# Statistical Methods for Managing Missing Data: Application to Medical Diagnosis

**Souad Guessoum, Hadjer Zaayout, Nabiha Azizi, Nadjet Dendani and Hayet Djellali**

**Abstract** This paper presents a work consisting in realizing a decision support system based on the technique of case-base reasoning and dedicated to the diagnosis of a very dangerous pulmonary pathology: lung cancer. The system is realized for the oncology department of Ibn roch hospital of Annaba (Algeria) and will help young oncologist physicians in their activity by providing them with the experience of experts in the same domain. The principle issue in this work is the missing data in the system memory relating to the patient's state. Indeed, missing values prevent the achievement of the diagnosis process. The problem is treated by proposing two statistical approaches in addition to re-evaluate in this new domain some ones which have been already proposed and evaluated in a previously domain. The validation is made on a base of 40 real cases collected from the archive of oncology department. Cases are collected as paper documents.

**Keywords** Artificial intelligence · Case-based reasoning technique · Decision support system · Medical diagnosis · Respiratory diseases · Missing data problem · Lung cancer

S. Guessoum (✉) · H. Djellali
LRS Laboratory, Badji Mokhtar University, Annaba, Algeria
e-mail: souadguessoum@yahoo.fr

H. Djellali
e-mail: hayetdjellali@yahoo.fr

H. Zaayout
Informatics Department, Badji Mokhtar University, Annaba, Algeria
e-mail: zaayouthadjer@yahoo.fr

N. Azizi (✉) · N. Dendani (✉)
Labged Laboratory, Badji Mokhtar University, Annaba, Algeria
e-mail: nabiha111@yahoo.fr; azizi@labged.com

N. Dendani
e-mail: ndendani@yahoo.fr



N. Dey et al. (eds.), *Medical Imaging in Clinical Applications*,
Studies in Computational Intelligence 651, DOI 10.1007/978-3-319-33793-7_16

## 1 Introduction

The lung cancer is a very dangerous disease which represents the cause of 1.3 million deaths a year in the world. Its principal cause is the chronic exposure to tobacco smoke, including passive smoking. The incidence of lung cancer among no-smokers, represents 15 % of cases and is often attributed to a combination of genetic factors and pollution air. According to the World Health Organization, the lung cancer is the most frequent cause of cancer death among men, and after breast cancer among women. The lung cancer is characterized by a set of clinical symptoms. Its diagnosis needs in addition some radiological examinations to search data required for accurate diagnosis. So physicians may ask for different examinations like thoracic scanner, abdominal scanner, endoscopy, MRI (Magnetic Resonance Imaging), and thoracic radiograph. Provide assistance to physicians in the diagnosis of this disease is the aim of the system presented in this paper.

To solve problems of daily life, we naturally use our old experiences, by remembering similar situations already encountered, which are compare with the current situation for building a new solution which in turn, be added to our experience. This human reasoning is often used by physicians during their activity of diagnose and treat patients. Since Artificial Intelligence (AI) is interesting in reproducing human reasoning on machine, there was a technique named Case-Based Reasoning (CBR) which reproduces the human behavior in his recourse to his past experiences to solve his new problems. So, the global idea of CBR is to solve a new problem by finding similar case(s) in a knowledge base and to adapt it (them) to the new situation.

This work is a continuation of a former one presented in [1], which involved the development of a support system dedicated to the medical decision. It was applied to the diagnosis of a dangerous respiratory disease caused by tobacco: Chronic Obstructive Pulmonary Disease (COPD). The system is called RespiDiag and is based on CBR technique principles. As its name suggests, RespiDiag (for Respiratory diseases Diagnosis) is supposed go beyond the diagnosis of COPD for extending to other pathologies of the same field. In this vision, the work is expanded this time to the diagnosis of lung cancer. However, and currently, the implementation of the system is done separately of RespiDiag pending its integration in a future step.

In this work, we are cooperating with specialist physicians of the oncology department in Ibn Rochd hospital of Annaba (Algeria) for giving a decision support system that can help young clinicians in the diagnosis of lung cancer. The system will give help for future oncologist physicians because it gathers experiences of many experts in the domain who are not always available. Some statistical methods are integrated in this system for managing the problem of missing data in its case base.

Knowing that the aim of the work is twofold: first, the realization of the CBR decision support system with the implementation of all phases of CBR cycle, and

secondly, the managing of the problem of missing data that can appear in the case base and/or in the new problem, the paper is organized as follows: Sect. 2 provides an overview on some cbr medical systems that can be found in the literature, while Sect. 3 presents principles and cycle of the CBR technique. Section 4 presents the work methodology and the application field.

The contribution of the paper begins in Sect. 5 that describes all the details on the CBR process with all similarity metrics proposed and used in the system. The same section introduces the problem of missing data, and gives all details on the second contribution of the paper with the proposed approaches for managing the missing data. Experimentations results are presented, compared and evaluated in Sect. 6. The paper is concluded by the last section.

## 2 Related Works

In the literature, many works focalized on the medical CBR systems can be found. This is motivated mainly by the fact that the case-based reasoning is very similar to clinical reasoning. Indeed, given a patient to diagnose, physician uses his past experiences to look for any resemblance between former patient's symptoms and those of the new patient. Such resemblance (if it exists) can help immensely in the decision about the new patient, in term of making the most precise diagnosis or proposing the most efficient treatment. And it is in this way that the competence of a physician relies heavily on his own experiences. So, CBR systems which modelise very well the reasoning on experiences, can be so helpful to support physicians in their decision about diagnoses/therapies.

Many medical fields of diagnosis or therapy, have benefited of CBR decision support systems. Each of these ones have targeted a specific problem with the CBR process or the considered field.

Among many medical CBR systems we can mention some ones as CASEY [2] dedicated to the diagnosis of heart failure, FM-Ultranet [3] which diagnoses fetal deformations, KASIMIR [4] which provides a treatment for breast cancer, the system developed in [5] for the diagnosis of acute bacterial meningitis and RespiDiag [1] for the diagnosis of COPD.

In the last one, the retrieval phase has been developed in depth and saw the proposal of some similarity metrics for specific attributes to the field of COPD, for which conventional metrics were inappropriate. Another problem has been studied in the same phase, that of the missing data which prevented the completion of its process. The work has seen the proposal and the evaluation of several approaches to manage it. And the present work is a continuation for RespiDiag that would see an enlargement to other respiratory diseases (as its name suggests).

In [6], it can be found a work that compares bayesian inferences and CBR applied to the diagnosis of acute bacterial meningitis. The comparison of results shows that the CBR system diagnosis was more accurate than the bayesian system

one. Results of the CBR system before and after the reuse phase were compared too and conclude that the adaptation step gives more flexibility and more robustness to the system.

