

An Intelligent Risk Prediction System for Breast Cancer Using Fuzzy Temporal Rules

U. Kanimozhi¹ · S. Ganapathy² · D. Manjula¹ · A. Kannan³

Received: 14 March 2016/Revised: 12 March 2017/Accepted: 31 July 2018/Published online: 19 October 2018
© The National Academy of Sciences, India 2018

Abstract Online prediction of risk on breast cancer is a challenging task in the area of health care during the past decade. Since the existing statistical and data mining methods have limitations with respect to the prediction of breast cancer, there is a need for proposing more effective predictive models which can predict the breast cancer more effectively. In this paper, we propose a new intelligent online risk prediction model for predicting the breast cancer using fuzzy temporal rules more accurately. Moreover, this intelligent system determines the contributing attributes from the dataset using intelligent fuzzy temporal rules and also performs prediction by applying fuzzy rule-based classification with temporal constraints. Moreover, the rules are validated using a domain expert and the experiments conducted in this work using questionnaire, rule-based classification and consultation with domain expert have proved that the proposed system provides more

accurate results for risk prediction than the other existing systems.

Keywords Risk assessment · Prediction · Fuzzy rules · Breast cancer · Classification

Breast cancer is the most common and serious disease of women in many countries and regions of the world in the recent years. Moreover, the patients diagnosed with breast cancer are increasing in South and East Asian countries. Hence, many researchers focused their work in detection and prevention of breast cancer more efficiently. In spite of all these efforts, it is still the second leading cause of death among women after cervical cancer due to the limitations on the existing prediction systems. Therefore, it is necessary to create awareness among people and to develop more efficient disease diagnosis and prediction techniques by proposing effective techniques for feature selection and classification of the breast cancer dataset.

In the past, many researchers proposed techniques for risk analysis, identification and prediction of breast cancer based on past data and current scenario. Among them, McCarthy et al. [2] discussed the reasons behind the incidence of breast cancer in American and African women. Their work focused on the analysis of breast cancer data for risk identification. However, a system which predicts the risk using past history and present data is more important necessity. Elpiniki et al. [4] proposed a new decision support system for evaluating the familial breast cancer risk factors and to assess the risk for grading them accurately. Their model is useful in assisting the clinical oncologists for decision-making. However, other type of people who are exposed to the risk for the first time must also be considered. Majid et al. [1] developed a hybrid

✉ U. Kanimozhi
kanimozhiu.03@gmail.com

S. Ganapathy
sganapathy@vit.ac.in

D. Manjula
manju@annauniv.edu

A. Kannan
kannan@annauniv.edu

¹ Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamilnadu 600025, India

² School of Computing Science and Engineering, VIT University, Chennai, Tamilnadu 600127, India

³ Department of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai, Tamilnadu 600025, India

prediction system for detecting proteins linked with human breast and colon cancers. The main advantage of their system is that they considered both feature spaces and the development of a prediction system. However, a rule-based system which can identify the symptoms at an earlier stage and performs a temporal analysis on their data will help to predict the risks more effectively. Castanho et al. [3] investigated the use of soft computing components to generate a genetic fuzzy system for predicting the pathological stage of prostate cancer. Their model is effective in the prediction of prostate cancer by applying fuzzy rules. However, the system can be enhanced with temporal constraints to make a better prediction model. Chen et al. [6] proposed a new expert system which is developed using a combination of rough sets and support vector machines for performing breast cancer diagnosis. Their model provides a better classification accuracy due to the use of hybrid method. However, a temporal modeling with a temporal reasoning facility can provide a better system for disease prediction and treatment planning. Ganapathy et al. [5] proposed a fuzzy temporal model for medical diagnosis for identifying the diabetes more efficiently. However, their model focused on developing fuzzy rule for predicting and treating diabetes disease. Therefore, a new fuzzy temporal rule-based model is proposed in this paper which performs feature selection and classification by considering the opinion of patients, the relatives and experts through questionnaire and interaction for identifying most important features.

The architecture of the intelligent system for breast cancer diagnosis is shown in Fig. 1. It consists of nine components, namely the user interface to interact with user, questionnaire-based interaction module for user interaction, question bank interaction and interaction with decision manager, the question bank to store the questionnaire,

expert validation module for rule validation, fuzzy inference engine for rule firing, rule matching and rule execution, temporal analyzer for performing temporal constraint satisfaction, decision manager for overall control and coordination, fuzzy temporal rule base for storing the rules and breast cancer dataset.

