A Biomedical Patient Data Driven Approach for the Prediction of Tumor Motion

Jonanthan Klinginsmith[1], Malika Mahoui[2], Huanmei Wu[2], Yuqing Wu[1]
[1] Indiana University, Bloomington
[2] Indiana University – Purdue University, Indianapolis

Purpose: IGRT is a recent advancement in the treatment of cancer that presents a great potential to increase the efficiency of treatment of tumor in the lower abdomen and lungs. However, the efficacy of treating tumors with radiation in these locations is often degraded by tumor respiratory motion. Therefore, the characterization and prediction of tumor motion aids in the precision of radiotherapy treatment. We hereby propose a knowledge discovery solution based on the correlation of the patient biomedical data and the tumor motion data for accurate tumor motion characterization and prediction.

Method and Materials: For the analysis of biomedical data, we worked through the main steps involved in a typical knowledge discovery analysis. An important phase includes the analysis of a large spectrum of biomedical data falling into several categories such as tumor description data, and patient treatment data in order to select the set of features to be considered for the mining process. We used clustering techniques such as K-means clustering to group patients based on a selected set of biomedical data attributes.

Results: Comprehensive preprocessing of the raw clinical data and several experiments were performed to identify stable patient clustering. The clustering results were graphically represented using the tumor location of patients for further analysis, which clearly demonstrate certain consistency among the grouping of patients based on their biomedical information. We have compared our clustering results with the current tumor location representation based on bronchopulmonary segments.

Conclusion: Patient biomedical data is a rich set of information that has the great potential in tumor characterization and predication especially for the treatment of patients with little or no tumor motion data. Combining the biomedical information and tumor motion data to explore the correlation among them will yield more accurate tumor motion predication.

1. Introduction

Radiation therapy is a common treatment for cancers in the thoracic and abdominal regions. The goal of radiation therapy is to ensure precise radiation delivery to kill tumor cells. To avoid side effects, radiation to surrounding healthy tissues and critical structures must be minimized. However, the quality of radiation treatment is complicated by intra-fraction motion, especially for abdominal or lung cancers. Intra-fraction organ motion is mainly caused by patient respiration, sometimes also by skeletal, muscular, cardiac, or gastrointestinal systems. In radio-therapy, intra-fraction organ motion is perceived as tumor motion. Several studies have shown that tumor motion has an impact on the quality of radiotherapy treatment. Several approaches are proposed to characterize tumor motion. The outcome is used for tumor motion prediction. One particular approach that we have explored is to leverage on the wealth of patients biomedical data available. Analysis can be performed on biomedical data to discover correlations between patients. These correlations, therefore, aid in characterization of tumor motion.

2. Methods

2.1 Description of patient biomedical information

Each patient that undertakes radiotherapy is described with tumor motion data generated during therapy sessions as well as with a rich set of biomedical data. This later type of data falls into several categories, including (1) basic patient information, such as patient's weight, height and age, (2) tumor description, such as tumor volume, anatomy, tumor location, pathology and tumor adhesion to the cardiac or aortic wall, (3) patient health information, such as complications which include asthma, emphysema or others, (4) patient treatment history such as CDDP, DOC, VSD, VAT or surgery, (5) marker description such as marker location, size, and the distance between marker and tumor, (6) radiation treatment information, such as duration of treatment and radiation dose, and (7) patient physiological condition during treatment, such as heartbeat, blood pressure and fullness of stomach. Our hypothesis is that the correlation between the biomedical information and the tumor motion information exist and can be discovered. Furthermore, such a correlation can be leveraged to improve the quality of the following medical services.

- For patients with little or no motion data such as in the case of new patients; if the group to which the patient belongs can be identified based on biomedical similarities, reliable tumor motion pattern can be predicted, based on the motion information of patients in the group.
- Combining the biomedical information and respiratory motion information of patients, more accurate models can be generated which will improve the prediction of tumor motion.
- Changes in patients respiratory motion pattern will be used in helping identify potential changes in the patients' biomedical condition.

