Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 91 (2016) 927 – 934

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2016)

A Hybrid Model to Support the Early Diagnosis of Breast Cancer

Davi Carvalho*, Plácido Rogerio Pinheiro, Mirian Calíope Dantas Pinheiro

University of Fortaleza, Graduate Program in Applied Informatics
Av. Washington Soares, 1321 - Bl J Sl 30 - 60.811-905, Fortaleza, Brazil

Abstract

Decision-making is a human behavior aiming at the selection of an alternative from groups of real alternatives. Breast Cancer is top cancer of a woman both in developed and developing the world, furthermore breast cancer is the second most frequent cause of death for women in the United States as well as in Asia. Moreover, the early diagnosis is vital to a treatment with a better chance of success. Multiple variables are involved in the process of diagnosis. This study aims to use a Hybrid model to support the early diagnosis of breast cancer. We have proposed the utilization of a hybrid model structured in methodologies build a Bayesian Network to calculate the condition probability of a given person having breast cancer and to support decision (Multi-Criteria Decision Analysis - MCDA) with the goal to achieve optimal to identify the most influential attributes and have a more accurate result than using Bayesian Network alone.

Peer-review under responsibility of the Organizing Committee of ITQM 2016

Keywords: Breast Cancer, Bayesian Networks, Multicriteria Analysis

1.Introduction

With the advance of sciences applied to the health area in the last few years, there has been a considerable increase early diagnosis in the Breast Cancer of the population. Such fact can be stated based on demographic studies in developed and developing countries, which have showed a progressive and significant increase in the population in the last years. Along with this fact, a major increase in the number of health problems among the woman can be noticed. Decision making is a special activity of human behaviour aimed at the conclusion of an objective. It consists in a result of a process of choice from an identified problem or from an opportunity of creation, optimization or improve in an environment. The conclusion of a decision making process is the selection of an alternative from a group of alternatives that can be applied to solve the problem [19]. Hybrid models have been applied with success in the support of decision making in the medical area, such models were used to support the early diagnosis of Alzheimer [2, 14]], Psychological disturbs [3, 20] and Diabetes [13, 22]. The measure often taken when there is a suspicion of a patient having breast cancer is to do a mastectomy

* Corresponding author.
E-mail address: placido@unifor.br.

Peer-review under responsibility of the Organizing Committee of ITQM 2016
doi:10.1016/j.procs.2016.07.112

[23]. Mastectomy is the medical term for the surgical removal of one or both breasts. This is a very traumatic procedure and a more accurate diagnosis could prevent both the development of breast cancer and an unnecessary mastectomy. This study aims to use a Hybrid model to support the early diagnosis of breast cancer.

The study consists of two parts, the first one being Bayesian Networks and the second Multi-Criteria Decision Making, more specific the MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) method. The 1990s marked the beginning of lots researches about Bayesian Network. Since then Bayesian Network are being used with success in Medical diagnosis [8]. Following this line of research Bayesian Network has given great results in the medical and more research can give even better results. The second part of the study is aimed to still give support to the diagnosis even when the Bayesian Network fails to do so.

This application can be used to develop computational support tool to reduce the subjectivity of diagnosis. Furthermore it can be adapted to support other diseases. These methodologies can be of great help if applied in databases to go through thousands of patients with a fraction of the cost and time of a doctor doing the same task thus freeing the doctor for the cases that really need his/hers attention.

Therefore, the paper is structured as follows: section 2 gives a brief view of the Breast Cancer is performed; section 3 gives an overview of the Breast Cancer (its history, proposals, and structure), which provided the battery of tests being used in this paper; section 4 presents a step-by-step of the model structuring; moreover section 5 presents analysis and results; finally, conclusions and futures works are shown in Section 6.

2. Breast Cancer

Breast Cancer is a cancer that develops from breast tissue. It accounts for 25% of all cancers in woman [4]. In 2012 it resulted in 1.68 million cases and 522,000 deaths. The survival rates of breast cancer patients in the developed world are high. Nevertheless in developing countries the survival rates can be really poor. Since 1960s, the gold standard for early detection of breast cancer has been, and still is mammography [5]. Event though mammography has help with the early diagnosis of breast cancer thus decreasing the mortality rate it still have limitations and a considerable rate of false positives. In developed countries this is less of an issue because the patients can often, can afford further testing but in developing countries this can be a real problem with life changing consequences. One of the ways to support the science of diagnosis is a statistical approach, which aims to analyze and understand information to develop efficient methods under imprecise data [6].

