following expression (we use AFP value as an example).

$$SimAFP(i,j) = \frac{1}{1 + |AFP(i) - AFP(j)|}$$

where $|AFP(k)| = |AFP_1(k), AFP_2(k)|$ and $\cdot$ is the dot product.

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).

$$SimDamage(i,j) = damage(i) \cdot damage(j).$$

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

$$SimLtreat(i,j) = \frac{Ltreat(i) \cdot Ltreat(j)}{|Ltreat(i)| \cdot |Ltreat(j)|}.$$

where $|Ltreat(k)| = |Ltreat_1(k), Ltreat_2(k), Ltreat_3(k), 1|$ and $\cdot$ 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.

$$SimLocLesion(i,j) = \frac{side(i) \cdot side(j)}{|side(i)| \cdot |side(j)|}.$$
where side(k) = [side1(k), side2(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).

\[
SimLFT(i, j) = \frac{\|lft(i)\| \cdot |lft(j)|}{\|lft(i)\| \cdot \|lft(j)\|},
\]

where lft(k) = [lft1(k), …, lft4(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.

\[
f(x) = \sum_{i=1}^{n} \alpha_i K(x(i), x) + b.
\]

where \(\alpha_i\) 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, \(\phi(u)\) and \(\phi(v)\), when the Mercer’s condition is satisfied [9]. Commonly used kernels include linear kernel \((u \cdot v)\), polynomial kernel \(((u \cdot v) + 1)^r\) and radial basis function kernel \((e^{-\gamma |u-v|^2})\). The classifier can be rewritten by the following form.

\[
f(x) = w \cdot \phi(x) + b.
\]

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

\[
w = \sum_{i=1}^{n} \alpha_i \phi(x(i)).
\]

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.

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.

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

where \(z\) is the output of the SVM, \(SV_j\) is the \(j^{th}\) 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 $\gamma$ 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>
<thead>
<tr>
<th>SimSVM</th>
<th>Performance</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>RBF</td>
<td>0.600</td>
<td>0.500</td>
<td>0.667</td>
</tr>
<tr>
<td>Polynomial</td>
<td>0.578</td>
<td>0.482</td>
<td>0.642</td>
</tr>
<tr>
<td>Linear</td>
<td>0.667</td>
<td>0.648</td>
<td>0.679</td>
</tr>
</tbody>
</table>

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