MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)
© 2015 IMIA and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-564-7-862

A Hybrid Approach Using Case-Based Reasoning and Rule-Based Reasoning to Support Cancer Diagnosis: A Pilot Study

Renata M. Saraiva[a], João Bezerra[c], Mirko Perkusich[c], Hyggo Almeida[a], Clauirton Siebra[b]

[a] Department of Computing and Systems, Federal University of Campina Grande, Campina Grande, PB, Brazil
[b] Department of Informatics, Federal University of Paraíba, João Pessoa, PB, Brazil
[c] Federal Institute of Paraiba, Monteiro, PB, Brazil

Abstract

Recently there has been an increasing interest in applying information technology to support the diagnosis of diseases such as cancer. In this paper, we present a hybrid approach using case-based reasoning (CBR) and rule-based reasoning (RBR) to support cancer diagnosis. We used symptoms, signs, and personal information from patients as inputs to our model. To form specialized diagnoses, we used rules to define the input factors' importance according to the patient's characteristics. The model's output presents the probability of the patient having a type of cancer. To carry out this research, we had the approval of the ethics committee at Napoleão Laureano Hospital, in João Pessoa, Brazil. To define our model's cases, we collected real patient data at Napoleão Laureano Hospital. To define our model's rules and weights, we researched specialized literature and interviewed health professional. To validate our model, we used K-fold cross validation with the data collected at Napoleão Laureano Hospital. The results showed that our approach is an effective CBR system to diagnose cancer.

Keywords:

Medical Decision Support Systems; Case Based Reasoning; Rules Based Reasoning.

Introduction

According to the World Health Organization (WHO), cancers figure among the leading causes of morbidity and mortality worldwide[1]. Researchers from the WHO and the International Agency for Research on Cancer (IARC) claim that in 2012, there were 14.1 million new cancer cases and a total of 8.2 million deaths from cancer worldwide. It is expected that the number of cancer patients will continue to rise by about 70% over the next 2 decades[1]. Early diagnosis of cancer is a big challenge because it is a disease with multiple locations and clinicopathological aspects while having no pathognomonic (i.e., specific to each disease) signs or symptoms[2]. Therefore, it can be detected in various stages of histopathological and clinic evolution. Healthy lifestyle and early diagnosis of this disease can reduce its mortality rate, according to the International Union of Cancer Control (UICC)[3].

Many researchers have applied Artificial Intelligence (AI) techniques to create health-related systems or models, such as the diagnosis or classification of diseases [1-4]. To diagnose cancer, researchers applied different computational techniques. Salem and El Bagoury [5] proposed a hybrid case-based adaptation model, that combines transformational and hierarchical adaptation techniques with artificial neural networks and certainty factors for the diagnosis of thyroid cancer. Zubi and Saad [6] combined data mining techniques with neural networks for the early diagnosis of lung cancer. For the diagnosis of breast cancer, Keles, Keles and Yavuz [7] used neuro-fuzzy rules while Sharaf-elDeen et al. [8] used a hybrid approach that combined case-based reasoning (CBR) with rule-based reasoning (RBR).

In this paper, we present a hybrid approach using CBR and RBR to assist healthcare professionals in the early diagnosis of patients with cancer. We use CBR and RBR because rules and cases are complementary [9]. Instead of using RBR as an alternative solution to CBR, as Sharaf-elDeen et al. [8] did, we use it to improve the probability of CBR to converge to the best solution. The proposed model may act both as a decision support system for less experienced clinicians and also as a second opinion for experts.

To carry out this research, we received the approval of the ethics committee at Napoleão Laureano Hospital. To define our model's cases, we collected real patient data at Napoleão Laureano Hospital, which is a reference for oncology in Brazil. We represented the case with patients' personal information, signs (i.e., objective findings that can be described by a health-care provider), symptoms (i.e., subjective complaints reported by patients), and their diagnoses. To define our model's rules and weights, we researched specialized literature [10] and interviewed a general practitioner.

For the purpose of this research, we collected data from patients with gastrointestinal cancer. More specifically, patients with the following gastric neoplasms: anal, colorectal, esophagus, and stomach. To validate our model, we developed a prototype and used the K-fold cross validation method. The final results show that our approach has increased the accuracy of the diagnosis by 22.92% when compared to using only CBR.