For this study the Wisconsin Breast Cancer Dataset [7] was selected for the testing. It consists of 699 cases. Each entry has 9 attributes that influence the outcome, benign or malignant. These attributes are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The results of the database arrived periodically in 8 groups from January 1989 to November 1991. There are 16 instances where there is a single missing attribute. Nevertheless this missing attribute is not an issue because of the Bayesian Network ability to work with incomplete data. This database has been used in a variety of relevant studies in different areas of computer science [7].

3. Bayesian Networks Model

The Bayesian network is a graphic model that has a graph directed acyclic. The nodes and arcs of the model that represents, respectively, the universal variables $U = (A_1, A_2,...,A_n)$ and the dependencies among the variables. In the network that was constructed for the medical problem modelled, the direction of the arcs represents the relations of consequence-cause among the variables. For example, we have an arc between an arc A to an arc B, we said that and a node A represent, semantically, a cause of B and use with the name that A is a

of the fathers of B. A good approach for this problem is to use probabilities to make the best decision possible with partial information. Bayesian Networks are solid methods do deal with uncertainty.

Bayesian Network represents knowledge through a directed acyclic graph in which a directed edge represents a direct influence between the vertices. The calculates the probability of an event 'A' happening given that an event 'B' happened. Each event has a given probability that can change if another event happens. For instance there is a probability of raining in a given day but if we have the additional information that the weather is cloudy the probability of raining changes. For the theorem there are two kinds of probabilities. Unconditional probability: P (A) which is initial probability of 'A' that does not depend of any other events and there is the conditional probability: P (A|B) which is the probability o 'A' given that 'B' happened. Bayes rule is P (A|B) = P (B|A) P (A)/ P (B)

The software Netica [9] was chosen for modelling the problem and calculating the conditional probabilities because of the good results it has given in similar problems. For each entry of the database the conditional probability was calculated a checked against the real diagnosis. A successful prediction was considered if the probability found was higher than 0.9 and the real life result was the same as predicted. If the probability found was smaller than 0.9 that was considered a doubtful result. If the probability was higher than 0.9 and the real life result was not the same as predicted, this was considered a failure. Using this method using Bayesian Network alone there was a 95.7% of success, 1.7% of doubt and 2.8% of failure.

