ONCO-i2b2: improve patients selection through CBR techniques with heterogeneous distance functions

Daniele Segagni[1]✉, Matteo Gabetta[2], Valentina Tibollo[1], Arianna Dagliati[3], Alberto Zambelli[2], Cristiana Larizza[2], Silvia G Priori[1], Riccardo G Bellazzi[2]

[1]Laboratorio di Informatica e Sistemistica per la Ricerca Clinica, IRCCS Fondazione Salvatore Maugeri, Pavia, Italy
[2]Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Pavia, Italy
[3]IUSS, Istituto Universitario di Studi Superiori, Pavia, Italy

Motivation and Objectives

The University of Pavia (UNIPV) and the IRCCS Fondazione Salvatore Maugeri hospital (FSM) in Pavia have recently started an information technology initiative to support clinical research in oncology called ONCO-i2b2. This project aims at supporting translational research in oncology and exploits the software solutions implemented by the Informatics for Integrating Biology and the Bedside (i2b2) research center. The ONCO-i2b2 software is designed to integrate the i2b2 infrastructure with the hospital information system, with the pathology unit and with a cancer biobank that manages both plasma and cancer tissue samples. Exploiting the medical concepts related to each patient, we have developed a novel data mining procedure that allows researchers to easily identify patients similar to those found with the i2b2 query tool, so as to increase the number of patients, compared to the patient set directly retrieved by the query. This allows physicians to obtain additional information that can support new insights in the study of tumors.

Methods

ONCO-i2b2 is based on the software developed by the Informatics for Integrating Biology and the Bedside (i2b2) research center. I2b2 has delivered an open source suite centered on a data warehouse, which is efficiently queried to find sets of interesting patients through a query tool interface.

The ONCO-i2b2 system gathers data from the FSM pathology unit (PU) database and from the hospital biobank, and integrates them with clinical information from the hospital information system (HIS).

One of the main functionalities of the ONCO-i2b2 project is related to the ability of gathering data about patients and samples, collected during the day-to-day activities of the Oncology I department of the FSM. ONCO-i2b2 also makes these data available for research purposes in an easy, secure and de-identified way. When a patient is hospitalized, he/she is invited to sign an informed consent to make available for research the samples, specimens and data collected for clinical purposes. Specimens obtained surgically are first analyzed by the pathologists of the PU, who may decide to send the specimens exceeding their expertise (together with the signed informed consent) to the laboratory of experimental oncology. The next step consists in the biobank storage of bio-specimens. ONCO-i2b2 is activated when a biopsy is performed to obtain a detailed diagnosis, and a report is generated. The report contains the cancer diagnosis, including the cancer 'stages' and the size of the tumor. These pieces of information are extracted using a dedicated natural language processing (NLP) module and will be used as concepts for running queries within the i2b2 web client. During this phase the selected samples are de-identified through the use of a new barcode, which does not contain any direct information about the donor.

At the same time the system integrates clinical data automatically from the FSM HIS and matches this information to the biobank samples. This information is then stored in the i2b2 Clinical Research Chart (CRC), the star schema data warehouse on which i2b2 is based.

Within the ONCO-i2b2 project, a case-based reasoning procedure has been developed, in order to allow researchers to enhance the patient selection process with an information retrieval procedure that uses the whole medical concept space related to a patient set to identify a group of similar patients. This functionality supports the extension of the original patient set obtained with the i2b2 query tool, and allows the extraction of the most similar patients to a specific patient on the basis of a set of variables.

At the current stage of the project we are mainly focused on the comparison between patients'

clinical data and we are going to expand the CBR system to allow analysis based on heterogeneous (binary, nominal and continuous) variables. Binary variables refer to the presence/absence of diseases or signs/symptoms. Nominal and continuous variables, instead, represent discrete/continuous values of clinical observations.

Concerning binary variables, to calculate the distance between two patients we exploit the Unified Medical Language System (UMLS) Metathesaurus to model their relationship, in order to create a uniform structure that can be used to compare patients, based on a normalized layer. After a patient set has been retrieved using the i2b2 query tool, the procedure finds all concepts related to patients' binary observations (disease, signs, symptoms) by means of an array containing UMLS concepts. Each concept is represented by its Concept Unique Identifier (CUI), a code that identifies concepts in the UMLS Metathesaurus, and by a boolean modifier, which indicates if the variable referring to the CUI is asserted or negated. The distance computed between cases exploits the semantic similarity between concepts in the UMLS ontology. For this reason we consider such a distance, a Semantic Distance (SD).