## 3 The Case-Based Reasoning

The main idea of CBR consists of reusing the solution of a former similar problem to solve a new one. All past experiences are gathered in the memory system called "case base", where the case is a couple of descriptors of the problem and its solution. In this technique, the new problem is called "target case" and the former cases which are already solved and saved in the case base are named "source cases". When we develop a CBR system, we don't need to know how the expert thinks for resolving the problem, because the knowledge in such system consists just of establishing a description of a problem and its solution.

CBR cycle is identified by Aamodt and Plaza in [7] as a process of four steps:

- The Retrieval phase: which is the most important phase of the cycle. It consists of calculating the similarity between the current problem and all previous problems gathered in the case base, in order to retrieve one or more most similar case(s). The number of retrieved cases depends on the decision of the constructor of the system. The process of this first phase is mainly based on similarity metrics,
- The Reuse phase: it is the most delicate step; it consists in adapting (if need be) the solution of the most similar source case to the target problem. Its difficulty is mainly due to the strong dependence of the heuristics and knowledge adaptation of the application fields. For this reason, the collaboration of experts is required during all throughout the conception process of these heuristics that are usually in rules form. For certain CBR systems dedicated particularly to the medical diagnosis, this phase is ignored, and the process consists just in finding the most similar diagnosis,
- The Revise phase: during which the adapted solution is presented to the user who will decide on his validity. In the affirmative case the last phase is begun. This step gives then, the possibility to the user of changing the details of proposed solution according his opinion. It will be a new experience for the system.
- The Retain phase: it consists of adding the new problem with its validated solution to the case base. And so the system learns of its new experiences!

Figure 1 gives the CBR cycle.

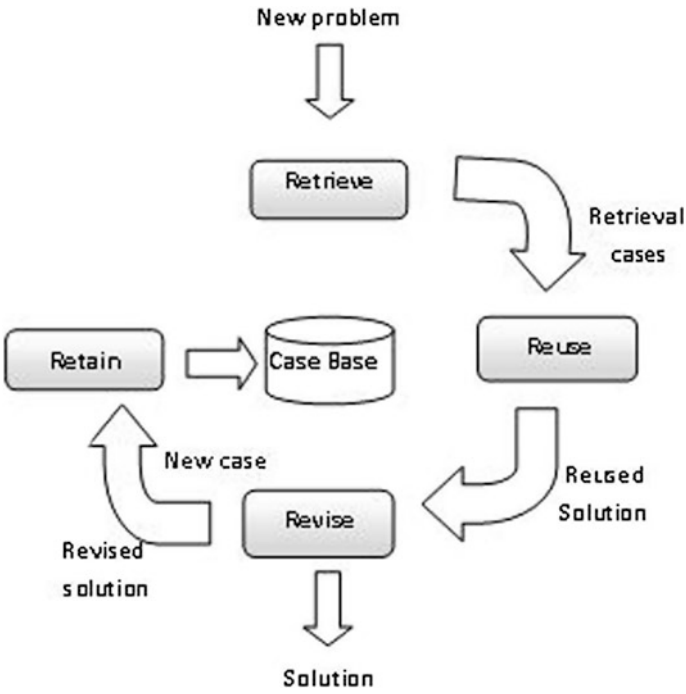


**Fig. 1** CBR cycle

## 4 Methodology

The system presented in this work is based on CBR principles, and its process involves the four phases of CBR cycle that are detailed in Sect. 3. And so, the first step is consisting in estimating similarity between the new problem and former situations gathered in the knowledge base of the system, in order to select the most similar one. This step is realized in this work by proposing general and some specific metrics for the field in question. The second phase is the reuse which adapts the retrieved solution to the new problem. it is conceived in the system by modeling a set of rules established with the collaboration of experts. The last two phases are realized in a simple way and their details are not subject of this paper.

The principal issue in this work is that the completion of the retrieval phase is prevented by a very common problem in medical systems, which is the missing data relating to patient's informations. This issue was already addressed in RespiDiag where two types of strategies have been proposed. The first one is called *online approaches* and aims to find an outcome for the retrieval process of the system by assigning values to local similarities when missing data appears. It's about *pessimistic*, *optimistic* and *medium* approaches.

The second strategy type is called *offline approaches*, and aims for its part to fill the void in the case base with plausible values estimated according to the principle of the proposed approaches. It's about the *statistical*, and *CBR* approaches.

The current work re-evaluates former approaches in the new case base containing lung cancer cases with different missing data rates of those in the first application. The work proposes also two others statistical methods. They are variants of the first ones which see changes made in their principles to obtain the *online statistic* and the *offline statistic** approaches. A comparison is dressed at the end of the evaluation of all these strategies.

## 4.1 Considered Data and Symptoms

The following set of data and symptoms is considered in the work:

- *age*: lung cancer patients are generally aged over 40 years. For physicians, this data is also necessary to specify the treatment,
- *sex*: in Algeria, men are more exposed to attrap lung cancer than women, because the majority of smokers are men,
- *profession*: this data informs physicians on the possibility for the patient to attrap lung cancer due to its polluted workplace,
- *toxic exposure*: this data informs that the patient is or not alcoholic, because alcohol increases the risk of developing lung cancer,
- *smoking*: including passive smoking,
- *packet number per year*: it is the average of number of smoked packets per year. The danger starts when this number exceeds 120 packets.
- *former health disorders*: that can be asthma, pulmonary tuberculosis, diabetes… etc.
- *chronic tiredness*,
- *anorexia*: that means loss of appetite,
- *sudden weight loss*,
- *night sweat*,
- *fever*,
- *cough*: that does not disappear and that intensifies with time,
- *thoracic pain*: that is constant and that is intensifies when breathing deeply,
- *dyspnea*: that intensifies with effort,
- *pleural effusion*: that means accumulation of fluid around the lungs,
- *hemoptysis*: that is rejection of blood from the respiratory tract following coughing,
- *swollen lymph nodes*: in the neck or over the clavicle,
- *opacity*: that means the presence of opaque spots on the thoracic radiograph,
- *hyper intensity*: that means the presence of an hyper intensity on the scanner image. We distinguish here hyper intensity on the thoracic scanner, named in this work *hyperintensity* and the hyper intensity on the abdominal scanner,

named in this work *hyper intensity's* that indicates the presence of adrenal, biliary or liver cancer which has metastasized to the lungs,
- *tumor mass*: this data informs on the presence/absence of the tumor mass seen in a review of endoscopy,
- *mass length, mass width* and *mass height*: are the measures of the tumor mass,
- *hyper fixation*: this data informs on the fixation of slightly radioactive substance (injected in the body) seen on the review of the bone scintigraphy. It indicates that the cancer has bone origin and has metastasized to the lungs,
- *hyper signal*: that indicates the presence of brain cancer that has metastasized to the lungs. This information is located on the review MRI,
- *metastasis*: this data informs on if the lung cancer has or not metastasized, it can be brain, liver, liver, bone biliary or adrenal metastasis. There may also be multiple metastases at the same time.