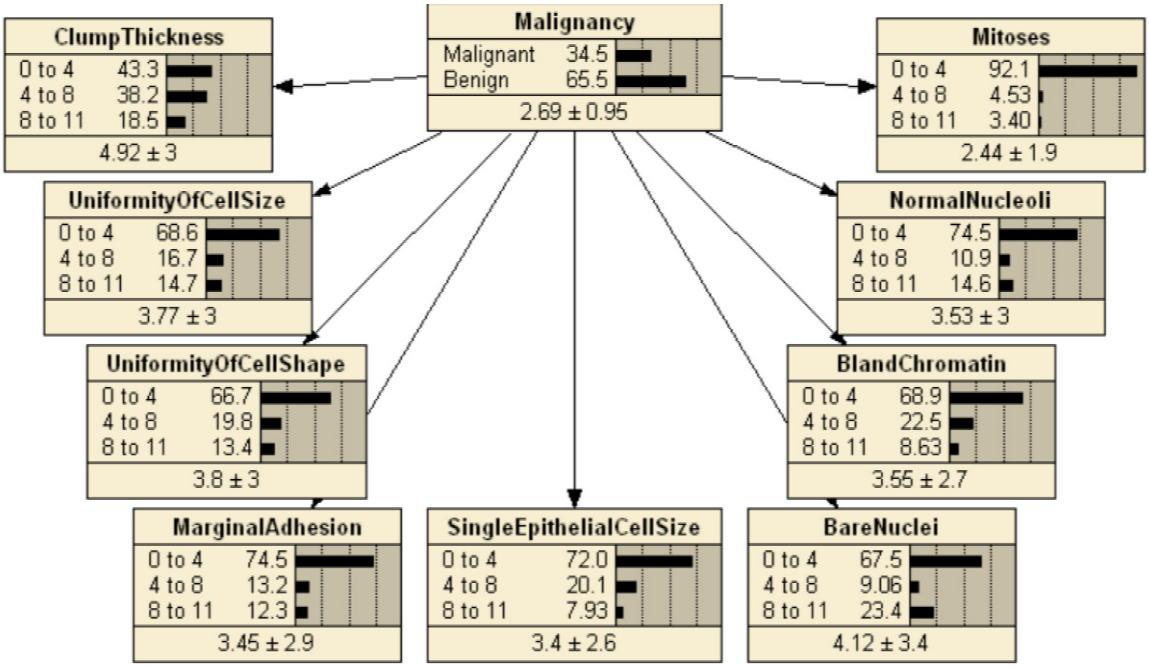

Figure 1: Bayesian Network

4. Multicriteria Model

According to [18], in decision making it is necessary to look for elements that can answer the questions raised in order to clarify and make recommendations or increases the coherency between the evolution of the process and the objectives and values considered in the environment. In this study we used the Multi-Criteria

Decision Analysis (MCDA) that is a way of looking at complex problems that are characterized by any mixture of objectives, to present a coherent overall picture to decision makers. It's important consider that the purpose of the Multicriteria methods is to serve as an aid to thinking and decision-making, but not to make the decision.

As a set of techniques, Multi-Criteria Decision Analysis provides different ways of measuring the extent to which options achieve objectives. A substantial reading on MCDA methods can be found in [15] and [16] where the authors address the definitions and the problems that are involved in the decision making process. Although not to be a simple way, the task of construction of the value tree is extremely facilitated with the aid of the Bayesian network. A great volume of information and inter-relations of the raised concepts are provided through of the network.

In this study, we used two Multi-Criteria Decision Analysis tools for to help in the resolution of the problem: Hiview and M-Macbeth for MCDA (www.m-macbeth.com). Hiview supports the appraisal and evaluation of options. It's particularly useful in helping decision makers faced with conflicting objectives to choose the best way forward. Through Hiview, we created a value tree including all decision criteria. The evaluation process is composed of the construction of judgment matrixes and constructing value scales for each Fundamental point of view (FPV) already defined. The construction of cardinal value scales will be implemented through the Macbeth methodology developed by [12]. There are times there are too many criteria and we have to find a way to decide which criteria are the most relevant. The Multi Criteria Decision-Making Method (MCDM) involves making preference decisions (such as evaluation, prioritization, selection, and so on) over the available alternatives that are characterized by multiple, usually conflicting, criteria [10]. The Bayesian Network results alone are very impressive. Nevertheless it fails for 4.3% of the times in the selected dataset. It consists of 30 people only in this dataset. This part of the study is aimed at selecting which of the attributes have more and less influence in the diagnosis using the MACBETH method implemented through the software Hiview [11].

Each attribute was divided in three groups, adding up to a total of 27 groups. The groups are as follows: a1 (Clump Thickness 1-3), a2 (Clump Thickness 4-7), a3 (Clump Thickness 8-10), b1 (Uniformity of Cell Size 1-3), b2 (Uniformity of Cell Size 4-7), b3 (Uniformity of Cell Size 8-10), c1 (Uniformity of Cell Shape 1-3), c2 (Uniformity of Cell Shape 4-7), c3 (Uniformity of Cell Shape 8-10), d1 (Marginal Adhesion 1-3), d2 (Marginal Adhesion 4-7), d3 (Marginal Adhesion 8-10), e1 (Single Epithelial Cell Size 1-3), e2 (Single Epithelial Cell Size 4-7), e3 (Single Epithelial Cell Size 8-10), f1 (Bare Nuclei 1-3), f2 (Bare Nuclei 4-7), f3 (Bare Nuclei 8-10), g1 (Bland Chromatin 1-3), g2 (Bland Chromatin 4-7), g3 (Bland Chromatin 8-10), h1 (Normal Nucleoli 1-3), h2 (Normal Nucleoli 4-7), h3 (Normal Nucleoli 8-10), i1 (Mitoses 1-3), i2 (Mitoses 4-7), i3 (Mitoses 8-10).