The CBR system we are developing computes the distance between patients considering both SD and the Interpolated Value Difference Metric (IVDM) distance proposed by Wilson and Martinez (1997) designed to handle applications with nominal attributes, continuous attributes, or both. It combines the two distances to derive the distance between two patients on the basis of any combination of binary, nominal and continuous variables. The distance function for the Interpolated Value Difference Metric for an attribute a on two patients x and y is defined as:

$$IVDM(x,y) = \sum_{a=1}^{m} ivdm_a(x_a, y_a)^2$$

where $ivdm_a$ is defined as:

$$ivdm_a(x,y) = \begin{cases} \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^2 & \text{if a is discrete} \\ \sum_{c=1}^{C} \left| P_{a,c}(x) - P_{a,c}(y) \right|^2 & \text{otherwise} \end{cases}$$

Results and Discussion

In this phase of the project we have tested the accuracy of the IVDM metric using a specific dataset derived from the amount of cancer data the ONCO-i2b2 CRC contains. Data used in the test phase are related to breast cancer patients, classified by histopathological attributes and concerning cells receptor status: estrogen receptor (ER), progesterone receptor (PR) and HER2. Cells with or without these receptors are called ER positive (ER+), ER negative (ER-), PR positive (PR+), PR negative (PR-), HER2 positive (HER2+), and HER2 negative (HER2-). Cells with none of these receptors are called basal-like or triple negative (TN).

We used as Case Base a cohort of 300 patients, classified in Luminal A (ER+ and low grade), Luminal B (ER+ but often high grade) and TN and a set of 60 patents has been used to validate the IVDM method. Table 1 shows the results of the validation phase.

Table1: this table describes the accuracy of the IVDM method and the number of similar patients rightly predicted using a test set of 60 patients (20 cases for each class).

Class	Similar patients found	Accuracy
Luminal A	12	60%
Luminal B	16	80%
Triple Negative	16	80%
Total	44	73%

The future step of this work consists in combining the SD with the IVDM in order to be able to handle in the distance computation any kind of variables. The next effort will be to combine the two distances in a function that weights them through a coefficient λ to be defined in dependence on the relevance of the two set of variables as:

$$dist = \lambda \times SD + (1-\lambda) \times IVDM$$

At this time the ONCO-i2b2 CRC contains the data of about 7,000 patients related to breast cancer diagnosis (about 600 of them have at least one biological sample in the cancer biobank), totaling about 50,000 visits and 120,000 observations recorded using 960 concepts. This very huge data set will represent a very relevant mean for validating our CBR system. The patient retrieval time is in the order of a

few seconds for patient sets up to 1000 patients. The performance decreases for larger patient sets. The implementation of such heterogeneous distance function we expect will enhance the i2b2 framework allowing the exploitation of the overall patient set for a most flexible patients retrieval from the ONCO-i2b2 CRC. Further evolutions of the system are related to import clinical data coming from the ordinary medical activity like haematochemical or instrumental.

Acknowledgements

The Onco-i2b2 project is funded by the "Regione Lombardia" in Italy. We gratefully acknowledge Prof. Carlo Bernasconi and the Collegio Ghislieri in Pavia for their active support. This paper revises and extends the paper "ONCO-i2b2: improve patients selection through case-based information retrieval techniques", by D. Segagni et al, presented at the DILS 2012 conference in Washington DC.

References

1. Betsy L Humphreys, Donald A B Lindberg, Harold M Schoolman, G Octo Barnett (1998) The Unified Medical Language System: An Informatics Research Collaboration. J Am Med Inform Assoc. 5, 1-11
2. Caviedes J, Cimino J (2004) Towards the development of a conceptual distance metric for the UMLS. J. Biomed. Inform. 37, 77-85
3. Mate S, Bürkle T, et al. (2011) Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. Stud Health Technol Inform. 169,502-506
4. Melton GB, et al. (2006) Inter-patient distance metrics using SNOMED CT defining relationships. J.Biomed. Inform. 39(6), 697-705
5. Murphy SN, Weber G, et al. (2010) Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 17(2), 124-130
6. Segagni D, et al. (2012) ONCO-i2b2: improve patients selection through case-based information retrieval techniques. Lecture Notes in Computer Science 7348, 93-99
7. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP (2012) Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Inform Assoc. [Epub ahead of print] PubMed PMID: 22822041
8. Wilson DR, Martinez TR (1997) Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research 6, 1-34