### *4.2 Considered Diagnoses*

Two types of lung cancer are essentially considered, they are of variable severity:

- *small-cell lung carcinoma* [8]
- *non-small-cell lung carcinoma* [9] which are essentially of three types:
  - *adenocarcinoma*
  - *squamous cell lung carcinoma*
  - *large cell lung carcinoma*

**With the aim not to encumber the reader with medical information, details on these diagnoses are given in the Appendix**.

## 5 The System Process

As already mentioned, the work aim to realize a decision support system for the diagnosis of lung cancer based on the principles of case-based reasoning. For the representation of cases, a set of 29 symptoms (descriptors) of the patient state is established with the collaboration of physicians. It is denoted in this work by `Attributes`. The new problem is denoted `tgt` where the former case is denoted `srce`.

A case is a couple of descriptions of the problem and its solution: case = (`pb`, `sol(pb)`) where `pb` is a problem describing the patient's conditions and `sol(pb)` is a diagnosis solution associated to `pb`:

- `pb` is described by the following set of attributes: `age`, `sex`, `fever`, `cough`, `dyspnea`, `profession`, `pollutedWorkplace`, `toxicExposure`,

chronicTiredness, formerHealthDisorders, nightSweat, anorexia, suddenWeightLoss, smoking, packetNumberPerYear, pleuralEffusion, thoracicPain, opacity, hemoptysis, swollenLymphNodes, hyperIntensity, hyperIntensitys, hyperFixation, hyperSignal, tumorMass, massLength, metastasis, massWidth and massHeight.

- sol(pb) is one of following diagnoses: *small cell lung carcinoma*, *pulmonary adenocarcinoma*, *bronchial adenocarcinoma*, *broncho-pulmonary adenocarcinoma*, *squamous cell lung carcinoma*, *large cell lung carcinoma* and *solitary fibrous tumor.*

The attribute pollutedWorkplace is added intentionally as a binary data for the need to estimate similarity between different values of the attribute profession. See Sect. 5.1.1 for more details. So in total we have a set of 29 Attributes.

## 5.1 The Retrieval Phase

By entering data of the tgt problem, the system will extract the most similar srce case existing in the CaseBase. This process is realized by comparing the attributes of tgt to attributes of srce and the comparison is done by assessing similarity expressed by:

$$\mathcal{S}(\texttt{srce}, \texttt{tgt}) = \frac{\sum_{a \in \texttt{Attributes}} w_a \times \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)}{\sum_{a \in \texttt{Attributes}} w_a} \tag{1}$$

where $w_a > 0$ is the weight of the attribute $a$ and $\mathcal{S}_a$ is a similarity measure defined on the range of $a$. The estimation of $\mathcal{S}_a$ depends on the type of $a$.

### 5.1.1 Similarity Metrics for Different Types of Attributes

When the type of $a$ is boolean, it is defined by:

$$\mathcal{S}_a(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{else} \end{cases} \text{for} \quad x, y \in \{\texttt{false}, \texttt{true}\} \tag{2}$$

It's the case of following attributes: sex, pollutedWorkplace, smoking, dyspnea, fever, nightSweat, pleuralEffusion, cough, hemoptysis, toxicExposure, chronicTiredness, anorexia, suddenWeightLoss, thoracicPain, opacity, swollenLymphNodes,

hyperIntensity, hyperIntensitys, hyperFixation, hyperSignal, and tumorMass.

If the attribute $a$ is of a numeric type, $\mathcal{S}_a$ is defined by

$$\mathcal{S}_a(x, y) = 1 - \frac{|y - x|}{B_a} \tag{3}$$

where $B_a$ is the "breadth" of the range of $a$, i.e., it is the difference between the maximal value of this range and its minimal value.

This equation is valid for the following numerical attributes: age, packetNumberPerYear, massLength, massWidth and massHeight.

In this application we dispose of three symbolic attributes which are profession, metastasis and formerHealthDisorders. Each of these attributes has an enumerated list of possible values.

For the estimation of similarity between the values of the attribute profession, we propose the following formula based on the profession value in addition to the information given in the attribute pollutedWorkplace. So, for the attribute profession we use:

$$\mathcal{S}_a(\texttt{tgt}, \texttt{srce}) = \begin{cases} 1 & \text{if } \texttt{tgt}.\text{profession} = \texttt{srce}.\text{profession} \\ 0.8 & \text{if } \texttt{tgt}.\text{profession} \neq \texttt{srce}.\text{profession and} \\ & \texttt{tgt}.\text{pollutedWorkplace} = \texttt{srce}.\text{pollutedWorkplace} \\ 0 & \text{else} \end{cases} \tag{4}$$

The attribute metastasis contains the name of the organ where the lung cancer has metastasized. Knowing that the basic values of the attribute: liver, adrenal, bone, biliary, or absence of metastasis, we note that can contain multiple informations at the same time. Indeed, cancer metastasis may be in one or more organs at once.

The separation between these different values for the same patient, is done simply by commas. We have chosen that the similarity between two elementary values of this attribute is to be estimated by the formula 2.

In order to calculate the similarity between composed values of srce. metastasis and tgt. metastasis, we proposed using the following formula:

$$\mathcal{S}_a(\texttt{srce}, \texttt{tgt}) = \frac{\sum_{a \in \texttt{Attributes}} \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)}{MaxCard} \tag{5}$$

where MaxCard is the maximum number of elementary values in srce. metastasis and tgt. metastasis.

For the attribute formerHealthDisorders, the same principle used for the metastasis attribute is kept. So, binary similarity is calculated between each two basic values in tgt and srce, after, the sum of all these is estimated and

divided by the maximum cardinality of the two sets of composite values in `srce` and `tgt`.

