Hierarchical visual case-based reasoning for supporting breast cancer therapy

Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, Brigitte Séroussi

This work was funded by the European Union’s Horizon 2020 research and innovation program under grant agreement No 690238.
Introduction

Breast cancer
- One of the most common types of cancer that affects women in Europe
- High survival rate at 10 years

Artificial Intelligence supports the diagnostic of breast cancer
- Deep learning
- SVM
- Image analysis

But supporting the therapy is more complex !!!
- Many treatments exist, with 4 main categories:
  - surgery, chemotherapy, endocrine therapy and radiotherapy
- Many clinical data need to be considered
  - Clinical data are often not structured, contrary to medical images
- Difficult to produce a learning base
  - For a patient, the best treatment is never known

Source: MedicalXPress
Introduction

The problem of explanations

- Physicians need to understand the rationale of a recommendation in order to follow it.
- For diagnosis systems, an annotated image can make a decent explanation.

But for therapy, explanations are much more difficult to produce:
  - And time is limited (3 minutes per patient in breast cancer unit).

=> Explainable Artificial Intelligence (XAI)
Introduction

The DESIREE European H2020 project
Decision Support and Information Management System for Breast Cancer

Objectives:
- To help clinicians with the management of patient data and images
- To support primary breast cancer therapeutical decision

A web-based platform with 3 decision-support modules:
- Clinical practice guidelines implementation using formal ontologies
- Statistical machine learning through rule-learning
- Case-based reasoning (CBR)
Introduction

Case-based Reasoning (CBR)

- A form of analogical reasoning
  - No learning: CBR does not try to learn a model
  - Typical example: kNN (k nearest neighbor)

3 steps:
- Retrieve similar older cases from a database, including cases with known solutions
- Adapt their solutions to the new case
- Retain the new case in the case database

In the therapeutic context
- A case = a patient
- A solution = a treatment
Introduction

Case-based Reasoning (CBR)

- Particularly interesting for producing explanations (XAI)
- The old cases can be used as explanations
  - This way of reasoning is familiar to physicians

=> Explanations may consist is the presentation of 2-50 similar cases

- But 2-50 breast cancer patient records represent a huge volume of data!

- A solution is the use of information visualization for displaying the cases
An automatic/visual approach to CBR

Previous works
- An be automatic or visual
  - Translate visually the CBR reasoning
- Displays case similarities (new case vs old ones)
- Qualitative similarity
- Quantitative similarity

Case database is a relational database

HL7 FHIR standard is used for communication with the clinical platform

Cases are retrieved using jColibri
Cases are retrieved using jColibri

Computes a distance matrix between cases

<table>
<thead>
<tr>
<th>Query</th>
<th>Similar #1</th>
<th>Similar #2</th>
<th>Similar #3</th>
<th>Similar #4</th>
<th>Similar #5</th>
<th>Similar #6</th>
<th>Similar #7</th>
</tr>
</thead>
<tbody>
<tr>
<td>dim #1</td>
<td>val #1</td>
<td>val #1b</td>
<td>val #1c</td>
<td><strong>val #1</strong></td>
<td>val #1d</td>
<td>val #1e</td>
<td>val #1f</td>
</tr>
<tr>
<td>dim #2</td>
<td>val #2a</td>
<td>val #2a</td>
<td>val #2b</td>
<td><strong>val #2a</strong></td>
<td>val #2d</td>
<td>val #2e</td>
<td><strong>val #2b</strong></td>
</tr>
<tr>
<td>dim #3</td>
<td>val #3</td>
<td>val #3</td>
<td>val #3</td>
<td>val #3c</td>
<td>val #3d</td>
<td>val #3e</td>
<td>val #3f</td>
</tr>
<tr>
<td>dim #4</td>
<td>val #4</td>
<td>val #4</td>
<td>val #4</td>
<td>val #4c</td>
<td>val #4d</td>
<td>val #4e</td>
<td><strong>val #4</strong></td>
</tr>
<tr>
<td>dim #5</td>
<td>val #5a</td>
<td>val #5b</td>
<td>val #5c</td>
<td><strong>val #5</strong></td>
<td>val #5</td>
<td>val #5e</td>
<td>val #5f</td>
</tr>
<tr>
<td>dim #6</td>
<td>val #6</td>
<td>val #6a</td>
<td>val #6b</td>
<td>val #6c</td>
<td>val #6d</td>
<td>val #6e</td>
<td>val #6f</td>
</tr>
</tbody>
</table>

```
<table>
<thead>
<tr>
<th></th>
<th><strong>Query</strong></th>
<th>Similar #1</th>
<th>Similar #2</th>
<th>Similar #3</th>
<th>Similar #4</th>
<th>Similar #5</th>
<th>Similar #6</th>
<th>Similar #7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar #1</td>
<td>2.0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar #2</td>
<td>2.1</td>
<td>1.5</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar #3</td>
<td>2.0</td>
<td>5.0</td>
<td>4.8</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar #4</td>
<td>1.9</td>
<td>5.1</td>
<td>4.9</td>
<td>1.1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar #5</td>
<td>4.5</td>
<td>5.2</td>
<td>5.2</td>
<td>6.0</td>
<td>6.1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar #6</td>
<td>4.2</td>
<td>1.7</td>
<td>1.8</td>
<td>5.5</td>
<td>5.6</td>
<td>5.5</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Similar #7</td>
<td>2.0</td>
<td>5.3</td>
<td>5.1</td>
<td>5.4</td>
<td>5.3</td>
<td>3.1</td>
<td>5.2</td>
<td>-</td>
</tr>
</tbody>
</table>
```
Visualization of quantitative similarities

