Computers in Biology and Medicine 92 (2018) 168–175

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach

Shuo Liu [a,1], Jinshu Zeng [b,1], Huizhou Gong [c], Hongqin Yang [d,**], Jia Zhai [e], Yi Cao [f], Junxiu Liu [g], Yuling Luo [h], Yuhua Li [i], Liam Maguire [g], Xuemei Ding [a,g,*]

[a] Faculty of Mathematics and Informatics, Fujian Normal University, Qishan Fuzhou, 350108, China
[b] Department of Ultrasonic Medical, The First Affiliated Hospital of Fujian Medical University, Fuzhou, 350005, China
[c] College of Foreign Languages, Fujian Normal University, Cangshan Fuzhou, 350007, China
[d] Fujian Provincial Key Laboratory for Photonics Technology, Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Cangshan Fuzhou, 350007, China
[e] Business School, University of Salford, Manchester, M5 4WT, UK
[f] Department of Business Transformation and Sustainable Enterprise, Surrey Business School, University of Surrey, Surrey, GU2 7XH, UK
[g] Faculty of Computing, Engineering and Built Environment, Ulster University, Londonderry, BT48 7JL, UK
[h] Faculty of Electronic and Engineering, Guangxi Normal University, Guilin, 541004, China
[i] School of Computing, Science and Engineering, University of Salford, Manchester, M5 4WT, UK

ARTICLE INFO

Keywords:
Clinical decision support
Data modelling
Bayesian network
Quantitative analysis
Diagnostic contribution
Breast cancer diagnosis

ABSTRACT

Background: Breast cancer is the most prevalent cancer in women in most countries of the world. Many computer-aided diagnostic methods have been proposed, but there are few studies on quantitative discovery of probabilistic dependencies among breast cancer data features and identification of the contribution of each feature to breast cancer diagnosis.

Methods: This study aims to fill this void by utilizing a Bayesian network (BN) modelling approach. A K2 learning algorithm and statistical computation methods are used to construct BN structure and assess the obtained BN model. The data used in this study were collected from a clinical ultrasound dataset derived from a Chinese local hospital and a fine-needle aspiration cytology (FNAC) dataset from UCI machine learning repository.

Results: Our study suggested that, in terms of ultrasound data, cell shape is the most significant feature for breast cancer diagnosis, and the resistance index presents a strong probabilistic dependency on blood signals. With respect to FNAC data, bare nuclei are the most important discriminating feature of malignant and benign breast tumours, and uniformity of both cell size and cell shape are tightly interdependent.

Contributions: The BN modelling approach can support clinicians in making diagnostic decisions based on the significant features identified by the model, especially when some other features are missing for specific patients. The approach is also applicable to other healthcare data analytics and data modelling for disease diagnosis.

1. Introduction

Breast cancer is the most prevalent cancer in women around the world. It has been reported that approximately 1.3 million women worldwide have been diagnosed with breast cancer since 2011, and approximately 465,000 women die from breast cancer each year [1]. In China, 214,360 women had died from breast cancer by 2008. It has been estimated that the number of Chinese women with breast cancer will reach 2.5 million by 2021 [2]. According to a report published by the Chinese National Cancer Centre in 2017, breast cancer is the most common cancer in Chinese women. Following lung, stomach, liver, oesophageal and colorectal cancers, breast cancer is the sixth largest killer in small- and medium-sized cities, with a mortality rate of 8.44% and 9.59%, respectively, while the mortality rate from breast cancer in large-sized cities is 12.78%, making it the fifth most common cause of death among all cancer types in Chinese women [3]. Due to the rapid

* Corresponding author. Faculty of Mathematics and Informatics, Fujian Normal University, Qishan Fuzhou, 350108, China.
** Corresponding author. Fujian Provincial Key Laboratory for Photonics Technology, Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Cangshan Fuzhou, 350007, China.
E-mail addresses: hqyang@fjnu.edu.cn (H. Yang), xuemeid@fjnu.edu.cn, x.ding@ulster.ac.uk (X. Ding).
[1] These authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2017.11.014
Received 23 May 2017; Received in revised form 15 November 2017; Accepted 15 November 2017

increase in the number of breast cancer patients, early identification of women at risk of developing breast cancer is currently an international priority [4].

In order to improve diagnostic accuracy and help domain experts to make more effective decisions, many computer-aided diagnosis (CAD) systems have been developed [5–7]. They provide new computational algorithms combined with domain knowledge to support clinical diagnosis. Zeng et al. [8–10] proposed different nonlinear state-space models for lateral flow immunoassay, which have been commonly used in clinical diagnosis. In clinical medicine, breast cancer could also be diagnosed via several different techniques, such as ultrasound, fine-needle aspiration cytology (FNAC) and magnetic resonance imaging (MRI) scanning.

Other CAD algorithms in [11–14] were also proposed to detect breast cancer. For example, Eltoukhy et al. [12] proposed a feature extraction method based on a statistical *t*-test for breast cancer diagnosis from a digital mammogram. They used wavelet and curvelet methods to transform digital mammography data into vector coefficients and then employed a support vector machine (SVM) algorithm for breast cancer diagnosis. As a result, the highest diagnostic accuracy based on wavelet and curvelet coefficients was 96.56% with 1238 features and 97.30% with 5663 features [12].