Background

Case-based reasoning

Case-Based Reasoning is a paradigm for solving problems that is fundamentally different from other major AI approaches. Instead of relying solely on general knowledge of a problem domain, or making associations along generalized relationships between problem descriptors and conclusions, CBR is able to use specific knowledge of previous experiments from concrete problems (cases) [11]. In CBR, a new problem is solved by reusing the solution of a previous

[1] http://www.who.int/mediacentre/factsheets/fs297/en/

[2] http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/estomago/diagnostico_profissional

[3] http://www.uicc.org/national-cancer-leadership-congress-2014

similar problem. A second important difference is that CBR is an incremental approach. This means that each time a problem is solved, this new experience is retained, making it immediately available for future problems [12].

The processes involved in CBR can be represented by a schematic cycle (Figure 1), which is comprised of the tasks of retrieving the most similar case, reuse/adapt the case to try to resolve the problem, revise the proposed solution if necessary, and retain new solution as part of a new case.

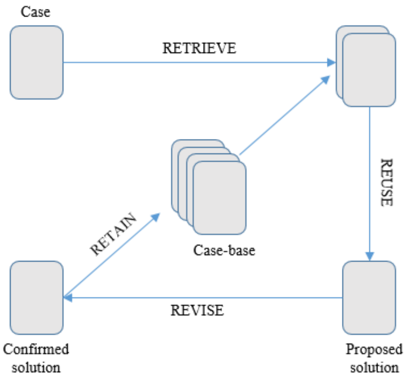


Figure 1 – CBR cycle, introduced by Aamodt and Plaza [11]

CBR can be integrated with other techniques. Marling et al. [13] present CBR integration with RBR and constraint-satisfaction problem (CSP) solving. Furthermore, they discuss CBR integration with model-based reasoning (MBR), genetic algorithms, and information retrieval. RBR was the first modality to be successfully integrated with CBR [13].

Rule-based reasoning

RBR is a methodology whose representation of knowledge is in the form of IF–THEN rule statements. Rules are patterns, so the RBR engine searches for patterns in the rules that match patterns in the data. RBR is an ideal approach for solving simple problems in which there are few rules [14]. In RBR, the problem solving complexity is directly proportional to the number of rules necessary to match the pattern of data. Furthermore, RBR lacks the ability to learn due to the difficulty of acquiring new expertise in pattern matching or new rules [14].

The basic form of a rule is the following:

```
IF <conditions>
THEN <conclusion>
```

where <conditions> represents the rule conditions, and can be connected by logical operators such as AND, OR, NOT, etc., forming a logical function. When rule conditions are satisfied, the <conclusion> is derived and the rule is said to "trigger" [9].

Methods

In medical decision support systems, the use of CBR or RBR methodologies is common [15]. In the proposed approach, we used CBR as the main reasoning process, and RBR was used to improve part of this process. The idea is that our approach can be used in a system that assists the physician in the early diagnosis of cancer. During a medical appointment, the patient tells the doctor some personal data and the symptoms that he/she is feeling. The physician will add this information to the system along with the signals perceived by the patient. The system will search in the database for the most similar case to that of the patient. Based on this result, the doctor may state the prognosis, and request tests to confirm the presence or absence of disease (Figure 2). The proposed model may act both as a decision support system for less experienced clinicians and as a second opinion for experts.

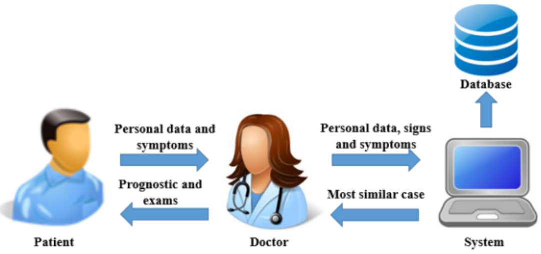
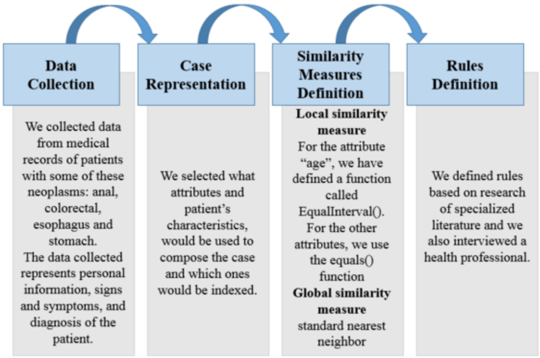


Figure 2 – System representation

Our methodology is composed of four main steps: data collection, case representation, similarity measures definition and rules definition (Figure 3). The first step is necessary for both the CBR application and also the RBR. The second and third steps correspond to two basic elements of a CBR system [12] and the fourth stage corresponds to RBR methodology. We applied RBR to define the case's attribute weights, used in the global similarity function.