For this study two Fundamental Points of View (FPV) are needed. The first one (FPV1) for identifying the groups that have little influence in the diagnosis. It has a minimal value of 0 and a maximum value of 1. The second one FPV2 or identifying the groups that have a major influence in the diagnosis has a minimal value of 1.01 and a maximum value of 2. Another important step is to use descriptors that can be defined as a set of levels of impact that can be used de measure the impact of the actions in terms of each criteria [12]. Because the decisions are being made considering a database the descriptor used to measure the level of impact is based on the number of people each criteria affect. The descriptor is used as shown in the following table 1.

Table 1: Descriptors

Level of Impact	Description	Order
N2	Many people	1
N1	Few people	2

5. Analysis of the Results

The descriptor should be ordered in a way that the level with higher impact is always preferred that the level with low impact. In this case the preferred state is the one with the most number of cases. This happens to make sure that the decision is made backed up by the most number of cases possible.

The number of people in the dataset obtained FPV1 without breast cancer divided by the number of people in a given group without breast cancer. For example if there were 10 people without breast cancer and 3 people in the a1 group without breast cancer, FPV1 for a1 would be 0.33.

FPV2 was obtained by the conditional probability of the person having breast cancer given that this person is in the given group. For example FPV2 for a1 is P (malignant | a1).

According in the Table 2, the Macbeth method quantifies the relative attractiveness of options, comparing two elements at a time, using a matrix of judgment, in order to create a robust and consistent decision model. The difference of attractiveness between the groups as judged as the difference of its point of view functions.

Table 2: Attractiveness of Options

	a1	**a2**	**…**	**an**
a1	0	(a1, c) - (a2, c)	…	(a1, c) - (an, c)
a2		0	…	(a2, c) - (an, c)
…			0	…
an				0

To use the matrix on Hiview the differences of were sorted into intervals (0-null, 1-very weak, 2-weak, 3-moderate, 4-strong, 5-very strong, 6-extreme), in the table 3.

Table 3: Intervals

Macbteh	criteria model
0	0 - 0
1	0.1 - 0.18
2	0.19 - 34
3	0.35 - 0.51
4	0.52 - 0.67
5	0.68 - 0.84
6	0.85 - 1.00

A software was developed do generate the judgment values for each FPV and save the formatted results in a excel file. The results obtained for FPV1 and FPV2 were put into Hivew and had their consistency checked.