## 5.2 Attributes Weights

The `Attributes` weights have been estimated with the collaboration of experts according to the respective importances of each one for the diagnosis. Table 1 shows the importance of attributes estimated by physicians who have expressed importance by the following expressions:

**Table 1** The weights estimated by physicians

| Attribute | Type | Weight |
|---|---|---|
| age | Numeric | 0.4 |
| sex | Boolean | 0.8 |
| profession | Symbolic | 0.1 |
| pollutedWorkplace | Boolean | 0.1 |
| formerHealthDisorders | Symbolic | 0.4 |
| smoking | Boolean | 0.8 |
| packetNumberPerYear | Numeric | 0.8 |
| dyspnea | Boolean | 0.4 |
| fever | Boolean | 0.2 |
| nightSweat | Boolean | 0.2 |
| pleuralEffusion | Boolean | 0.8 |
| cough | Boolean | 0.2 |
| thoracicPain | Boolean | 0.4 |
| hemoptysis | Boolean | 0.8 |
| chronicTiredness | Boolean | 0.2 |
| toxicExposure | Boolean | 0.2 |
| suddenWeightLoss | Boolean | 0.8 |
| anorexia | Boolean | 0.4 |
| swollenLymphNodes | Boolean | 0.8 |
| opacity | Boolean | 0.4 |
| metastasis | Symbolic | 0.8 |
| hyperIntensity | Boolean | 0.8 |
| hyperIntensitys | Boolean | 0.8 |
| hyperFixation | Boolean | 0.8 |
| hyperSignal | Boolean | 0.8 |
| tumorMass | Boolean | 0.8 |
| massLength | Numeric | 0.2 |
| massWidth | Numeric | 0.2 |
| massHeight | Numeric | 0.2 |

- Very High Importance (VHI) reflected into the system by the value 0.8
- Average Importance (AI) reflected in the system by the value 0.4
- Low Importance (LI) reflected into the system by the value 0.2

It is noteworthy that the importance of `profession` data is relative to the working environment which can be polluted, so it becomes a risk factor for catching a lung cancer. In the system object of this paper, this data was reinforced by another information `pollutedWorkplace` of the binary type (yes, no). Thus the weighting coefficient of the attribute `profession` which was estimated by physicians to 0.2 (low importance), was divided by two: between `profession` and `pollutedWorkplace` with the aim to maintain balance between the weights estimated by physicians.

## 5.3 The Problem of Missing Data

During cases collecting from the archive of oncology department of Annaba, it has been observed that patient's files contain missing data which can be about different informations. Indeed, sometimes physicians may avoid certain examinations for a given patient, depending on its state or on its examinations results already obtained. It is noteworthy here that patient's files on which this work has been done are even in paper format. Naturally, the missing data in patient's files are reflected in the case base of the system by missing values that can be appeared in different attributes.

The problem of missing data raises a difficulty to achieve the case retrieval process. Indeed, the similarity value $\mathcal{S}(\texttt{srce}, \texttt{tgt})$ cannot be computed if for at least one attribute $a$, $\texttt{srce}.a$ and/or $\texttt{tgt}.a$ is unknown. In the following, the notation "$\texttt{pb}.a = ?$" (resp., "$\texttt{pb}.a \neq ?$") means that the value of $a$ for the problem pb is unknown (resp., is known).

The case base is structured in twenty nine attributes corresponding to the problem descriptors plus one attribute corresponding to the solution descriptor (the diagnosis). With the collaboration of physicians, 40 real cases were collected from the archive of the oncology department. This set of cases has been selected so as to have maximum diversity in symptoms for the same diagnosis, it is the set of the source cases. In the following, the case base is denoted by `CaseBase`. Table 2 shows some statistics on the `CaseBase`.

Another set of 20 real cases has been selected from the same archive. This set will serve as a test sample for the system. Cases of this set also contain gaps left by

**Table 2** Statistics on the `CaseBase`

| | Number | Percentage |
|---|---|---|
| Cases | 40 | |
| All data required | 1160 | 100 % |
| Present data | 872 | 75.17 % |
| Missing data | 288 | 24.82 % |

**Table 3** Statistics on the sample test

| | Number | Percentage |
|---|---|---|
| Cases | 20 | |
| All data required | 580 | 100 % |
| Present data | 447 | 77.06 % |
| Missing data | 133 | 22.93 % |

missing data, and constitute the targets cases that will allow us to compare and evaluate results of all proposed approaches.

Table 3 shows some statistics on the sample test.

## 5.4 *Previously Approaches Re-evaluated in This Application*

In this section, some approaches proposed and evaluated in the previously work [1] are summarized. They aim to manage the problem of missing data simultaneously in the case base and in the new problem. These selected approaches are included in this work for a second evaluation in the new application field which has another rates of missing data. This work propose in addition its own two new statistical approaches which are presented in Sect. 5.5.

In [1], can be found the *online approaches*, which are invoked during the system retrieval process, and each time where a missing value in the `tgt/srce` case appears. These *online approaches* are intended to assign a value to the local similarity without filling the gap left by the missing value. Three *online* strategies of them named each according to its principle, the *optimistic, pessimistic*, and *medium* approaches are reused in this work.

In the same reference, it has been proposed another type of strategies called *offline approaches* which aimed to fill the gaps in the case base. These approaches are executed outside of the system process, hence their name of *offline*, and are only concerned with the missing data in the case base (and not in the target problem). So, they consist in attributing a plausible value $\upsilon$ to $\texttt{srce}.a$, when $\texttt{srce}.a = ?$. This attribution is denoted by $\texttt{srce}.a := \upsilon$.

### 5.4.1 The Online Optimistic Approach

This approach proposes to give the most possible optimistic value to the similarity assumption that the missing value can be as close as possible to the present value, and the contribution of the local similarity $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$ to the global similarity $\mathcal{S}(\texttt{srce}, \texttt{tgt})$ is maximal, so:

$$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := 1 \tag{6}$$

### 5.4.2 The Online Pessimistic Approach

A pessimistic strategy would consist in assuming that $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$ is minimal:

$$\begin{aligned}
&\text{if } \texttt{srce}.a = \,?\, \text{ and } \texttt{tgt}.a \neq \,?\, \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \inf_{x \in \texttt{range}(a)} \mathcal{S}_a(x, \texttt{tgt}.a) \\
&\text{if } \texttt{srce}.a \neq \,?\, \text{ and } \texttt{tgt}.a = \,?\, \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \inf_{y \in \texttt{range}(a)} \mathcal{S}_a(\texttt{srce}.a, y) \\
&\text{if } \texttt{srce}.a = \,?\, \text{ and } \texttt{tgt}.a = \,?\, \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \inf_{x,y \in \texttt{range}(a)} \mathcal{S}_a(x, y)
\end{aligned} \tag{7}$$

### 5.4.3 The Online Medium Approach

This approach considers the balance between the last two approaches. So, it consists in estimating the average of the pessimistic value (denoted by $\upsilon_p$ in the following) and the optimistic one (i.e., 1):

$$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \frac{1 + \upsilon_p}{2} \tag{8}$$

### 5.4.4 The Offline Statistical Approach

This approach is based on statistics. Let $\texttt{SPWKV}_a$ be the set of the source problems with known values for attribute $a$:

$$\texttt{SPWKV}_a = \{\texttt{srce}' | (\texttt{srce}', \texttt{sol}(\texttt{srce}')) \in \texttt{CaseBase} \text{ and } \texttt{srce}'.a \neq \,?\,\}$$

When $a$ is a numerical attribute, the value of the attribute $a$ for srce consists simply in estimating the average of all values of $a$ for the source problems for which these values are known:

$$\texttt{srce}.a := \frac{\sum_{\texttt{srce}' \in \texttt{SPWKV}_a} \texttt{srce}'.a}{\text{card}\, \texttt{SPWKV}_a} \tag{9}$$

where card $X$ is the number of elements of the finite set $X$.

