*Advances in Intelligent Systems Research, volume 134*
2nd International Conference on Control, Automation, and Artificial Intelligence (CAAI 2017)

# ISE: An Algorithm to Screen out the High-Risk Group of Breast Cancer

Fei Chang[1] and Rui Wang[1, 2,*]
[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China;
[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083, China;
*Corresponding author

***Abstract*—In the study for the prevention and the control of breast cancer, using mobile devices to design questionnaires and applying models to screen out high-risk groups has great significance. However, some existing models are not suitable for the Asian women. In this work, we proposed an ISE algorithm to derive the conditions of breast cancer by analyzing the relationship among the influential factors of breast cancer and the strength of the relationship. Based on this algorithm, we can obtain the high-risk groups of breast cancer and finally construct a model for the prevention and the control of breast cancer, especially for Asian women.**

***Keywords-mobile and sever; data analysis; model of high -risk groups in breast cancer***

## I. INTRODUCTION

Studies indicated that if breast cancer could be prevented and diagnosed early, it could have a significant impact on the incidence of breast cancer [1]. Therefore, working on the prevention and the control of breast cancer has a great positive impact. In order to diagnose and treat breast cancer early, many researchers have been working on building a model for prevention and control of breast cancer. Currently, there are numerous widely used models, especially the Gail model [2], but they are not suitable for Asian women. In order to build a better prevention and control model of breast cancer for Asian women, we proposed an algorithm named ISE to screen out high -risk groups of breast cancer, as shown in FIGURE I. In this study, individual information of breast cancer will be collected through many mature information statistics Apps which run on mobile terminal equipment [3]. But the process of building the model is difficult due to the following challenges:

1. To better screen out the high-risk groups of breast cancer ,we need to know the main influential factors of breast cancer at first. Once the factors are determined, we can go on to the next step.

2. After obtaining the influential factors of breast cancer, finding the relationship among these factors, and the relationship between these factors and breast cancer are the main tasks we should complete. We are unable to achieve this through prior knowledge, which makes it a challenge in this study.

3. At the end of building the model, how to get the range of values of each factor that can cause breast cancer is a major problem. Only with this information can we get the conditions for breast cancers to screen out the high-risk groups.

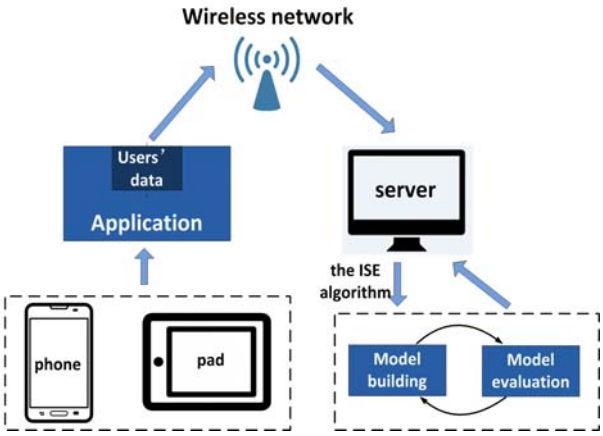


FIGURE I. THE FRAMEWORK OF OUR MAIN WORK

## II. ISE ALGORITHM

### A. Input Data

There are many existing Apps to collect data concerning breast cancer with the type of questionnaire such as Sojump, which runs on users' mobile phones or tablets. Users are data provider and data consumer at the same time. We referred to the prior medical knowledge to complete the first challenge that we listed above[4]. The main factors we acquired that influence breast cancer are shown in Figure II. Here, we combined the main factors with real statistical data as the input data of the ISE algorithm.

| | |
|---|---|
| A | age |
| B | weight |
| C | the first child-bearing age |
| D | breast benign tumor |
| E | times of abortion |
| F | satisfaction of current life |
| G | living environment |
| H | family history of breast cnacer |

FIGURE II. THE MAIN INFLUENCE FACTORS OF BREAST CANCER

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

*B. Algorithm Description*

To meet the challenges mentioned in part one, we proposed an algorithm named ISE which was based on the principle of Data Analysis Technology (DAT)

Based on the input data, we could determine whether there was a relationship between factors by computing the their mutual information [5]. We set the node A and node B which represent two main factors as an example, their mutual information I (a;b) was given by:

$$I(A;B) = H(A) - H(A \mid B) \qquad (1)$$

Here, H (A) represents the information entropy of node A, H(A|B) represents the conditional entropy of node A. If the value of $I(A;B)$ reaches the threshold we defined before, it shows that the two nodes are interdependent and they will be connected with an undirected edge. Then we combine it with the existing simulated annealing algorithm [6] to give a direction of each edge. In detail, we use the simulated annealing algorithm to add、delete edges and combine whether there will be a ring with changing the direction of them to select the direction of each edge. Finally, we will conclude the relationship about the eight main factors of breast cancer.

It is worth mentioning that we suggested combining mutual information with the simulated annealing algorithm. In comparison with the method that only used the simulated annealing algorithm, this method reduced searching space on the side and it improved the efficiency of the results. It is shown in TABLE I.