For these cases our hypothesis relies on the identification of patient groups based on the motion information and biomedical information, and the discovery of the correlations between motion patterns and biomedical patterns. The approach we adopted and started experimenting for the preliminary results consists of:

1. Mining patient biomedical data and producing biomedical information-based grouping of patients.
2. Mining motion data and producing motion-based grouping of patients.
3. Discovering correlation of motion data and biomedical data using the clustering generated in the first two steps.

This presentation focuses on clustering patients based on biomedical data. This clustering process is part of a more elaborated process generally known as the knowledge discovery process, which also includes several steps known as pre-processing, such as data cleaning, data integration, data selection and data transformation. Unsupervised clustering techniques such as k-means, hierarchical clustering or neural network are candidates for the clustering task in the first two steps. For the preliminary results we used k-means clustering algorithm for its simplicity and the quality of grouping it may achieve. After patient biomedical data is clustered, a post processing phase is deployed to analyze the newly acquired knowledge. Data presentation is also part of this phase which applies visualization techniques to illustrate the findings.

2.2 Data preprocessing

One set of patient biomedical data is available to us with 32 features (table 1). Because the data was not originally collected and stored for the purpose of automatic processing but rather to be archived and used by human beings (medical doctors) when needed, a data cleaning phase is a paramount step in the knowledge discovery process.

The analysis of the data showed that many attributes in the patient biomedical information did not have a properly defined domain of values. For example, the attribute *adhesion* which specifies the adhesion of the tumor includes values such as "attached to the posterior chest wall." We are dealing with sentences which are difficult to map into discrete values (necessary for the clustering process). Moreover, multiple terms are used in expressing the same concept, and some sentences even contain typos. Part of the cleaning phase is to properly define the attributes and their domains. For example, we restrict the value of one dimension of lung tumor location to a set {posterior, anterior, middle}. A special care was attributed to the *anatomy* field that specifies the biomedical location of the tumor. Two domains were considered during the clustering process depending on whether we distinguish between the left and right side of the lungs. Based on the redefined attributes and their domains, patient records are automatically checked for errors.

Basic patient information	**Tumor description**	**Patient health info**	**Patient treatment**	**Marker description**	**Radiation treatment**	**Patient physiological condition**
age	location, volume, anatomy, pathology, location, adhesion	complications Heart beat		distance from the cardiac wall	days (duration)	heat beat,

Table 1. Sample of the initial set of patient biomedical data

Missing values is another important issue that has impact on the quality of the mining result. Missing values exist because some features may not be possible to be measured when patient information is gathered. For example, the tumor location was difficult to be specified for some patients. Several alternative approaches are available in dealing with missing values, including ignoring patient data with missing values, automatically assigning a default value such as the attribute mean, or assigning a global default value. Depending on the nature of the attribute, different approaches are considered. For example, for a patient whose tumor position is not available, the record is dropped from the dataset before the analysis. On the other hand, missing values about patient's age is assigned using the attribute mean.

Original Field Label	**Description**	**Example Data**
Age	patient's age	77
Pathology	tumor pathology	squamous cell
Anatomy	lung bronchopulmonary segment	LS3
Location (Cranio-caudal, ventro-dorsal)	cranio-caudal value, ventro-doral value	middle, middle
Tumor adhesion to the cardiac or aortic wall(minimum)	description of adhesion information	far from the heart, attached to the lateral chest wall
Distance from the cardiac (aortic) wall to the marker(cm)	distance in cm	7.2
Heart beat(bpm)	heart beats per minute	82
Indication	cancer was indicated	after myocardial infraction
Tx before XRT	multiple surgery, VAT	multiple surgery, VAT
Complications	other complications	Aortic aneurysm
volume (cm^3)		376.2
Days	days of treatment	13