Scatter plot: 2D projection of the distance matrix
- 1 dot = 1 patient
- Colors = classes
- Target facilitates distance evaluation

Various methods for scatter plot
- MDS (Multi-Dimensional Scaling), PCA, tSNE, ...

Here, 2 types of distances:
- A - Between the new patient and a similar patient (more important!)
- B - Between two similar patients

=> we used polar MDS
- Preserve distances of type A to the detriment of those of type B
Polar MDS scatter plot

- Origin $O =$ new patient
- Each similar case $S$ is defined by their polar coordinates $(L, \theta)$
- $L$ is already known: it is the distance between $S$ and $O$
- $\theta$ is determined by solving an optimization problem:
  - Find the best values $\theta$ that minimize the stress function:

$$S_p(d) = \sum_{2<i<j} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}}$$

=> no information loss for the distances involving the new patient

Number 1 is the new patient
AFB metaheuristic

Artificial Feeding Birds (AFB) [Lamy JB. Artificial Feeding Birds (AFB): a new metaheuristic inspired by the behavior of pigeons, Advances in nature-inspired computing and applications 2019, Springer]

Simple ➔ Performant ➔ Generic

Can solve any optimisation problem defined by a triplet of functions (cost(), fly(), walk() )
Visualization of qualitative similarities

Rainbow boxes

- A recent set visualization technique
- Elements are patients
- Sets are shared characteristics
- Set of patients with “age > 60”
- Only the two major therapeutic decisions are kept
- Numeric values are discretized using the Minimum Description Length Principle (MDLP)
- Only the boxes with the highest Mutual Information (MI) are kept
Rainbow boxes

- Elements => columns
- Sets => rectangular boxes
- Color => one color per element
- Box color is the mean of its elements color
- Non contiguous element in a set => box hole
- Elements are ordered so as to minimize the number of holes
- Boxes are stacked vertically by size

Rainbow boxes for patient similarity

- Column height = similarity with the new patient
- Box height = importance of the box (MI)
- Box color = weighted mean of the header’s color
  - Indicate toward which therapy orientates the box

Visual reasoning

<table>
<thead>
<tr>
<th>Similar #6</th>
<th>Similar #2</th>
<th>Similar #1</th>
<th>New patient</th>
<th>Similar #3</th>
<th>Similar #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>dim #2 = val #2b</td>
<td>dim #2 = val #2a</td>
<td>dim #3 = val #3</td>
<td>dim #1 = val #1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dim #5 = val #5</td>
<td></td>
</tr>
</tbody>
</table>
function classify(q, X, s, n, m):
    q is the query case
    X is the case database (we assume that q ∈ X)
    s is the dissimilarity measure (a function taking 2 cases and returning their dissimilarity, e.g. Euclidean distance)
    n ≥ 2 is the total number of cases considered (query + similar cases)
    m ≥ 1 is the maximum number of boxes selected

For each case i in X, compute s(q, Xi)
Let X' be the set of selected cases, X' contains the n elements of X with the lowest dissimilarity s(q, i)
We assume that X'_1 = q and X'_2 to X'_n are the similar cases

Let d be the distance matrix between cases in X'
For each case i in X':
    For each case j in X':
        d_{i,j} = s(i, j)

Let w be the weights of the similar cases
w_i = \begin{align*}
    1, & \text{ if } s_{\text{max}} = s_{\text{min}} \\
    \frac{s_{\text{max}} - s(q, X'_i)}{s_{\text{max}} - s_{\text{min}}}, & \text{ otherwise}
\end{align*}
with s_{\text{min}} = \min(s(q, X'_i)) and s_{\text{max}} = \max(s(q, X'_i))

Let y_1 and y_2 be the two best classes in X' (determined by a majority vote over similar cases, weighted by w_i)
Let X'' be the subset of X' displayed in rainbow boxes
X'' = X' \cap (\{q\} \cup y_1 \cup y_2)

Let B be the set of candidates boxes (currently empty)
For each dimension Z:
    If Z has numeric values:
        Discretize Z
        For each value v that Z takes in cases X'':
            If q has value v for dimension Z:
                Add Zv = \{x ∈ X'' \mid x_Z = v\} into B