We collected data from medical records of patients with some of these neoplasms: anal, colorectal, esophagus and stomach.
The data collected represents personal information, signs and symptoms, and diagnosis of the patient.

We selected what attributes and patient's characteristics, would be used to compose the case and which ones would be indexed.

Local similarity measure
For the attribute "age", we have defined a function called EqualInterval(). For the other attributes, we use the equals() function
Global similarity measure
standard nearest neighbor

We defined rules based on research of specialized literature and we also interviewed a health professional.

Figure 3 – Main steps of the methodology

In the following subsections, we present in more detail the four steps of our methodology.

Data Collection

With the approval of the Napoleão Laureano Hospital ethics committee, we collected data from the medical records of patients. We kept the privacy of their data. We collected the medical records ID from the agenda of the specialist in gastrointestinal cancer and we focused, more specifically, on patients with the following gastric neoplasms: anal, colorectal, esophagus and stomach.

The data collected corresponds to some personal information such as age, family history, signs and symptoms such as cutaneo-mucosal pallor and dysphagia, respectively, and the diagnosis of the patient. It was not possible to collect all the data we would have liked because some records had little information about the signs and symptoms of the patient. Furthermore, many of them had little digitalized infomation, hindering our collection process.

In this research, we used forty-eight cases from real patients: six cases of patients with anal cancer, six with esophageal cancer, fifteen with colorectal cancer, and twenty-one with stomach cancer.

Case Representation

To represent the cases, we used a set of [attribute – value]. An [attribute – value] system is a basic knowledge representation framework comprising a table with columns representing attributes and rows representing objects. Each table cell therefore designates the value (also known as the "state") of a particular attribute of a particular object. In this work, each object is a case and can be represented by a problem, personal data, signs and symptoms, and also by a solution, the diagnosis (Figure 4).

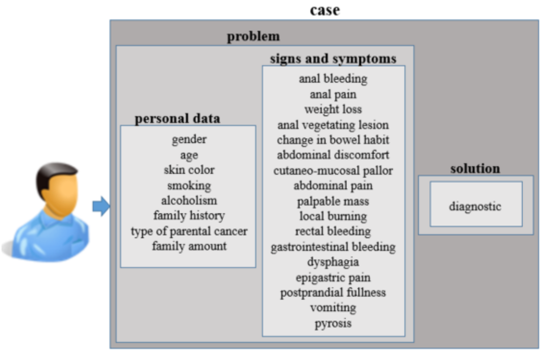


Figure 4 – Representation of the case in the proposed approach

In our approach, a case consists of twenty-six attributes, most of them boolean (i.e., they assume true or false values). Types and the respective values of the remaining attributes can be viewed in Table 1.

Table 1 – Attributes Configuration

Attribute	Type	Value
gender	String	male / female
age	Integer	1 to 110
skin color	String	white / brown / black
smoking	String	non-smoking / smoking / ex-smoking
alcoholism[4]	String	non-alcoholic / alcoholic / ex-alcoholic
family history	String	none / 1º degree / 2º degree / 3º degree
type of parental cancer	String	none / mouth / colorectal / stomach / eye / ovary / lung
family amount	String	0 / 1 / >1

Similarity Function Definition

Each attribute in a case is a piece of information about the case. An important attribute for the recovery process is called the "attribute index". Each index has a set weight, which represents the importance of that attribute in the recovery process, and is typically instantiated by the user with a value between 1 and 10.

The recovery process is based on a similarity function. In this work, we used the standard nearest neighbor method [12]:

$$\boldsymbol{sim(Q,C)} = \frac{\sum_{i=1}^{n} \boldsymbol{f(Qi,Ci)x\,wi}}{\sum_{i=1}^{n} \boldsymbol{wi}} \qquad (1)$$

This function returns the global similarity value between two cases, Q and C. Qi corresponds to the attribute value i of a new case and Ci corresponds to the same attribute i of the case recovered from the case-base. Wi is the weight of i attribute and n is the number of attributes of a case. Thus, n is equal to twenty-three, since the attributes "family history," "number of family", and "prognosis" are not used in the calculation of the equation.

In (1), "$\boldsymbol{f}$" is the function of local similarity of each attribute. All of them use the local similarity function Equal(), except for the age attribute. For this, a new function called EqualInterval(), which calculates the similarity of the attribute according to age ranges (<= 60, 61-70 and >= 71) defined by us. Thus, for this attribute, the values of 74 and 80 are coded identically.