The domain expert has given the values for identifying benign and malignant type of tumors which are in the range of 0–10. They are normalized to 0-1 using a simple normalization process.

Moreover, the fuzzy IF...THEN rules were formulated based on expert knowledge of the predictive model and associated risks. Here, we utilize the triangular membership function shown in Eq. (1).

$$\mu_A(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x-a}{b-a} & a \leq x \leq b \\ \frac{c-x}{c-b} & b \leq x \leq c \\ 0 & x > 0 \end{cases}, \tag{1}$$

where A is the fuzzy set, μ_A is the membership function, x is the universe of discourse, a is the lower limit, b is the modal value, and c is the upper limit. Table 1(a) and (b) show the input and output membership functions incurred for different linguistic variables in the dataset.

Based on the membership function, the output values are computed as follows for triangular membership function with $x = 96$, $a = 40$, $b = 80$ and $c = 120$.

$$f(x; a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right)$$

$$f(x; a, b, c) = \max\left(\min\left(\frac{96-40}{80-40}, \frac{120-96}{120-80}\right), 0\right)$$

$$= \max(\min(1.4, 0.6), 0)$$

$$= 0.6.$$

Fig. 1 Overall system architecture

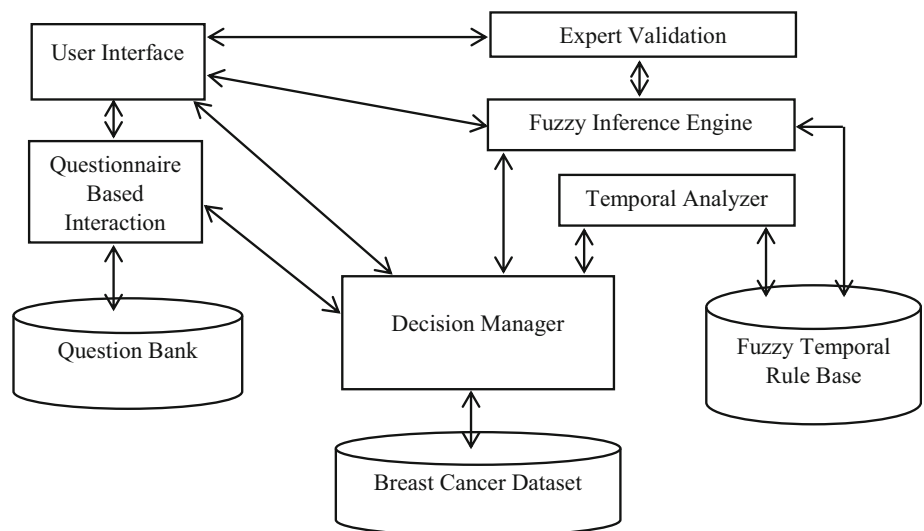


Table 1 **a** Input variable membership functions. **b** Output variable membership functions. **c** Performance evaluation on test data

(a)	
Linguistic value	Fuzzy numbers
Small/regular (very low)	(0, 0, 0.3)
Intermediate/moderate (low)	(0, 0.3, 0.5)
Large/marked (medium)	(0, 0.5, 0.8)
Present/conspicuous/diffuse (high)	(0.5, 0.8, 1.0)
Absent/thick (very high)	(0.8, 0.9, 1.0)

(b)	
Linguistic value	Fuzzy numbers
Benign	(0, 0, 0.5)
Malignant	(0, 0.5, 1.0)

(c)			
	C4.5 classification	Fuzzy temporal rule-based classification without expert validation	Fuzzy temporal rule-based classification after expert validation
Sensitivity	0.837	0.886	0.993
Specificity	0.727	0.886	0.982
Precision	0.696	0.888	0.993
Accuracy (%)	78.200	88.567	99.133

The mapping between the fuzzy variables and membership values without temporal constraints for identifying the uniformity in cell size (UCS) is shown in Fig. 2.

