Artificial Intelligence in Medicine 57 (2013) 73–86

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

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

On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks

Marina Velikova[a,*], Peter J.F. Lucas[a,b], Maurice Samulski[c], Nico Karssemeijer[d]

[a] Institute for Computing and Information Sciences, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[b] Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[c] Datec Norge AS, Mediahuset, N-7075 Tiller, Norway
[d] Department of Radiology, Radboud University Nijmegen Medical Centre, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 12 January 2012
Received in revised form 26 October 2012
Accepted 8 December 2012

Keywords:
Bayesian networks
Data discretisation
Structure learning
Computer-aided detection
Medical image interpretation
Mammography

ABSTRACT

Objectives: To obtain a balanced view on the role and place of expert knowledge and learning methods in building Bayesian networks for medical image interpretation.
Methods and materials: The interpretation of mammograms was selected as the example medical image interpretation problem. Medical image interpretation has its own common standards and procedures. The impact of these on two complementary methods for Bayesian network construction was explored. Firstly, methods for the discretisation of continuous features were investigated, yielding multinomial distributions that were compared to the original Gaussian probabilistic parameters of the network. Secondly, the structure of a manually constructed Bayesian network was tested by structure learning from image data. The image data used for the research came from screening mammographic examinations of 795 patients, of whom 344 were cancerous.
Results: The experimental results show that there is an interesting interplay of machine learning results and background knowledge in medical image interpretation. Networks with discretised data lead to better classification performance (increase in the detected cancers of up to 11.7%), easier interpretation, and a better fit to the data in comparison to the expert-based Bayesian network with Gaussian probabilistic parameters. Gaussian probability distributions are often used in medical image interpretation because of the continuous nature of many of the image features. The structures learnt supported many of the expert-originated relationships but also revealed some novel relationships between the mammographic features. Using discretised features and performing structure learning on the mammographic data has further improved the cancer detection performance of up to 17% compared to the manually constructed Bayesian network model.
Conclusion: Finding the right balance between expert knowledge and data-derived knowledge, both at the level of network structure and parameters, is key to using Bayesian networks for medical image interpretation. A balanced approach to building Bayesian networks for image interpretation yields more accurate and understandable Bayesian network models.

1. Introduction

Bayesian networks have become the state-of-the-art for the representation of and reasoning with uncertain knowledge of a clinical problem. They have a sound statistical basis, yet allow exploiting available background knowledge in a way superior to many other formalisms for statistical machine learning. Even when no data at all are available, it is often still possible to develop a Bayesian network, guided by a mixture of expert knowledge and information from literature. If data are available, one can also learn the network structure and parameters from data. As this holds for any medical domain, it also holds for medical imaging. However, medical imaging has its own characteristics: methods are applied, ranging from image segmentation via region detection to lesion determination, as part of the *image processing pipeline*. At the very end of this pipeline we find image interpretation; methods for image interpretation are, thus, clearly dependent on the previous processing steps. Some of the characteristics of the image processing steps, such as that image features are continuous, have particular implications for image interpretation that has a foundation in medical knowledge of the structure—histology and anatomy—and function—physiology. As in the end, medical images need to tell something about the patient, medical knowledge offers a natural start for computer-aided detection. However, exploiting explicit representations of

* Corresponding author. Tel.: +31 24 3652104; fax: +31 24 3652525.
E-mail addresses: marinav@cs.ru.nl (M. Velikova), peterl@cs.ru.nl (P.J.F. Lucas), msamulski@gmail.com (M. Samulski), n.karssemeijer@rad.umcn.nl (N. Karssemeijer).

0933-3657/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.12.004

medical knowledge into medical image interpretation has so far met with significant challenges.

These challenges bring us back to the relationship between manual construction and learning from data of Bayesian network, a topic discussed repeatedly in the past, without giving rise to scientific consensus. New in this paper is that we address this issue from the point of view of image interpretation. We critically exam the assumptions made in the expert-knowledge-guided development of a Bayesian network for medical image interpretation by the use of image data. Both the assumptions made in choosing the probabilistic parameters and in designing the graphical structure are studied.