When $a$ is non numerical (i.e., boolean or symbolic), the statistical approach consists in making a vote:

$$\texttt{srce}.a := \operatorname*{argmax}_{\upsilon \in \texttt{range}(a)} \text{card}\{\texttt{srce}' \in \texttt{SPWKV}_\texttt{a} | \, \texttt{srce}'.a = \upsilon\} \tag{10}$$

### 5.4.5 The Offline CBR Approach

In this approach, the CBR process itself is used to propose a value to $\texttt{srce}\,.\,a$. In this way, a source case is decomposed in a different way as before: a problem is defined by all the attributes of $\texttt{srce}$ except $a$ and a solution is a value for $a$.

A new similarity measure $\mathcal{S}^a$ has been defined between these new source cases.

$$\mathcal{S}^a(\texttt{srce}_1, \texttt{srce}_2) = \frac{\sum_{b \in \texttt{Attributes} \setminus \{\texttt{a}\}} w_b^a \times \mathcal{S}_b(\texttt{srce}_1\,.\,b, \texttt{srce}_2\,.\,b)}{\sum_{b \in \texttt{Attributes} \setminus \{a\}} w_b^a} \tag{11}$$

The main difference between the *offline statistical approach* and the *offline cbr approach* one is that the values $\texttt{srce}'.\ a$ are weighted by $\mathcal{S}^a(\texttt{srce}', \texttt{srce})$.

That means that if $a$ is a numerical attribute, then:

$$\texttt{srce}\,.\,a := \frac{\sum_{\texttt{srce}' \in \texttt{SPWKV}_\texttt{a}} \mathcal{S}^a(\texttt{srce}', \texttt{srce}) \times \texttt{srce}'\,.\,a}{\sum_{\texttt{srce}' \in \texttt{SPWKV}_a} \mathcal{S}^a(\texttt{srce}', \texttt{srce})} \tag{12}$$

And if $a$ is a non numerical attribute, then a weighted vote approach is used:

$$\texttt{srce}\,.\,a := \underset{\upsilon \in \texttt{range}(a)}{\text{argmax}} \sum \{\mathcal{S}^a(\texttt{srce}', \texttt{srce}) | \texttt{srce}' \in \texttt{SPWKV}_\texttt{a} \text{ and } \texttt{srce}'.\,\texttt{a} = \upsilon\} \tag{13}$$

where $\sum X = \sum_{x \in X} x$ for any finite set of numbers $X$.

## 5.5 *New Proposed Approaches*

We were inspired by these last offline approaches, to propose and evaluate two new strategies in this work. The first one, is a variant of the offline statistic approach but is used during the online process of the system, while the second is a variant of the offline CBR approach and used outside the system process.

### 5.5.1 The Online Statistic Approach

This approach is inspired by the *offline statistical approach*, whose principle is detailed in Sect. 5.4.4. Indeed, in this strategy, the principle of the average of all informed values of $a$ in source problems is kept. So, the same formula (9) mentioned in the precedent section for the numerical attributes is used here. The

principle of the vote concerning binary or symbolic attributes (formula 10) is also kept.

The first modification brought to the *offline statistical approach* is that it is executed during the *online* system process, specifically during the retrieval phase, hence its name *online statistical approach*. It treats simultaneously the missing value in `tgt` and `srce` cases, while the first one concerned only the missing data in the case base.

The second change is in the method principle itself. Indeed, when a missing value appears in the `tgt` and/or `srce` case, it will be replaced by a plausible value *virtually* and not physically as in the first approach. The virtual replacement of a missing value means that the plausible value is considered in the calculation of local similarity without being stored in the concerned attribute. So the gap is still existing, whereas the retrieval process can be completed and therefore the most similar case can be selected from the case base despite the missing data.

Note here that if we have the missing data to both in the same attribute of `tgt` and `srce` cases, gaps will be replaced virtually by the same value, hence the local similarity will necessarily be equal to 1, which joined here the principle of the *optimistic approach*.

The virtual replacement is motivated primarily by the fact that the case base is in permanent enrichment, which can give better estimates of the missing values for future needs and therefore more reliable and more accurate diagnoses. The disadvantage of this approach is the requirement to recalculate averages that may slow down the system response time.

### 5.5.2 The Offline CBR Approach*

This *offline approach* is executed outside the system process and is concerning only the missing data in the case base. It's inspired of the strategy summarized in the Sect. 5.4.5. A modification is brought on its principle consisting in reusing the value $\texttt{srce}'.a$ of the `srce` case closest to `srce`, according to $\mathcal{S}_a$, instead of taking into account *all* the source cases $\texttt{srce}' \in \texttt{SPWKV}_a$ like in our old cbr approach.

Indeed, in this new approach, the cbr principle is used: it consists in estimating similarity between `srce` cases and `tgt` case, for selecting the most similar cases to the `tgt` one. In this approach the number of retrieved cases depends of the type of missing data to bridge. It is equal to 2 when the missing value is of numerical type and so, their average is considered for filling the gaps. If this missing value is binary, three (3) most similar cases are retrieved from the case base, and the approach proceeds to the vote for selecting the value whose number of occurrences is maximal and which will replace the gap. When the missing value is symbolic, only one most similar case is selected, and its value fills the gap.