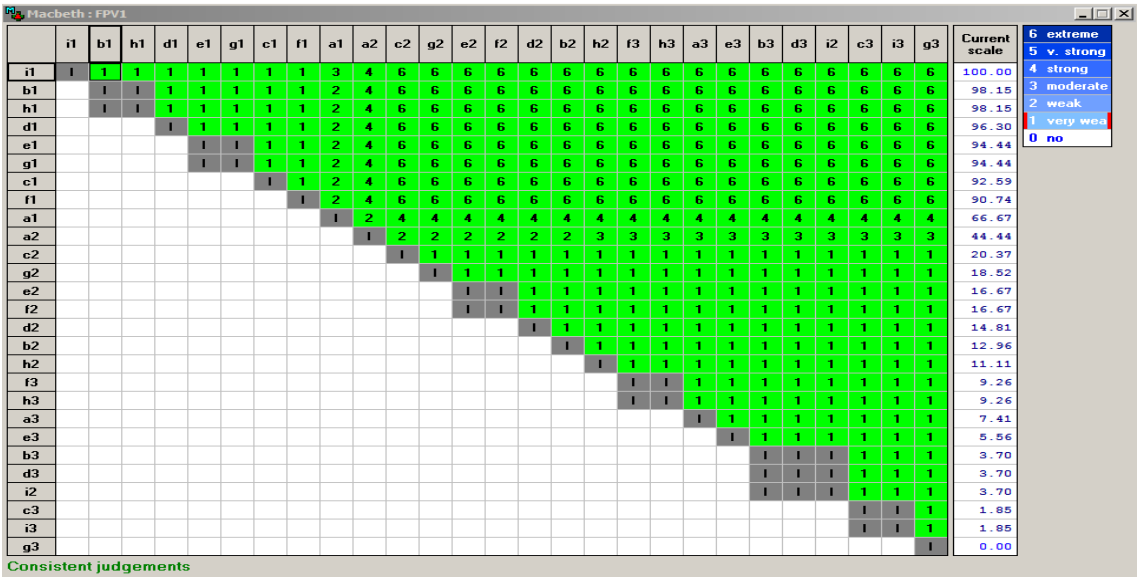

Figure 2: matrix judgment value 1

It is possible to see the difference of attractiveness between the values. It becomes clear that i1, b1 and d1 are considerable more attractive for PFV1 than g3.

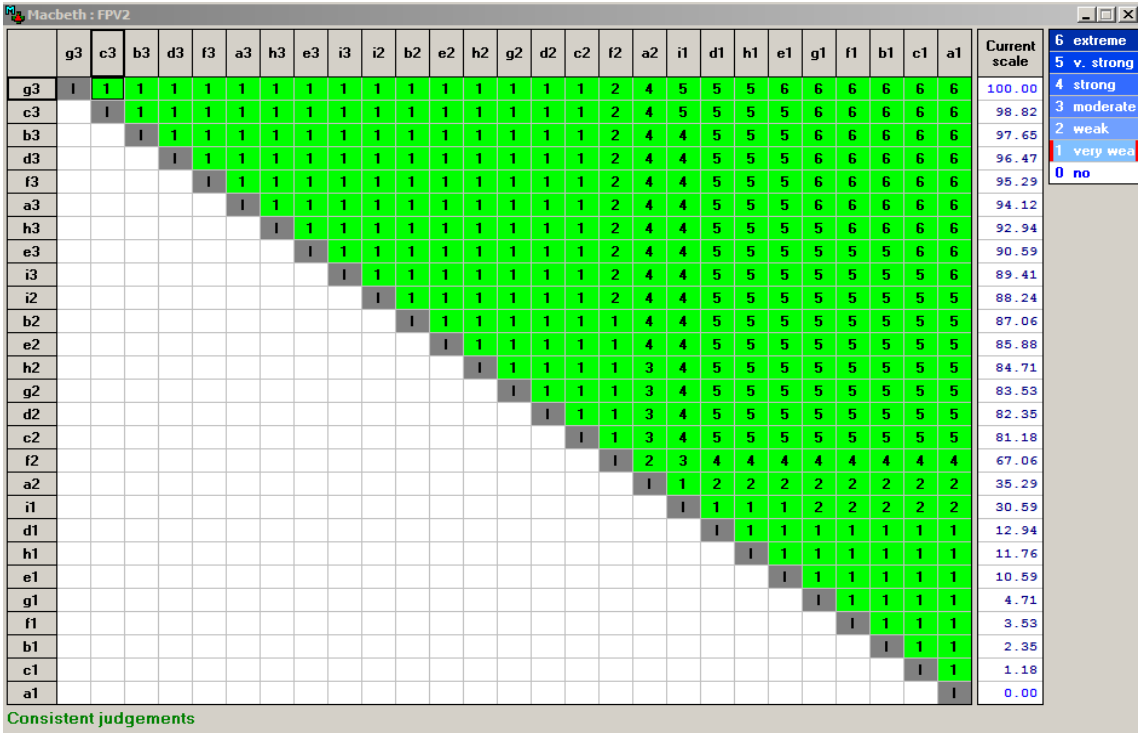

Figure 3: matrix judgment value 2

Analyzing both the matrixes it is possible to see that the most influential groups are g3 (Bland Chromatin 8-10), c3 (Uniformity of Cell Shape 8-10), b3 (Uniformity of Cell Size 8-10), d3 (Marginal Adhesion 8-10), f3

(Bare Nuclei 8-10), a3 (Clump Thickness), h3 (Normal Nucleoli 8-10), e3 (Single Epithelial 8-10) and i3 (Mitoses 8-10). And the least influential are i1(Mitoses 0-1), b1(Uniformity of Cell Size 0-1), h1(Normal Nucleoli 0-1), d1(Marginal Adhesion 0-1), e1(Single Epithelial Cell Size 0-1), g1(Bland Chromatin 0-1), c1(Uniformity of Cell Shape 0-1), f1(Bare Nuclei 0-1).