The research was carried out in a concrete clinical setting: the interpretation of breast-cancer screening mammograms. Breast-cancer detection is a hard medical image interpretation task. With the digitisation of medical images in the last decade, there has been considerable progress in computer-aided interpretation of mammograms where most of the improvement have come from the development of new pattern-recognition techniques to better detect potentially suspicious breast regions. However, existing systems still exhibit limitations in attaining the required clinical accuracy, i.e. with respect to presence or absence of cancer in the patient. The major reason for this is their failure to explicitly represent the working principles and knowledge of human experts; expert radiologists normally compare image parts and different images of the breasts to each other, i.e. they interpret potentially suspicious regions of the breasts in the context of all other available image information. It is only recently that researchers have started to study ways to incorporate such principles into computer-aided detection (CAD) systems [1,2].

As part of the research we constructed a Bayesian network that incorporated the most important image features and their relationships as used by radiologists to interpret mammograms. Thus, the resulting Bayesian network can be looked upon as a knowledge representation of mammogram interpretation in terms of breast tissue architecture and signs of abnormality. As image features are continuous variables, we used Gaussian distribution to model their uncertainty.

In a well-cited paper by Pradhan et al., published in 1996, it was experimentally established that the network structure is the single most important factor determining the Bayesian network's performance [3]. In time, this insight has become general wisdom underlying much of Bayesian network modelling. The results of this paper were in particular compelling as they were based on an extensive study of a variety of large, real-world networks. In our paper, we challenge the conclusions from the paper by Pradhan et al. and aim at offering a more balanced view on this important problem. It is also the right time to reexamine this problem, as considerable progress has been made in Bayesian network technology since 1996. Other recent research [4] also suggests that the problem of the sensitivity of Bayesian networks to imprecision in their parameters is domain-dependent and requires more careful investigation.

We emphasise that the problem of medical image interpretation we tackle in this paper is particularly challenging as the input to the network is based on image features *automatically* extracted by a CAD system through image processing, which in itself is a complex task and ongoing research. Even though the continuous nature of the features obtained in this way is understandable from a physical point of view, their relationships to the clinical abnormalities detected in the image are not straightforward from the radiologist's point of view. In contrast, the features provided by radiologists after visual inspection and interpretation are discrete; they have a specific semantics, although prone to subjective variation [5,6]. Furthermore, the manual network contained two features obtained from the CAD system's output, that are assumed to have direct causal relationships with the variable that indicates whether or not an abnormality is present. Again, the inclusion of such variables is novel in comparison to available benchmark datasets for breast cancer and their relationships have not been studied before.

Hence, the novelty of our research lies in the thorough investigation of both the quantitative part (probability distribution) and the qualitative part (structure) of the manual network to obtain insight into the appropriateness of the assumptions made in developing a Bayesian network for a highly complex task: medical image interpretation. The selected task of mammogram interpretation is sufficiently similar to other complex medical image interpretation tasks to act as an example problem for the research. As breast cancer is a major disorder that is associated with enormous research efforts, techniques for the automated detection of breast cancer reflect the state of the arts of the field of CAD.

In this study in particular, we build upon our results from the work presented in [7], where we discretised the continuous mammographic features automatically extracted from the CAD system to check whether the probabilistic parameters in the initial expert network were optimal and correctly reflecting reality. It was shown that the parameters play an essential role in the network's performance. Therefore, after preliminary investigations [8], in the current study we provide an extensive and thorough investigation of learning Bayesian network structures, both restricted and unrestricted, from the discretised image data to gain detailed insight into the feature dependencies and independencies assumed in the manual model. The performance of the learned networks is compared with that of the manual network in terms of classification accuracy and knowledge representation.