### 5.6 The Reuse Phase

This phase consists in adapting the solution of the most similar case retrieved from the case base to the `tgt` case. It is the most delicate phase of the cbr process because of its strong dependence on the application field. With the collaboration of the experts physicians, a set of rules is established for adapting the diagnosis of the retrieved case (`srce`, `sol(srce)`). These rules are based mainly on the attributes having the highest weights. An example of rules is given below.

$$\begin{aligned} \textit{if } \texttt{sol(srce)} = \textit{bronchial} - \textit{adenocarcinoma and} \\ \texttt{tgt}.\textit{tumorMass}, = \text{``}\textit{no}'' \textit{ and} \\ \texttt{tgt}.\textit{hyperIntensity}, = \text{``}\textit{Yes}'' \textit{ and} \\ \texttt{tgt}.\textit{opacity}, = \text{``}\textit{no}' \textit{ then} \\ \texttt{sol}.(\texttt{tgt}) = \text{``}\textit{broncho} - \textit{pulmonary} - \textit{adenocarcinoma}'' \end{aligned}$$

Some work remains to be done about the adaptation process. However, the adaptation rules acquired so far have improved the performance of the system, when compared to a null adaptation approach (see Sect. 6).

## 6 Evaluation, Comparison and Discussion

The evaluation of the system considers three case bases: Base A, Base B and Base C. Only the first one contains missing data, whereas the two others are completely filled.

- Base A is the original case base which contains missing data (cf. Table 2),
- Base B is obtained by applying the *offline statistical approach* on Base A (cf. Sect. 5.4.4),
- Base C is obtained by applying our new *offline cbr approach** on Base A (cf. Sect. 5.5.2).

The evaluation consists first in gathering the diagnoses given by the system. These diagnoses correspond to each $\texttt{CaseBase} \in \{\text{Base A, Base B, Base C}\}$ each of the four *online approaches* (Sects. 5.4.1–5.4.3 and 5.5.1) each of the 20 `tgt` problems taken from the sample test.

Thus $3 \times 4 \times 20 = 240$ diagnoses are generated by the system. Then, these diagnoses are compared to the real diagnoses of the sample test given in the following section.

The evaluation of the system is conducted in two steps:

- after the input of data relating to the target problem, the most similar diagnosis is given without the application of adaptation rules. It is saved with the aim to evaluate the online approaches.

- In the second time adaptation is launched on the same diagnosis to give the final result of the system.

## 6.1 Results of Approaches

Tables 4, 5 and 6 show the different diagnoses given by the online statistic approach when applied on bases A, B and C (respectively). Tables 7, 8 and 9 show the different results given by the online pessimistic approach when applied on the bases A, B and C (respectively), while Tables 10 and 11 show the different results given by the online medium approach when applied on the bases A and B (respectively).

## 6.2 Comparison and Discussion

The following (Tables 12, 13, 14, 15, 16, 17, 18 and 19) gives the different percentages of each response quality, of each approach applied on the three bases. Note here that to assess the results quality of the system before and after the adaptation process, diagnoses were grouped into class according to their likeness. A diagnosis result is considered "good" when it is exactly the same as the real diagnosis of the test case. It is considered "average" when it belongs to the real diagnosis class, and it is "low" when it is out this class.

By observing the first and the second Tables 12 and 13, we can see that the online statistical approach is combined, significantly better with the offline statistical approach that fulfilled the base B, where it gives its best rate of "Good response" that reaches 30 % before the adaptation and which passes to 45 % after adaptation. This can be interpreted by the rapprochement between the principles of the two statistical approaches offline and online which apparently complement well. We note that the second approach was inspired from the first one.

The results of this approach in this context is not absolute, a future evaluation of it can be done by considering some existing data as missed and compare this data with the virtual values estimated by the approach. An estimation of response time is also possible in a future work to assess the potential slowdown when the case base is growing in number of records.

For the pessimistic approach (Tables 14 and 15), note that adaptation was able to double the rate of correct responses on the bases B and C. Indeed it rises from 20 to 40 % and from 15 to 30 % (respectively), while for the base A that contains the missing data, rates made a big leap from 10 to 15 %. This brings into focus the efficacy of rules adaptation that are established in the system. We also can see that the pessimistic approach gives its better score of good response when it is applied on the base B.

**Table 4** Results of the online statistic approach on the Base A

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.71 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.90 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.74 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Lung adenocarcinoma | 0.83 | Broncho-pulmonary adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Broncho-pulmonary adenocarcinoma | 0.78 | Broncho-pulmonary adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.80 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Small cell carcinoma | 0.77 | Small cell carcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.93 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.85 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.74 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.80 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Small cell carcinoma | 0.88 | Small cell carcinoma |
| 13 | Bronchial carcinoma | Small cell carcinoma | 0.88 | Small cell carcinoma |
| 14 | Solitary fibrous tumor | Small cell carcinoma | 0.90 | Small cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.71 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Squamous cell carcinoma | 0.88 | Squamous cell carcinoma |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.82 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.93 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Lung adenocarcinoma | 0.86 | Lung adenocarcinoma |
| 20 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.81 | Lung adenocarcinoma |

Finally, for the medium and optimistic approaches (Tables 16, 17, 18 and 19), it can observed that before adaptation, they give almost the same rate of good responses on the three bases, neighboring 15 %, whereas they give a rate of 45 % if we consider responses of average quality, which all become good ones after the adaptation process.

**Table 5** Results of the online statistic approach on the Base B

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.71 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.90 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Lung adenocarcinoma | 0.74 | Broncho-pulmonary adenocarcinoma |
| 4 | Lung adenocarcinoma | Lung adenocarcinoma | 0.81 | Broncho-pulmonary adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Broncho-pulmonary adenocarcinoma | 0.79 | Broncho-pulmonary adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.80 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Small cell carcinoma | 0.78 | Small cell carcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.87 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Squamous cell carcinoma | 0.83 | Squamous cell carcinoma |
| 10 | Small cell carcinoma | Bronchial adenocarcinoma | 0.71 | Bronchial adenocarcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.75 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Small cell carcinoma | 0.86 | Small cell carcinoma |
| 13 | Bronchial carcinoma | Small cell carcinoma | 0.85 | Small cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.89 | Small cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.71 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Squamous cell carcinoma | 0.81 | Squamous cell carcinoma |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.82 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.87 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Lung adenocarcinoma | 0.86 | Lung adenocarcinoma |
| 20 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.82 | Lung adenocarcinoma |

For the offline cbr approach*, we can observe that the rate of good responses is always between the rates corespondent to the bases A and B, which means that this approach that gave the base C completely filled, improved the quality of results regarding the original case base that contains the missing data.