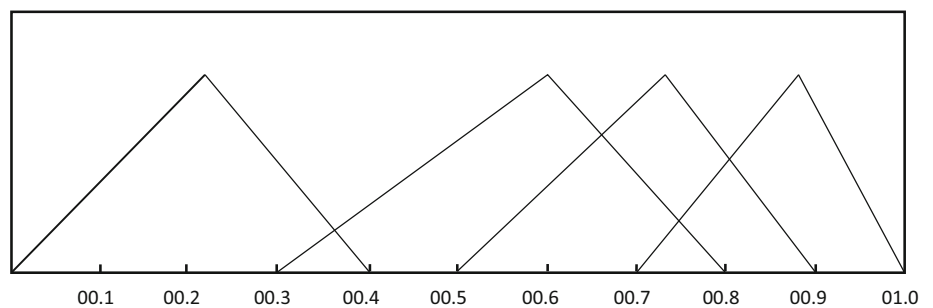
In Fig. 2, the cell sizes were measured between the interval [t1, t2], and hence further temporal constraints are not included in the inference process. The cells are classified into regular, moderate, high moderate and marked types of cells.

Figure 3(a) shows the mapping of fuzzy rules based on fuzzy variables and the classified results by applying temporal constraints on epithelial cell size (ECS) which are grouped into small, intermediate and large.

In Fig. 3(a), two time intervals, namely [t1, t2] and [t3, t4], were used as base times for the purpose of constraint formulation. The fuzzy variables are named as small, intermediate and large in the two time intervals leading to four types of fuzzy variables, namely small, intermediate, later large and large.

The proposed algorithm consists of two phases, namely the feature selection phase and the prediction phase. The steps of the proposed algorithm are as follows:

Fig. 2 Fuzzy membership diagram for uniformity of cell size (UCS)



Algorithm: Intelligent Fuzzy Temporal Rule based Prediction Algorithm

Input : Dataset, features and fuzzy temporal rules

Output : Tumor or not and Malignant/Benign

Function Feature_Selection(Number of attributes, Attribute_list)

Begin

Set Attribute_list = {}

For i = number of attributes do

Begin

1: Read attribute name from the dataset

2: Read opinions from m (100) patients, n (200) relatives and k (5) medical experts.

3: Compute score for current attribute by applying the formula

$$Decision\ Score\ (DS) = \frac{\left(\frac{W1 \times \sum_{i=1}^m P_i}{m} + \frac{W2 \times \sum_{i=1}^n R_i}{n} + \frac{W3 \times \sum_{i=1}^k E_i}{k} \right)}{W1 + W2 + W3}$$

4: Obtain threshold (Th) from expert.

5: If $DS_i > Th$ then

Attribute_list = Attribute_list U DS_i

End

Return Attribute_list

End

Function Prediction (Attribute_list, Fuzzy_Temporal_Rules, Dataset)

Begin

1: Read one record

2: Call decision tree (C4.5) algorithm [5] to form decision tree.

3: Classify the dataset into cancerous and noncancerous

4: If Cancerous Then

Apply fuzzy rules 1 to 18 using forward chaining inference

5: Generate start state and end state using current time interval [t1, t2]

6: Generate branching temporal tree with branches for [t2,t3], [t2,t4], [t3,t5], [t3,t6], [t4,t7] and [t4,t8].

7: Perform interval comparison based on the data values for all these intervals and by applying temporal reasoning operations as follows:

[t1,t2] before [t3,t4] iff [t1<t2, t2<t3 and t3<t4

[t1,t2] after [t3,t4] iff [t1<t2, t3<t4 and t4<t1

[t1,t2] overlaps [t3,t4] iff [t1<t3, t3<t2 and t2<t4.

8: Call C4.5 with these temporal constraints.