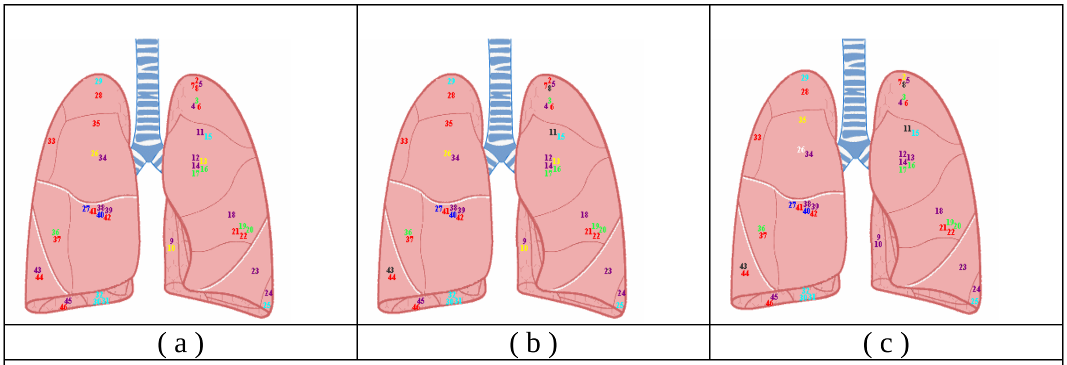
Table 2. List of attributes of the biomedical data selected for the clustering process

Not all attributes of the patient information may contribute to the clustering process. The next step in the preprocessing phase – data selection (also called dimension reduction) – aims to discard irrelevant or weak attributes that won't contribute positively to the clustering process. Relying on the expertise brought from medical and radiotherapy fields, the initial 32 fields were reduced to 12 fields, including tumor location, age, heartbeat, etc. (see table 2). We also ran automatic attribute selection techniques such as chi-squared method or PCA method to complement the manual attribute selection. Furthermore, values in some attributes are normalized by the software performing the clustering process.

2.3 Biomedical data clustering and analysis of the results

Weka software was selected as the software kit; and k-means clustering technique is used in the clustering analysis of the patient biomedical information. We choose the Euclidian distance similarity measure to specify how similar any two patients are, based on the selected features. Our experiments varied the parameter k, which specifies the maximum number of clusters to be generated, between 2 and 10.

The clustering results were analyzed to determine the stable points – we considered a clustering configuration stable if increasing the configuration k results in minimum changes in the clustering. For the 48-patient data set, we found that five, six and seven are the stable points. Based on this clustering, we assigned each patient a unique identification number (id) and plotted the anatomical tumor location of the patient on a two-dimensional space on the lungs' bronco-pulmonary segments, with the patients' id color-coded according to the clustering he/she belongs to, as shown in Figure 1.



(a) (b) (c)

Figure 1: Clustering results based on biomedical information
(a) 6 clusters (b) 7 clusters (c) 8 clusters.

3. Results

While the clustering results clearly demonstrate certain consistency among the grouping of patients based on their biomedical information, they pose more questions to our research. For example, whether the current tumor location representation based on broncho-pulmonary segments is the best way in describing the tumor location for the purpose of position prediction in radiation treatment? Will the clustering change if representation changes? Does the change make sense? Our next step is to present the results to medical and radiotherapy experts to find semantic support to the selected list of cluster sets. The feedback from the experts will be used to refine every step of the knowledge discovery process, including further refinement of the pre-processing procedure, redefining or refining data representation and transformation, exploring different similarly measures, readjusting the weights among attributes, etc. Future work includes combining clustering results with tumor motion data during the tumor prediction process. To conclude, the study revealed that patient biomedical data is a rich set of information that has the great potential in tumor characterization and predication especially for the treatment of patients with no or little tumor motion data. Combine the biomedical information and tumor motion data and explore the correlation among them will yield more accurate tumor motion predication.