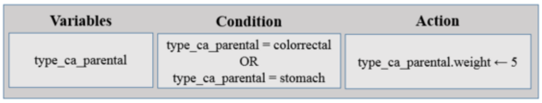
The twenty-three features used in the calculation of similarity have a default weight of "1". Some attributes have altered weights according to pre-established rules mentioned in the later section. We believe that the weights in general should be chosen by a specialist in gastrointestinal cancer in order to increase the accuracy of the system.

As we have a small case-base, the recovery method that we use is the sequential, that is, the measure of similarity is calculated for all cases of the base [12].

Rules Definition

The rules created and used in our approach were based on information extracted from a medical book [10] and the National Cancer Institute[5] (INCA) website. The weights used in the rules were decided with the help of a general practitioner. We interviewed the general practitioner and used a template to formulate the questions: "How important is the X attribute to diagnose cancer Y?", where X corresponds to the some case attribute and Y to some type of cancer. The answers were collected on a scale from 1 to 10.

1st rule: This corresponds to the patient's family history. Patients who have a relative who has or had a particular type of cancer is more likely to also have cancer. Among the types of cancer studied in this research, only colorectal and stomach cancer consider family history.

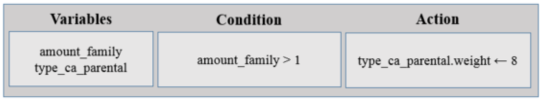


Variables	Condition	Action

type_ca_parental	type_ca_parental = colorrectal
	OR
	type_ca_parental = stomach
		type_ca_parental.weight ← 5

Figure 5 – First Rule

2nd rule: This also corresponds to the patient's family history. If the number of a patient's relatives who have cancer is greater than 1, then the patient will be more likely to also have this neoplasm.



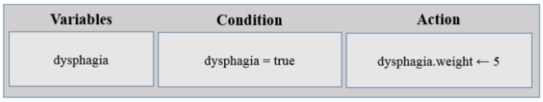
Variables	Condition	Action
amount_family type_ca_parental	amount_family > 1	type_ca_parental.weight ← 8

Figure 6 – Second Rule

3rd rule: Dysphagia is the main clinical manifestation that occurrs in patients with esophageal cancer.

[4] a chronic disorder characterized by dependence on alcohol

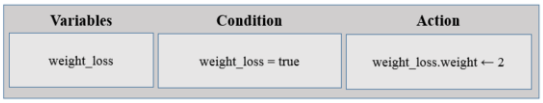
[5] National Cancer Institute http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home



Variables	Condition	Action
dysphagia	dysphagia = true	dysphagia.weight ← 5

Figure 7 – Third Rule

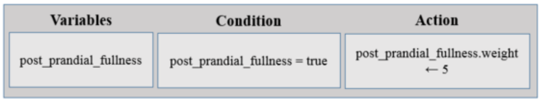
4th rule: According to the data from the literature [10], weight loss is a common feature in patients with colorectal, esophagus, or stomach cancers. However, there are cases of patients with anal cancer that also presented with weight loss.



Variables	Condition	Action
weight_loss	weight_loss = true	weight_loss.weight ← 2

Figure 8 – Fourth Rule

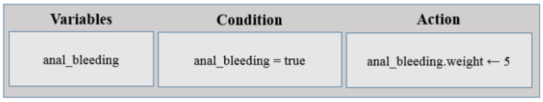
5th rule: Postprandial fullness is a feeling of stomach fullness, and is characteristic of patients with stomach cancer.



Variables	Condition	Action
post_prandial_fullness	post_prandial_fullness = true	post_prandial_fullness.weight ← 5

Figure 9 – Fifth Rule

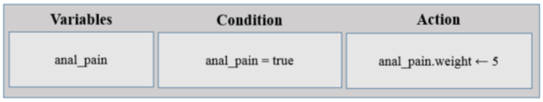
6th rule: Anal bleeding is the most common feature present in patients with anal cancer [10]. However, patients with colorectal cancer may also exhibit this symptom.



Variables	Condition	Action
anal_bleeding	anal_bleeding = true	anal_bleeding.weight ← 5

Figure 10 – Sixth Rule

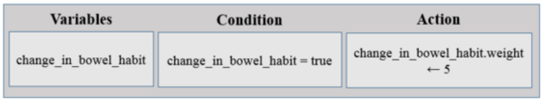
7th rule: Anal pain is a feature of patients with anal cancer, but it also appears as a symptom in patients with colorectal cancer.