The structure of the paper is as follows. We start with a review of the theory of Bayesian networks and related work in the areas of discretisation and structure learning in Section 2. Next, in Section 3, some background is provided on mammogram interpretation, the Bayesian network for mammogram interpretation that was developed by hand is presented, and we describe the data used for the experimental part in the research. Our previous work that examines the assumptions about the probabilistic parameters of the Bayesian network is shortly summarised in Section 4. This is done to provide the reader with a good understanding of the choices made about the discretisation of the data used for the research study on structure learning presented in Section 5. Finally, in Section 6 we return to the questions from which the research started and summarise what has been achieved.

2. Background

2.1. Bayesian networks

A *Bayesian network* (BN) is defined as a pair $\mathcal{B} = (G, P)$, where G is a directed acyclic graph (DAG) $G = (V, E)$ and P is a joint probability distribution of the random variables X_V [9–11]. There exists a 1-1 correspondence between the nodes $v \in V$ and the random variables $X_v \in X_V$; the (directed) edges, or arcs, $E \subseteq (V \times V)$ correspond to direct causal relationships between the variables: a node is a *parent* of a *child*, if there is an arc from the former to the latter. We say that G is an *I-map* of P if any independence represented in G, denoted by $A \perp\!\!\!\perp_G B \mid C$, with $A, B, C \subseteq V$ mutually disjoint sets of nodes, is satisfied by P, i.e.

$$A \perp\!\!\!\perp_G B \mid C \quad \Longrightarrow \quad X_A \perp\!\!\!\perp_P X_B \mid X_C \ ,$$

where A, B and C are sets of nodes of the DAG G and X_A, X_B and X_C are the corresponding sets of random variables, indexed by A, B and C. The acyclic directed graphical part of a Bayesian network G is by definition an I-map of the associated joint probability distribution P. A Bayesian network $\mathcal{B}$ offers a compact representation of the

joint probability distribution P in terms of local *conditional probability distributions* (*CPDs*) or *conditional probability tables* (*CPTs*), if the data are discrete, by taking into account the conditional independence information represented by the DAG. In these terms, a useful concept is the so-called *Markov blanket* of a node $v \in V$, consisting of its parents, children, and the children's parents. It can be proven that a node v is conditionally independent of all other nodes given its Markov blanket, which implies that this is the only knowledge needed to predict the behaviour of that node [9].

The structure and parameters of a BN can be determined manually using expert knowledge or learnt automatically from a dataset. Hybrid approaches are also common practice, as done in the current study, aiming to combine the knowledge acquired from human experience, on the one hand, and from the factual quantitative information, on the other hand. We next briefly discuss the fields of discretisation and structure learning methods, and review related studies of such methods for biomedical problems.

2.2. Discretisation

Discretisation of data has been studied for more than 20 years as one of the major preprocessing steps in data analysis. Its goal comprises the transformation of continuous variables into a finite number of discrete values, or ranges, to facilitate: (i) the improvement in classification performance, (ii) the induction process of a classifier, or (iii) the interpretability of the models learnt. In the context of Bayesian networks, automatic discretisation is used as method to reexamine the probabilistic parameters.

Two simple and often applied methods for discretisation are *equal frequency binning* and *equal width binning*, which determine the bin boundaries by first sorting the data on ascending values and subsequently divides the sorted data in n equally *sized* or *ranged* bins, respectively. Both methods are unsupervised as they do not use class information.

A well-known supervised discretisation technique is the method of Fayyad and Irani [12], which uses the class entropy to facilitate the induction of better decision trees. This method selects a bin boundary based on the minimisation of the class information entropy. The class entropy of a (sub)set S is defined as

$$\mathrm{Ent}(S) = -\sum_{i=1}^{k} P(C_i, S) \log P(C_i, S),$$

where $P(C_i, S)$ represents the proportion of instances in S with class C_i and k stands for the number of classes. For each candidate cut point T of an attribute A, a weighted average is calculated of the entropy of the two subsets S_1 and S_2 created by the cut point:

$$E(A, T; S) = \frac{|S_1|}{|S|}\mathrm{Ent}(S_1) + \frac{|S_2|}{|S|}\mathrm{Ent}(S_2),$$

where $|\cdot|$ represents the cardinality of a set. The candidate cut point for which this function is minimal is selected. This process can be repeated on the subclasses to create multiple bins, but the minimal description length criterion is used as a stopping criterion to avoid ending up with too many bins.