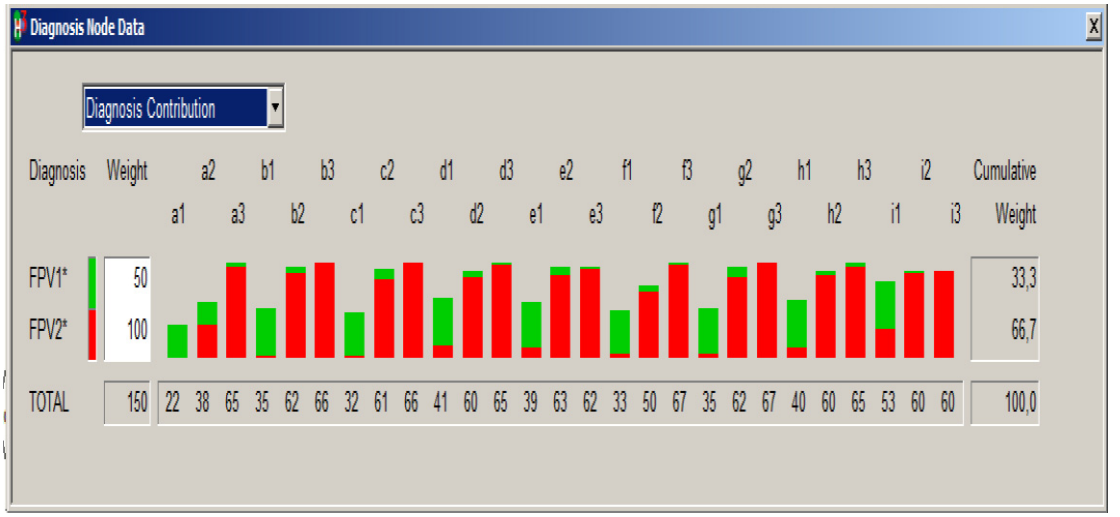

Figure 4: Node Contribution

6. Conclusion and future works

With these results it is possible to know the exact influence of each criteria. This is aimed to give support to diagnosis in cases where the Bayesian Network does not give a consistent result. This article is part of a study that has the objective to develop an automated tool to give an accurate diagnosis to general cases and support a decision to more difficult cases. Given a large enough database this tool will be able to run through thousands of cases in a few minutes and alert the cases with high chance of having breast cancer. This can be particularly useful in developing countries where there is a shortage of specialized doctors.

Furthermore with a larger database the results will be more consistent and accurate. We will continue to work to improve the results and reduce the number of people who have a late diagnosis or a unnecessary traumatic surgery. The availability of this kind of database is vital to future researches.

This methodology will be applied to another dataset with different attributes. Moreover it can also be extended to an Expert System using the Multicriteria results as entry to support even furthers the diagnosis of Breast Cancer.

Acknowledgements. The second author is grateful to National Counsel of Technological and Scientific Development (CNPq) via grants #304747/2014-9. The authors would like to thank Foundation Edson Queiroz/University of Fortaleza and Finep (Studies and Projects Financing Agency) for all the support.

References

[1] WHO, http://www.who.int/cancer/detection/breastcancer/en/ , 04/2016

[2] Castro, Ana Karoline Araujo de; Pinheiro, Placido Rogerio; Pinheiro, Mirian Caliope Dantas; Tamanini, Isabelle. Towards the Applied Hybrid Model in Decision Making: A Neuropsychological Diagnosis of Alzheimer's Disease Study Case. INT J COMPUT INT SYS, 4, 89-99, 2011.