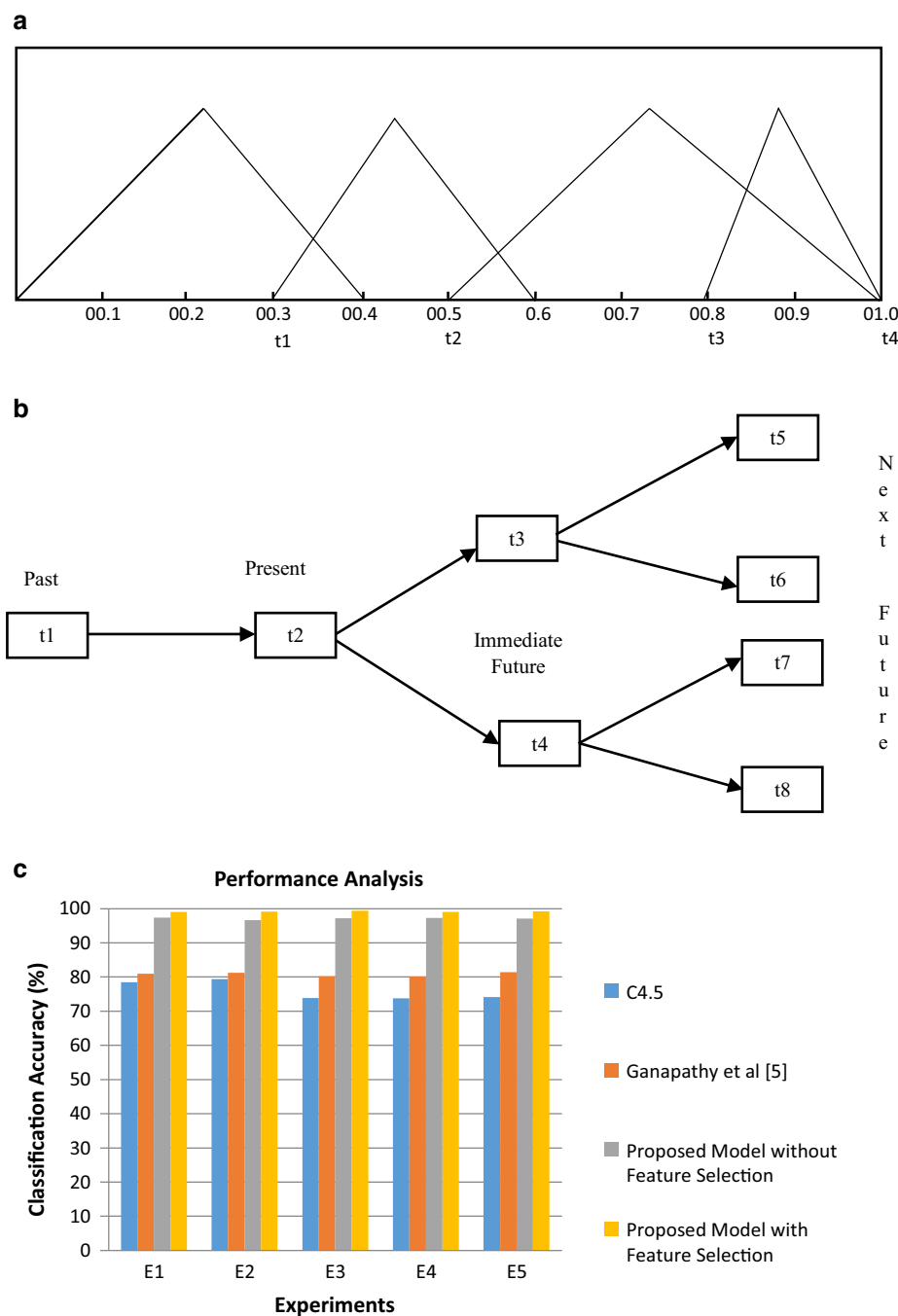
9: Fit Curve for a data series based on prime attributes used for prediction using Curve fitting method.

10: Propagate constraints by forming a constraint propagation graph.

11: Decide the type of tumor as Malignant (0.5 to 1.0) / Benign (0 to 0.5) using Threshold values.

End

Fig. 3 **a** Fuzzy membership diagram for epithelial cell size (ECS) **b** Temporal decision tree (ECS) **c** Comparative analysis of prediction accuracy



The tree used for performing prediction using temporal constraints for eight time instants, namely t_1, t_2, \dots, t_8 , is shown in Fig. 3(b).

The implementation has been carried out using the UCI Machine Learning Repository (Frank and Asuncion 2010), namely breast cancer Wisconsin original data have been used. It has 9 attributes with two possible outcomes, benign and malignant with 458 and 241 cases each. The potential of the proposed system has been thoroughly examined by using interactions with 100 patients, 200 relatives and 5

domain experts by showing the bench mark dataset and the validation rules used in this system. Finally, feature selection was performed based on expert opinion and patient and relative opinions. A sample set of rules were generated using decision tree algorithm and refined by the proposed fuzzy temporal rule-based prediction algorithm in order to detect the benign and malignant tumors.