For each box Zv in B, compute MI(ZvY) = \sum_{z \in \{Z = v, Z \neq v\}} \sum_{y \in \{y_1, y_2\}} p(z, y) \log \left( \frac{p(z, y)}{p(z)p(y)} \right)
with p(y) = \frac{|y|}{|X'' \setminus \{q\}|}, p(Z = v) = \frac{|Z_v|}{|X'' \setminus \{q\}|}, p(Z \neq v) = \frac{|Z_v|}{|X'' \setminus \{q\}|}, p(Z = v, y) = \frac{|Z_v \cap y|}{|X'' \setminus \{q\}|}, p(Z \neq v, y) = \frac{|X'' \setminus Z_v \cap y|}{|X'' \setminus \{q\}|}

Let B' be the set of selected boxes, B' contains the m elements of B with the highest MI(ZvY)

Compute S_{y_1} = \sum_{Z_v \in B'} \left( MI(Z_vY) \times \sum_{2 \leq i \leq n} w_i \mid x_i \in Z_v \cap y_1 \right)
Compute S_{y_2} = \sum_{Z_v \in B'} \left( MI(Z_vY) \times \sum_{2 \leq i \leq n} w_i \mid x_i \in Z_v \cap y_2 \right)

If S_{y_1} > S_{y_2}:
    return y_1
Else:
    return y_2
Boxes give arguments in favor one type of therapy

Physicians may choose a different option if he disagrees
Resulting interface

Boxes give arguments in favor one type of therapy
Physicians may choose a different option if he disagrees

Limited to 2-6 classes of therapy
=> Hierarchical approach dividing the decision in several smaller ones
Ontology of breast cancer therapy

We organized possible therapies in a formal ontology
- OWL format
- Owlready ontology-oriented programming module for Python

Class hierarchy: Lumpectomy

- 'Breast Cancer Procedure'
  - 'Breast Cancer loco-regional procedure'
  - 'Breast Cancer Radiotherapy'
  - 'Breast Irradiation'
  - 'Chest Wall Irradiation'
  - 'Lymph Node irradiation'
- 'Breast Cancer Surgical Procedure'
  - 'Breast Cancer Removal Surgery Procedure'
  - 'Breast Surgical Procedure'
  - 'Biopsy of Breast'
  - 'Breast Conservation Treatment'
- Lumpectomy
  - Quadrantectomy
  - 'Breast Re-Excision'
  - Mastectomy
  - 'Oncoplastic Breast Surgery'
  - 'Lymph Node Surgical Procedure'
  - 'Breast Plastic Surgery Procedure'
  - 'Breast Surgery Ancillary Procedure'
  - 'Breast Cancer Non Therapeutic Management'
- 'Breast Cancer Systemic Therapy'
  - 'Ancillary Systemic Therapy'
  - 'Endocrine Therapy'
  - 'Systemic Chemotherapy'

French book on Owlready!
JB Lamy
Python et les ontologies
ENI editions, 2019
Hierarchical approach

- Buttons allow to choose one of the two major classes of therapy
- Then, the visualization is limited to the similar patients with this therapy
- New classes are determined, according to the ontology
Lumpectomy

#3176
#3175
#2988
#3089
#2753
#2754
#3094
#0
Query

ERRResult ≥ 84
ki67Result: 8-40
tumor size ≤ 40

age: 60-70

tumor size atMammography

Compare Lumpectomy...

Mastectomy

#3250
#3259
#2995
#2978
#3162
#3154
#3110

ERRResult ≥ 84

age < 49

her2IHCResult =
breast co

Compare Mastectomy...
### Cyclophosphamide Doxorubicin Paclitaxel Trastuzumab standard therapy

<table>
<thead>
<tr>
<th>#</th>
<th>Query</th>
<th>#0</th>
<th>#3028</th>
<th>#3053</th>
</tr>
</thead>
<tbody>
<tr>
<td>2966</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2789</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **age:** 48-63
- **tumor size at Ultrasound:** 37.5-48.0
- **tumorBIRADSCategory at MRI = Birads5**

### Cyclophosphamide Docetaxel Epirubicin

- multiple tumors at Ultrasound

- **tumor size at Mammography:** = 40.0
Discussion and conclusion

- A hierarchical visual approach for explainable therapeutical decision-making

- Similar accuracy as kNN, but better explainability

- Main limits:
  - May require some training for the physicians
  - Number of similar cases is reduced at each iteration
    - Should we extract additional cases to compensate?

- Set visualization is an interesting approach to explainable artificial intelligence (XAI)

- Perspectives:
  - Clinical validation and evaluation
  - Adaptation to other domains
  - Extension to other AI techniques (deep learning, boosting)

JB Lamy. Artificial Feeding Birds (AFB): a new metaheuristic inspired by the behavior of pigeons. Advances in nature-inspired computing and applications 2019

JB Lamy. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artificial Intelligence in Medicine 2017;80:11-28