TABLE I. THE COMPARISON OF TWO METHODS

| Two Methods / Contrast index | Only The Simulated Annealing Algorithm | Combing Mutual Information With The Simulated Annealing Algorithm |
|---|---|---|
| ***Computing limit*** | All edges | The edges that reach the defined threshold |
| ***Computing time*** | About half a day | About five hours |
| ***Computing efficiency*** | low | high |
| ***Computing result*** | Low accuracy | High accuracy |

After obtaining the relative relationship between the main factors, we continued with the principle of parameter learning on Big Data [7] to determine the incidence of beast cancer. At first, we attributed the input data to a prior parameter $\theta$, and its expected value is given by :

$$L(\theta) = \sum_{X} \log P(Y, X \mid \theta) P(X \mid Y, \theta') \qquad (2)$$

It is considered that the statistical data may not complete. In the above formula, X represents missing data, Y represents the observable data, $P(X \mid Y, \theta')$ represents the conditional probability distribution of X according to Y and the current parameter $\theta'$. The parameter $\theta'$ can be obtained by maximizing the $L(\theta)$ :

$$\theta' = \arg\max L(\theta) \qquad (3)$$

Then, let the computed $\theta'$ to be $\theta$ again and iterate this formula until the result is converged. From this, we can find the range of values that will achieve the highest incidence of breast cancer and then screen out the high-risk population of breast cancer by comparing with the personal information provided by users. Thus, a model for prevention and control of breast cancer is built based on the research results. Meanwhile, we can verify the reliablility of the influence factors of breast cancer chose by prior knowledge by the re-search result. The relationship between the main factors and the incidence of breast cancer is shown in TABLE II, where 'Y' indicates there is a dependency relationship between two factors that are represented as letter nodes. The last line represents the probability of each factor causing breast cancer.

TABLE II. THE RELATIONSHIP AND ITS STRENGTH BETWEEN FACTORS AND BREAST CANCER

| Influential Factors | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| ***A*** | | Y | Y | Y | Y | Y | Y | Y |
| ***B*** | Y | | Y | Y | | | Y | Y |
| ***C*** | Y | Y | | Y | Y | | Y | Y |
| ***D*** | Y | Y | Y | | Y | Y | Y | Y |
| ***E*** | Y | | Y | Y | | Y | Y | Y |
| ***F*** | Y | | | Y | Y | | | |
| ***G*** | Y | Y | Y | Y | Y | | | Y |
| ***H*** | Y | Y | Y | Y | Y | | Y | |
| ***I*** | 88.16% | 97.24% | 97.31% | 97.48% | 98.88% | 98.27% | 98.67% | 98.54% |

## III. CONCLUSION

We adopt the method of knowledge driven, selecting the main factors that affecting breast cancer according to the experience of experts. Then, we propose an algorithm named ISE to explore the relationship among these factors and the relationship and its strength between these factors and breast cancer. Thus, we can obtain the high-risk groups of breast cancer and finally construct a model for the prevention and the control of breast cancer, especially for Asian women. In the process of building this model, the statistical data is multidimensional and large in quantity due to the vast number

of users. Therefore, the computational performance of the algorithm needs take into consideration. The model we build is evaluated by the real breast cancer patient data. It is found that the model has a good analytical performance and a good application in screening out the high-risk population of breast cancer. After evaluating the model, we apply this model to the initial App by client/server (C/S) mode and any user can know about her own breast health, as shown in Figure III.

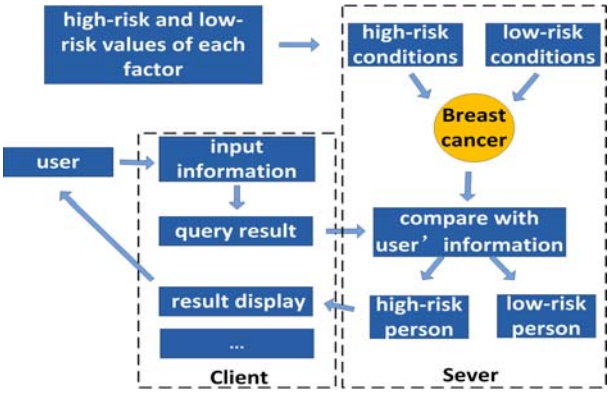


FIGURE III. THE MODEL APPLYS TO THE USERS' MOBILE DEVICES

Sever: When users log on this App through their mobile phones or tablets and fill in their personal information about breast cancer, the data will be uploaded to the sever. Then the sever compares the data with the model we built to get the result and return it to the client. This process can be completed in 4 seconds.

Client: Considering the difference in the performance of mobile phones [8], the analytic result can be shown in a user interface in 2 seconds on average. Then, users can know whether they are at high risk of having breast cancer quickly. At the same time, they can judge their own breast health according to the conditions of breast cancer.

We have obtained some preliminary research results in this study. We believe that improving the data aggregation process running on mobile devices will further improve the accuracy. The current study is in the preliminary stages but it holds promise for the future, when it may be applied to many applications in this area.

## REFERENCES

[1] DeSantis C, Ma J, Bryan L, et al. Breast cancer statistics, 2013[J]. CA: a cancer journal for clinicians, 2014, 64(1): 52-62.

[2] Tice J A, Cummings S R, Ziv E, et al. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population[J]. Breast cancer research and treatment, 2005, 94(2): 115-122.

[3] Shafagh H, Hithnawi A, Dröscher A, et al. Talos: Encrypted Query Processing for the Inter of Things[C]//Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems. ACM, 2015: 197-210.

[4] Liyuan Liu. Preliminary study for risk factors of Breast cancer and screening models of high-risk groups[D]. Jinan: Shandong University, 2010.

[5] Chen X W, Anantha G, Lin X. Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(5): 628-640.

[6] Xavier-de-Souza S, Suykens J A K, Vandewalle J, et al. Coupled simulated an- nealing[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010, 40(2): 320-335.

[7] McLachlan G, Krishnan T. The EM algorithm and extensions[M].John Wiley & Sons,2007.

[8] Salem A, Nadeem T. Colphone: A smart-phone is just a piece of the puzle[C] //Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubi- quitous Computing: Adjunct Publication. ACM, 2014:263-266