[3] Nunes, Luciano Comin; Pinheiro, P. R.; Cavalcante, T. P.; Pinheiro, M. C. D. . Handling Diagnosis of Schizophrenia by a Hybrid Method. Computational and Mathematical Methods in Medicine (Print), 1-13, 2015.
[4] Steward, W. Bernard; Wild, Christopher P.; World Cancer Report 2014. World Health Organization. 2014.
[5] Chaurasia, S., Chakrabarti, P., Chourasia, N.; An Application of Classification Techniques on Breast Cancer Prognosis, International Journal of Computer Applications, 59(3): 6-10, 2012
[6] Watthayu, W., Peng, Y. A Bayesian network based framework for multi-criteria decision making, In: Proceedings of the 17th International Conference on Multiple Criteria Decision Analysis, 2004.
[7] http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
[8] Le Zhang, Yuan Gao, Balmatee Bidassie, Vincent G. Duffy, Application of Bayesian Networks in Consumer Service Industry and Healthcare, Lecture Notes in Computer Science, 8529, 484-495, 2014
[9] Netica Application, Norsys Software Corp., http://www.norsys.com, 1998.
[10] Hwang, C.L., Yoon, K., (1981) Multiple Attribute Decision Making: Methods and Applications, Springer-Verlag, Berlin.
[11] Hiview, Catalyse Consulting, http://www.catalyzeconsulting.com/index.php/software/hiview3/
[12] Bana e Costa, C.A., De Corte, J.M., Vansnick, J.C. MACBETH, International Journal of Information Technology and Decision Making, 11(2), 359-387, 2012.
[13] Menezes, A. C, Pinheiro, Placido. R, Pinheiro, M. C. D, Cavalcante, T. P. Towards the Applied Hybrid Model in Decision Making: Support the Early Diagnosis of Type 2 Diabetes. In: 3rd International Conference on Information Computing and Applications (ICICA 2012), Lecture Notes in Computer Science, 7473, 648–655, 2012.
[14] Castro, A.K.A. de, Pinheiro, P.R., Pinheiro, M.C.D.: An Approach for the Neuropsychological Diagnosis of Alzheimer's Disease: A Hybrid Model in Decision Making. Springer Berlin. Lecture Notes in Computer Science vol. 5589
[15] PINHEIRO, P.R., SOUZA, G.G.C.: A Multicriteria Model for Production of a Newspaper. Proc: The 17th International Conference on Multiple Criteria Decision Analysis, Canada (2004) 315–325.
[16] Goodwin, P.,Wright, G.: Decision Analysis for Management Judgment, John Wiley and Sons, Chicester (1998).
[17] Russell, S., and Norvig, P., (1995) Artificial Intelligence: A modern approach. Prentice Hall
[18] Bana e Costa, C.A. Structuration, construction et exploitation dún modèle multicritère d'aide à la décision. Thèse de doctorat pour l'obtention du titre de Docteur em Ingénierie de Systèmes - Instituto Técnico Superior, Universidade Técnica de Lisboa, 1992.
[19] Gomes, Luiz Flávio Autran, Gomes, C. F. S. Tomada de Decisão Gerencial Enfoque multicritério. 5. ed. São Paulo: Atlas, 2014.
[20] Nunes, Luciano Comin; Pinheiro, Plácido Rogerio; Pequeno, Tarcisio Cavalcante; Pinheiro, Mirian Caliope Dantas; Towards an Applied to the Diagnosis of Psychological Disorders. Advances in Experimental Medicine and Biology, 696, 573-580, 2011.
[21] AMIN-NASERI, M.R., NESHAT, N. An Expert System Based on Analytical Hierarchy Process for Diabetes Risk Assessment (DIABRA). Lecture Notes in Computer Science, 6729, 252-259, 2011.
[22]MENEZES, A. C, PINHEIRO, P. R, PINHEIRO, M. C. D, CAVALCANTE, T. P. Towards the Applied Hybrid Model in Decision Making: Support the Early Diagnosis of Type 2 Diabetes. In: 3rd International Conference on Information Computing and Applications (ICICA 2012), LNCS 7473, pp. 648–655, 2012.
[23] Robert Hodgson, Sylvia H. Heywang-Köbrunner, Susan C. Harvey, Mary Edwards, Javed Shaikh, Mick Arber, Julie Glanville, Systematic review of 3D mammography for breast cancer screening, 27, 52-61, 2016