As a well-established probabilistic classifier, Bayesian network (BN) analysis has been used widely for data analytics and data modelling in many healthcare area, such as psychotic depression [15], Alzheimer's disease [16], heart disease [17] and social anxiety [18] as well as breast cancer. Wang et al. [19] proposed a three-layer BN for earlier diagnosis of breast cancer. They assessed the performance of the BNs constructed based on non-imaging features, imaging features and both. They found that the BN built on both imaging and non-imaging features performed well and that imaging features dominated BN performance. In 2007, Nicandro et al. [20] evaluated the performances of seven BN classifiers (i.e. Naïve Bayes classifier, Bayes-N, MP-Bayes, Greedy, MP-Bayes + Greedy, PC which is a procedure contained the Tetrad, and a CBL2 algorithm in Power Constructor, which is a software package containing CBL1 and CBL2 algorithms) for breast cancer diagnosis based on fine-needle aspiration from a breast lesion collected by a single observer and multiple observers. They found that the classifiers learnt from different data performed differently, which indicated that the observations would impact the breast cancer diagnostic result. Furthermore, in 2009 [21], Nicandro and his team made use of two decision trees and four different BNs for breast cancer diagnosis. Their study discovered interobserver variability in breast cancer cytodiagnosis, indicating that different observers would focus on different perspectives while making a diagnostic decision. Kalet et al. [22] designed a Bayesian model to detect a misdiagnosis made at the initial diagnostic stage of a disease such as lung, brain or female breast cancer. The BN model they designed produced a better AUC (0.98) than a decision made by clinical experts (0.90).

Additionally, BN was also used in other studies of breast cancer, such as risk factor estimation [23], and causal interaction detection [24]. Nicandro et al. [25] employed a score-based BN approach to estimate the power of thermography for breast cancer diagnosis. The BNs were learned using Naïve Bayes, hill-climber and repeated hill-climber algorithms with a minimum description length (MDL) metric. The BN learned by a repeated hill-climber algorithm provided the best accuracy for both cancer and non-cancer diagnosis (75.50 ± 6.99%) and sensitivity of cancer diagnosis (94%). Furthermore, their obtained BN identified five important features for breast cancer diagnosis: 1C (hottest point in only one breast), f unique (total number of hottest points), thermovascular network (number of veins with the highest temperature), curve pattern and asymmetry (temperature difference between the right and left breasts).

Although a BN modelling approach has been used for breast cancer diagnosis, a report on quantitative analysis among different breast cancer features, which is critically important for clinical decision making, is lacking. As a result, some researchers might ignore the relationships between different features, which may lead to a high misdiagnosis rate [19,26]. A clearly explained BN in a medical area can increase the understanding of disease pathology and provide valuable decision-making assistance to domain experts. This paper employed a BN modelling approach to discover the probabilistic relationships between different data features of breast cancer. We also analysed the contribution of each feature to breast cancer diagnosis. The data were focused on ultrasound and FNAC examinations obtained from The First Affiliated Hospital of Fujian Medical University, China and the Breast Cancer Wisconsin Dataset (BCWD) of the UCI machine learning repository [27].

BN modelling can be deconstructed into two sub-processes: structure learning and parameter learning. In this study, a K2 learning algorithm [28] with an MDL score metric was used to learn the BN structure. Our reasons for using a K2 algorithm were the following: 1) K2 is the most commonly used algorithm for BN structure learning [29], 2) K2 is relatively easy to implement [29], 3) K2 only needs to consider a subset of a directed acyclic graph (DAG) and can quickly find the variable with the local maximal score [30] and 4) a K2 algorithm makes good use of experts' knowledge to learn the BN structure.

The contributions of this study are 1) we discovered the most important features which can provide uninitiated observers and doctors objective and quantitative guidance to focus on specific features for early breast cancer diagnosis. 2) We analysed the probabilistic dependencies among different data features and identified the strength of the dependency, which can assist the domain experts in making a quantitatively accurate diagnosis, even using fewer available features. A focus on different features by different observers [21] may cause them to miss some important features, which can significantly influence diagnostic results. The above two contributions are helpful in decreasing the misdiagnosis rate. 3) Our study showed a potential translational application of the BN modelling approach to the breast cancer care pathway.

The remainder of this paper is organized as follows. Section 2 provides the basic theory of BN in detail, as well as a brief introduction about the technique of BN visualization. Section 3 presents the experimental results based on two real-world datasets. Section 4 discusses the results and evaluates the BN modelling approach in comparison with other methods. Finally, Section 5 concludes this paper and discusses potential extensions of the method in future work.

2. Methods and materials

Numerous approaches have been developed to support breast cancer diagnosis. BN analysis has been used widely to improve diagnostic accuracy and to discover probabilistic relationships among features and the influence of joint probability distribution inference.

2.1. Bayesian network

A BN represents a domain which explicitly provides a set of variables belonging to a specific domain and visualizes the relationships between the variables [31]. It can successfully represent uncertain knowledge in various fields [30]. A BN is usually represented using a DAG, where V denotes a set of nodes made up by a set of variables, and E denotes a set of edges between the nodes in V. No cycles are present in the DAG [32]. Each edge is directly linked from one node to another, and it indicates that the corresponding two nodes are mutually dependent. Otherwise, nodes are independent if there is no link between them.

Consider a given dataset D containing a set of variables $X = \{X_1, X_2, ..., X_n\}$, the joint probability distribution on X, $p(X_1, ..., X_n)$, is defined as

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i|\pi_i) \quad (1)$$

where π_i is the set of parents of X_i.