Variables	Condition	Action
anal_pain	anal_pain = true	anal_pain.weight ← 5

Figure 11 – Seventh Rule

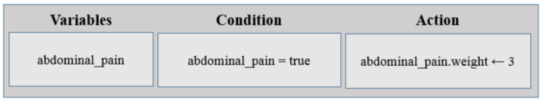
8th rule: A change in bowel habits (diarrhea or constipation) is a warning sign for colorectal cancer. There are cases of patients with stomach cancer who also have this symptom.



Variables	Condition	Action
change_in_bowel_habit	change_in_bowel_habit = true	change_in_bowel_habit.weight ← 5

Figure 12– Eighth Rule

9th rule: Abdominal pain and changes in bowel habits are symptoms of colorectal cancer. They are also often present in patients with stomach cancer.



Variables	Condition	Action
abdominal_pain	abdominal_pain = true	abdominal_pain.weight ← 3

Figure 13 – Ninth Rule

Validation Method and Results

To validate the system, we developed a prototype, and we used k-fold cross-validation [16]. As we have forty-eight cases and the number of cases for each neoplasm is a multiple of three, we split the data into three blocks, each one with sixteen cases.

We used each block as input to the system. First, we made tests without the use of rules (i.e., all attributes independent of the value had a weight of "1"). Then, we redid the tests using the rules previously mentioned. The result of the three iterations can be seen in Table 2, Table 3, and Table 4, respectively.

The diagnosis accuracies by fold and by type of cancer are shown in Table 5 and Table 6, respectively.

Table 2 – Results of the first test block

Nº	**Real diagnosis**	**Without rules**		**With rules**	
0	----------	**1° case**	**hit**	**1° case**	**hit**
1	**Anal**	anal	1	anal	1
2	**Anal**	anal	1	anal	1
3	**Colorectal**	esophagus	0	colorectal	1
4	**Colorectal**	colorectal	1	colorectal	1
5	**Colorectal**	esophagus	0	esophagus	0
6	**Colorectal**	anal	0	anal	0
7	**Colorectal**	stomach	0	colorectal	1
8	**Esophagus**	esophagus	1	esophagus	1
9	**Esophagus**	esophagus	1	esophagus	1
10	**Stomach**	stomach	1	stomach	1
11	**Stomach**	stomach	1	stomach	1
12	**Stomach**	colorectal	0	colorectal	0
13	**Stomach**	stomach	1	stomach	1
14	**Stomach**	stomach	1	stomach	1
15	**Stomach**	stomach	1	stomach	1
16	**Stomach**	stomach	1	stomach	1
sum	----------	----------	11	----------	13

Table 3 – Results of the second test block

Nº	**Real diagnosis**	**Without rules**		**With rules**	
0	----------	**1° case**	**hit**	**1° case**	**hit**
1	**Anal**	anal	1	anal	1
2	**Anal**	anal	1	anal	1
3	**Colorectal**	anal	0	anal	0
4	**Colorectal**	stomach	0	colorectal	1
5	**Colorectal**	colorectal	1	colorectal	1
6	**Colorectal**	anal	0	anal	0
7	**Colorectal**	stomach	0	stomach	0
8	**Esophagus**	stomach	0	stomach	0
9	**Esophagus**	esophagus	1	esophagus	1
10	**Stomach**	esophagus	0	stomach	1
11	**Stomach**	colorectal	0	stomach	1
12	**Stomach**	stomach	1	colorectal	0
13	**Stomach**	stomach	1	stomach	1
14	**Stomach**	colorectal	0	stomach	1
15	**Stomach**	stomach	1	stomach	1
16	**Stomach**	esophagus	0	stomach	1
sum	----------	----------	7	----------	11

Discussion and Conclusion

Many researchers have developed different approaches to predict, diagnose, and classify cancers, but in general, only a single type of cancer is discussed. In this research, we focused on four types of gastrointestinal cancer. In addition, we used rules to customize the cases. The diagnosis accuracies by fold and by type of cancer are shown in Table 5 and Table 6, respectively.

To assess if our approach increased the diagnosis accuracy compared to using only the CBR approach, we used the paired t-test with 95% confidence interval and got p-value = 0.02664, refuting the null hypothesis (H0 = The CBR performance is the same as the proposed hybrid approach). Given this, we confirmed our expectations.

The limitations of this study are related to the quantity and quality of the cases and weights. In addition, we believe that with the help of an oncologist, we could improve the rules and attribute weights. Even though the model training technique used data from a specific population group, the cross-validation results might not be enough to generate adequate data for a reliable model.