**Table 6** Results of the online statistic approach on the Base C

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.71 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.90 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.74 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.83 | Broncho-pulmonary adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Broncho-pulmonary adenocarcinoma | 0.78 | Broncho-pulmonary adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.80 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Lung adenocarcinoma | 0.77 | Small cell carcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.93 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.85 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.74 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.80 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Small cell carcinoma | 0.88 | Small cell carcinoma |
| 13 | Bronchial Carcinoma | Squamous cell carcinoma | 0.88 | Small cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.90 | Small cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Lung adenocarcinoma | 0.71 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Lung adenocarcinoma | 0.84 | Squamous cell carcinoma |
| 17 | Small cell carcinoma | Bronchial carcinoma | 0.82 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.93 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.86 | Lung adenocarcinoma |
| 20 | Bronchial adenocarcinoma | Small cell carcinoma | 0.81 | Lung adenocarcinoma |

**Table 7** Results of the pessimistic approach on the Base A

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.53 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.80 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.63 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.69 | Bronchial adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Bronchial adenocarcinoma | 0.44 | Bronchial adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.57 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Lung adenocarcinoma | 0.67 | Lung adenocarcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.74 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.68 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.59 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.57 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.68 | Squamous cell carcinoma |
| 13 | Bronchial carcinoma | Squamous cell carcinoma | 0.67 | Squamous cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.70 | Squamous cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.57 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Solitary fibrous tumor | 0.66 | Solitary fibrous tumor |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.55 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.67 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Squamous cell carcinoma | 0.64 | Squamous cell carcinoma |
| 20 | Bronchial adenocarcinoma | Squamous cell carcinoma | 0.52 | Squamous cell carcinoma |

**Table 8** Results of the pessimistic approach on the Base B

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.53 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.80 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.63 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.69 | Bronchial adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Bronchial adenocarcinoma | 0.44 | Bronchial adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.57 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Lung adenocarcinoma | 0.67 | Lung adenocarcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.74 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.68 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.59 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.57 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.68 | Squamous cell carcinoma |
| 13 | Bronchial carcinoma | Squamous cell carcinoma | 0.67 | Squamous cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.70 | Squamous cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.57 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Solitary fibrous tumor | 0.66 | Solitary fibrous tumor |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.55 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.67 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Squamous cell carcinoma | 0.64 | Squamous cell carcinoma |
| 20 | Bronchial adenocarcinoma | Squamous cell carcinoma | 0.52 | Squamous cell carcinoma |

**Table 9** Results of the pessimistic approach on the Base C

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.53 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.80 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.63 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.69 | Bronchial adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Bronchial adenocarcinoma | 0.44 | Bronchial adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.57 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Lung adenocarcinoma | 0.67 | Lung adenocarcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.74 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.68 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.59 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.57 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.68 | Squamous cell carcinoma |
| 13 | Bronchial carcinoma | Squamous cell carcinoma | 0.67 | Squamous cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.70 | Squamous cell carcinoma |
| 15 | Pulmonary small cell carcinoma | squamous cell carcinoma | 0.57 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Solitary fibrous tumor | 0.66 | Solitary fibrous tumor |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.55 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.67 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Squamous cell carcinoma | 0.64 | Squamous cell carcinoma |
| 20 | Bronchial adenocarcinoma | Squamous cell carcinoma | 0.52 | Squamous cell carcinoma |

**Table 10** Results of the medium approach on the Base A

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.65 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.86 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.73 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Lung adenocarcinoma | 0.78 | Bronchial-pulmonary adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Bronchial adenocarcinoma | 0.65 | Bronchial adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Bronchial adenocarcinoma | 0.73 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Lung adenocarcinoma | 0.79 | Lung adenocarcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.83 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.82 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.76 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.72 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Lung adenocarcinoma | 0.77 | Squamous cell carcinoma |
| 13 | Bronchial carcinoma | Lung adenocarcinoma | 0.78 | Squamous cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.80 | Squamous cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.68 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Squamous cell carcinoma | 0.74 | Squamous cell carcinoma |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.72 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.77 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Squamous cell carcinoma | 0.78 | Squamous cell carcinoma |
| 20 | Bronchial adenocarcinoma | Squamous cell carcinoma | 0.72 | Squamous cell carcinoma |

**Table 11** Results of the medium approach on the Base B

| Attribute | Real diagnosis | Diagnosis before adaptation | Similarity | Diagnosis after adaptation |
|---|---|---|---|---|
| 1 | Broncho-pulmonary adenocarcinoma | Bronchial adenocarcinoma | 0.65 | Broncho-pulmonary adenocarcinoma |
| 2 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.88 | Broncho-pulmonary adenocarcinoma |
| 3 | Lung adenocarcinoma | bronchial adenocarcinoma | 0.72 | Bronchial adenocarcinoma |
| 4 | Lung adenocarcinoma | Bronchial adenocarcinoma | 0.84 | Bronchial adenocarcinoma |
| 5 | Bronchial adenocarcinoma | Bronchial adenocarcinoma | 0.65 | Bronchial adenocarcinoma |
| 6 | Bronchial adenocarcinoma | Lung adenocarcinoma | 0.70 | Lung adenocarcinoma |
| 7 | Squamous cell carcinoma | Lung adenocarcinoma | 0.74 | Lung adenocarcinoma |
| 8 | Squamous cell carcinoma | Squamous cell carcinoma | 0.88 | Squamous cell carcinoma |
| 9 | Squamous cell carcinoma | Lung adenocarcinoma | 0.83 | Lung adenocarcinoma |
| 10 | Small cell carcinoma | Clear cell carcinoma | 0.73 | Clear cell carcinoma |
| 11 | Small cell carcinoma | Lung adenocarcinoma | 0.74 | Lung adenocarcinoma |
| 12 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.84 | Squamous cell carcinoma |
| 13 | Bronchial carcinoma | Squamous cell carcinoma | 0.84 | Squamous cell carcinoma |
| 14 | Solitary fibrous tumor | Squamous cell carcinoma | 0.83 | Squamous cell carcinoma |
| 15 | Pulmonary small cell carcinoma | Squamous cell carcinoma | 0.66 | Squamous cell carcinoma |
| 16 | Squamous cell carcinoma | Solitary fibrous tumor | 0.80 | Solitary fibrous tumor |
| 17 | Small cell carcinoma | Lung adenocarcinoma | 0.72 | Lung adenocarcinoma |
| 18 | Broncho-pulmonary adenocarcinoma | Lung adenocarcinoma | 0.87 | Broncho-pulmonary adenocarcinoma |
| 19 | Lung adenocarcinoma | Squamous cell carcinoma | 0.78 | Squamous cell carcinoma |
| 20 | Bronchial adenocarcinoma | Squamous cell carcinoma | 0.70 | Squamous cell carcinoma |

**Table 12** Results quality of the online statistic approach before adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 20 | 40 | 40 |
| Base B | 30 | 35 | 35 |
| Base C | 15 | 35 | 50 |