The optimal and important attributes that are suggested by experts among the other features available are:

1. Epithelial cell size (ECS) [small—1; intermediate—2; large—3] (*small* 0–3; *intermediate* 4–7; *large* 8–10 from the scale factor).
2. Uniformity of cell size (UCS) [regular—1; moderate—2; marked—3] (*regular* 0–3; *moderate* 4–7; *marked* 8–10).
3. Bare nuclei (BN) [present—1; absent—2] (*present* 0–5; *absent* 6–10).
4. Bland chromatin (BC) [present—1; absent—2] (*present* 0–5; *absent* 6–10).
5. Normal nuclei (NN) [conspicuous—1; absent—2] (*conspicuous* 0–5; *absent* 6–10).
6. Clump thickness (CT) [diffuse—1; thick—2] (*diffuse* 0–5; *thick* 6–10).
7. Marginal adhesion (MA) [sticky cells: present—1; absent—2] (*present* 0–5; *absent* 6–10).

Hence, the fuzzy rules are generated based on the threshold values set above.

Fuzzy Rules for Breast Cancer Risk Prediction

- Rule 1 If (ECS = 1 && UCS = 1 && BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Benign
- Rule 2 If (ECS = 1 && UCS = 2&& BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Benign
- Rule 3 If (ECS = 3 && UCS = 3&& BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Malignant
- Rule 4 If (ECS = 2 && UCS = 1&& BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Benign
- Rule 5 If (ECS = 3 && UCS = 1&& BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Malignant
- Rule 6 If (ECS = 2 && UCS = 2 && BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Malignant
- Rule 7 If (ECS = 2&& UCS = 3 && BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Malignant
- Rule 8 If (ECS = 3 && UCS = 2&& BN = 1&& BC = 1 && NN = 1 && CT = 1&& MA = 1) Then Malignant
- Rule 9 If (ECS = 3&& UCS = 3&& BN = 1 && BC = 1 && NN = 1 && CT = 1 && MA = 1) Then Malignant

- Rule 10 If (ECS = 1 && UCS = 1 && BN = 2 && BC = 2 && NN = 2 && CT = 2 && MA = 2) Then Malignant

Experiments have been conducted using the dataset, and the necessary summarized result over the dataset is shown in Table 1(c).

From Table 1(c), it is observed that the fuzzy temporal rules with expert validation provide better classification accuracy than the classification using C4.5 classification algorithm and the proposed fuzzy temporal rule-based classification algorithm without expert validation. From this, it is recommended that the prediction accuracy is more only when domain expert is consulted for making suitable decisions.

Figure 3(c) represents the risk prediction accuracy based on classification accuracy with respect to breast cancer on various numbers of experimental trials.

From Fig. 3(c), it is observed that feature selection improves the classification accuracy in the proposed model when it is compared with the existing models. Moreover, the use of fuzzy temporal constraints enhances the classification accuracy.

Acknowledgements The authors thank Dr. MohamadGouse M.S. and Dr. Nafeesa. Banu M.D. for providing the expert advice forming fuzzy rules pertaining to medical diagnosis.

References

1. Majid A, Ali S, Iqbal M, Kausar N (2014) Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. *Comput Methods Progr Biomed* 113:792–808
2. McCarthy AM, Armstrong K, Handorf E, Jones M, Chen J, Demeter MB, McGuire E, Conant EF, Domchek SM (2013) Incremental impact of breast cancer SNP panel on risk classification in a screening population of white and African American women. *Breast Cancer Res Treat* 138(3):889–898
3. Castanho MJP, De Re AM, Rautenberg S, Billis A (2013) Fuzzy expert system for predicting pathological stage of prostate cancer. *Expert Syst Appl* 40:466–470
4. Papageorgiou EI, Subramanian J, Karmegam A, Papandrianos N (2015) A risk management model for familial breast cancer: a new application using Fuzzy Cognitive Map method. *Comput Methods Progr Biomed* 122:123–135
5. Ganapathy S, Sethukkarasi R, Yogesh P, Vijayakumar P, Kannan A (2014) An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization. *Sadhana* 39(2):283–302
6. Chen H-L, Yang B, Liu J, Liu D-Y (2011) Fuzzy expert system for predicting pathological stage of prostate cancer. *Expert Syst Appl* 38:9014–9022