In future works, we will extend the case-base, and will seek help from a medical expert to validate the rules and the weights associated with them. Furthermore, we will discuss the remaining phases of the CBR cycle.

Table 4 – Results of the third test block

Nº	Real diagnosis	Without rules		With rules	
0	----------	1° case	hit	1° case	hit
1	**Anal**	colorectal	0	anal	1
2	**Anal**	colorectal	0	anal	1
3	**Colorectal**	stomach	0	stomach	0
4	**Colorectal**	colorectal	1	colorectal	1
5	**Colorectal**	stomach	0	anal	0
6	**Colorectal**	colorectal	1	colorectal	1
7	**Colorectal**	stomach	0	stomach	0
8	**Esophagus**	colorectal	0	esophagus	1
9	**Esophagus**	stomach	0	esophagus	1
10	**Stomach**	esophagus	0	esophagus	0
11	**Stomach**	colorectal	0	stomach	1
12	**Stomach**	stomach	1	stomach	1
13	**Stomach**	stomach	1	stomach	1
14	**Stomach**	stomach	1	stomach	1
15	**Stomach**	stomach	1	stomach	1
16	**Stomach**	stomach	1	stomach	1
sum	----------	----------	7	----------	12

Table 5 – Diagnosis accuracies by fold

	Folds			
	First	**Second**	**Third**	**Mean**
Without rules	68.75%	43.75%	43.75%	52.08%
With rules	81.25%	68.75%	75%	75%
Gain	12.50%	25%	31.25%	22.92%

Table 6 – Diagnosis accuracies by type of cancer

		Type of cancer		
	Anal	**Colorectal**	**Esophagus**	**Stomach**
Without rules	66.66%	26.66%	50%	66.66%
With rules	100%	46.66%	83.33%	85.71%
Gain	33.34%	20%	33.33%	19.05%

Acknowledgments

We would like to thank CAPES and Federal Institute of Paraiba for the financial support. We would like to extend our thanks to employees and friends of Napoleao Laureano Hospital.

References

[1] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 2001 June: 7: 673-679.

[2] Lin RH. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009 May: 47: p. 53—62.

[3] Hsu KH, Chiu C, Chiu NH, Lee PC, Chiu WK, Liu TH, et al. A case-based classifier for hypertension detection. Knowledge-Based Systems 2011 July: 24: 33-39.

[4] Chuang CL. Case-based reasoning support for liver disease diagnosis. Artificial Intelligence in Medicine 2011 June: 53: 15-23.

[5] Salem ABM, and El Bagoury BM. A Case-Based Adaptation Model for Thyroid Cancer Diagnosis Using Neural Networks. In Society, Florida Artificial Intelligence Research. Cairo: Egypt, 2003; pp. 155-9.

[6] Zubi ZS, and Saad RA. Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer. In Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Base. Tripoli: Libya, 2011; pp. 32-7.

[7] Keles A, Keles A, and Yavuz U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Expert Systems with Applications 2011: 38: 5719–26.

[8] Sharaf-elDeen DA, and Moawad IF, E. MK. A Breast Cancer Diagnosis System using Hybrid Case-based Approach. International Journal of Computer Applications 2013 June: 72: 14-9.

[9] Prentzas J, and Hatzilygeroudis I. Categorizing approaches combining rule-based and case-based reasoning. Expert Systems 2007 May: 24: 97-122.

[10] Engel CL. Tumores Gastrointestinais Writers EM, ed., 2010.

[11] Aamodt A, and Plaza E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 1994 March: 7: 39-59.

[12] Wangenheim GV, and Wangenheim V. Raciocínio Baseado em Casos Ltda EM, ed., 2003.

[13] Marling C, Sqalli M, Rissland E, Muñoz-Avila H, and Aha D. Case-Based Reasoning Integrations. AI Magazine 2002: 23: 69-86.

[14] Lim TP, Husain W, and Zakaria N. Recommender System for Personalised Wellness Therapy. International Journal of Advanced Computer Science and Applications 2013: 4: 54-60.

[15] Rossille D, Laurent JF, and Burgun A. Modelling a decision-support system for oncology using rule-based and case-based reasoning methodologies. International Journal of Medical Informatics 2005 June: 74: p. 299—306.

[16] Refaeilzadeh P, Tang L, and Liu H. Cross-Validation: Springer US, 2009.

Address for correspondence:

Renata M. Saraiva renatams@copin.ufcg.edu.br