**Table 13** Results quality of the online statistic approach after adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 40 | 20 | 40 |
| Base B | 45 | 20 | 35 |
| Base C | 30 | 20 | 50 |

**Table 14** Results quality of the pessimistic approach before adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 10 | 30 | 60 |
| Base B | 20 | 45 | 35 |
| Base C | 15 | 30 | 55 |

**Table 15** Results quality of the pessimistic approach after adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 25 | 15 | 60 |
| Base B | 40 | 25 | 35 |
| Base C | 30 | 15 | 55 |

**Table 16** Results quality of the medium approach before adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 20 | 30 | 50 |
| Base B | 20 | 45 | 35 |
| Base C | 20 | 35 | 45 |

**Table 17** Results quality of the medium approach after adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 40 | 10 | 50 |
| Base B | 40 | 25 | 35 |
| Base C | 35 | 20 | 45 |

**Table 18** Results quality of the optimistic approach before adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 15 | 35 | 50 |
| Base B | 20 | 45 | 35 |
| Base C | 20 | 35 | 45 |

**Table 19** Results quality of the optimistic approach after adaptation

| Base | Good response (%) | Average response (%) | Low response (%) |
|---|---|---|---|
| Base A | 40 | 10 | 50 |
| Base B | 40 | 25 | 35 |
| Base C | 35 | 20 | 45 |

## 7 Conclusion

This article presents a case-based decision support system dedicated to the diagnosis of lung cancer, which is a very lethal disease caused mainly by tobacco. The system describes a patient thanks to 29 numerical, boolean and symbolic attributes. The retrieval phase of the CBR process of the system is based on a similarity measure $S$ defined as a weighted average of local similarity measures $S_a$ associated with each attribute $a$, while the adaptation phase is based on a set of rules.

The main issue for this application is related to the missing data in the case base (about 25 %) and in the target problem (about 23 % in the sample test). In order to manage this problem, 2 approaches are defined in this paper, and are inspired by the previously strategies presented in [1], where some approaches are selected to be re-evaluate in the new domain of lung cancer diagnosis. All these approaches are distinguished into 2 categories: *online* and *offline* strategies. The first one aims at estimating, at runtime, the value of the local similarity of an attribute for which the values in the source case and/or in the target problem are unknown. The evaluation has shown that the new proposed approach online gives the best results on the sample test with a rate of 45 % of "Good responses" which coincide with the real diagnoses made by experts.

The second category is the offline strategies which aim at filling the gaps in the case base. The evaluation has shown that the new CBR offline approach* always gives the best results regarding the base case containing the missing data, but of lower quality relative to the base case filled with former statistical method of work.

All these results are related to a well-defined context data/absent and with well-defined rates. When these rates increase, the possibility of degradation of performance of one or the other approach is here. The same work can be redone, increasing the rate of missing data with the objective to evaluate the behavior of these approaches.

# Appendix

## *Considered Diagnoses*

There are essentially two types of lung cancer of variable severity:

- *small-cell lung carcinoma* [8] (SCLC): is a type of highly malignant cancer that arises most commonly within the lung, although it can occasionally arise in other body sites, such as the cervix, prostate, and gastrointestinal tract. Compared to non-small cell carcinoma, SCLC has a shorter doubling time, higher growth fraction, and earlier development of metastases. SCLCs represent about 20 % of lung cancers and are difficult to treat. They grow rapidly and, when diagnosed, it is common that cancer cells are already scattered throughout the rest of the body to form metastases (secondary tumors). In 95 % of cases, lung cancers are small cell linked to smoking.
- *non-small-cell lung carcinoma* [9] (NSCLC): is any type of epithelial lung cancer other than small cell lung carcinoma. As a class, NSCLCs are relatively insensitive to chemotherapy, compared to small cell carcinoma. When possible, they are primarily treated by surgical resection with curative intent, although chemotherapy is increasingly being used both pre-operatively and post-operatively. NSCLCs represent about 80 % of lung cancers and heal more easily because they grow more slowly. Lung cancer NON-small cell are essentially of three types:
  - *adenocarcinoma* (ADC), which account for 40 % of non-small cell cancers, sometimes affecting the alveoli and are slightly more common among non-smokers and women.
  - *squamous cell lung carcinoma*, which are more common in men than in women. They also represent 40 % of non-small cell cancers, they reach the bronchi and are associated directly to smoking,
  - *large cell lung carcinoma*, which represent 20 % of non-small cell cancers, with a faster growth than the other two types, that are caused by 90 % of tobacco consumption.

Each of these lung cancer may be bronchial, pulmonary or bronco-pulmonary. We note that there is another type of lung cancer benign, which is named solitary fibrous tumor.

# References

1. Guessoum, S., Laskri, M.T., Lieber, J.: RespiDiag: a case-based reasoning system for the diagnosis of chronic obstructive pulmonary disease. Exp. Syst. Appl. **41**(2), 267–273 (2014)
2. Koton, P.: Reasoning about evidence in causal explanations. In: Kolodner, J. (ed.) Proceedings of the Case-Based Reasoning Workshop, Clearwater Beach, Florida (1988)

3. El Balaa, Z., Strauss, A., Uziel, P., Maximini, K., Traphner, R.: FM-Ultranet: a decision support system using case-based reasoning applied to ultrasonography. In: McGinty, L. (ed.) Workshop Proceedings of the International Conference on Case-Based Reasoning, p. 3744 (2003)
4. Lieber, J., Bresson, B.: Case-based reasoning for breast cancer treatment decision helping. In: Blanzieri, E., Portinale, L. (eds.) Advances in case-based reasoning. In: Proceedings of the European Workshop on Case-Based Reasoning, EWCBR. Lecture Notes in Artificial Intelligence, vol. 1898, pp. 173–185 (2000)
5. Maurente, C., Edye, E.O., Delgado, S.H., GarcÃ-a, D.R.: Evaluation of case based reasoning for clinical decision support systems applied to acute meningitis diagnose. Innov. Adv. Comput. Sci. Eng. 259–264 (2010)
6. Ocampo, E., Maceiras, M., Herrera, S., Maurente, C., Rodriguez, R., Sicilia, M.: Comparing Bayesian inference and case-based reasoning as support techniques in the diagnosis of acute bacterial meningitis. Exp. Syst. Appl. **38**, 10343–10354 (2011)
7. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun. **7**(1), 39–59 (1994). IOS Press
8. https://en.wikipedia.org/wiki/Smallcell_carcinoma
9. https://en.wikipedia.org/wiki/Non-small-cell_lung_carcinoma
10. Lin, J.H., Haug, P.J.: Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J. Biomed. Inform. **41**, 1–14 (2008)