In [13] the authors investigated the reduction of the variance introduced by various discretisation techniques for decision tree induction. The results demonstrated that this reduction facilitates the interpretability and stability of the models learnt. In [14], a categorisation of 8 types of discretisation methods is provided such as supervised vs. unsupervised, parametric vs. non-parametric, global vs. local. In the same study, the authors propose a novel discretisation method based on the so-called wrapper approach where the accuracy of a naïve Bayes classifier is taken into account in the discretisation process in order to guide the search for the best ranges of all variables to improve the classification accuracy.

Comparative studies of various discretisation techniques on the performance of naïve Bayes classifiers are provided in [15–17], showing improvement in the results compared to the continuous baseline. In addition, in [18] the effectiveness of a number of discretisation methods is evaluated to provide a heuristic for the selection of the best discretisation method.

2.3. Structure learning

While the probabilistic parameters are one side of the coin, the independence relationships expressed by the graphical structure are clearly the other side. We will not review all work on structure learning but restrict ourselves to the basics; for a general overview, the reader is referred to the standard textbooks [9–11]. Structure learning is an optimisation problem that aims at finding the best graph representing the conditional independence relationships in the data. Use of exhaustive search for the optimal graph is, however, infeasible for most problems due to the explosive number of graphs for datasets with 5 or more variables [19]. More common is the use of greedy search, which searches the space of DAGs, or the space of *Markov equivalent* classes of structures, i.e. structures that encode the same conditional independence assumptions, called *essential graphs* (EG) [20]. The latter implies greedy search in EG space.

There are three main types of methods used for solving this problem: (i) *constrained-based*, where conditional independence tests are applied to determine relationship constraints, (ii) *score-and-search*, where a score is used to judge the fitness of the model, and a search method allows exploring the search space of acyclic directed graphs, and (iii) *hybrid*, which are a combination of the previous two. Markov blanket discovery of a variable is another typical problem, which aims to identify the minimal set of features that is needed for feature selection pre-processing or classification tasks; a number of efficient algorithms have been recently proposed in the literature [21–23]. The score measure always includes some measure of the likelihood of the data given the graph and its probabilistic parameters, $\Pr(D|G, P)$, where D are the data and the probabilistic parameters P are fixed. If only the fitness of the graph with respect to the data is investigated, the marginalised likelihood, $\Pr(D|G)$, is used. Here the parameters P are marginalised out. In addition, the score measures typically include the possibility to include a prior on the structure, $\Pr(G)$, and a penalty for unwanted complexity of graph structure. The score measure used in this research is the *Bayesian Dirichlet equivalent* (BDe) score, which is applicable only to discrete data [24]. This measure takes Markov equivalence into account.

2.4. Related work on discretisation and structure learning for biomedical problems

Quite a number of previous studies have been carried out that focussed on structure learning of BNs, discretisation of data, or both for biomedical problems. In [25] the authors proposed a search-and-score approach using the mutual information among all pairs of variables as a guide to restrict the creation of nonsignificant connections in the network structures. Constraints and causal grouping of variables are explored in a non-parametric, hierarchical approach proposed in [26] to facilitate the structure learning of BNs for small, sparse datasets. Based on medical knowledge, for example, the variables can be grouped into two classes—diseases and symptoms—and the influence from diseases to symptoms can be imposed as a constraint for the learner to reduce considerably the hypothesis space of structures. A recent study [27] presents a methodology for combining expert knowledge and structure learning to model heart failure using BNs. This combination is achieved by pre-classifying the available data as background, past and current health. To a large extent the data used in this study is

based on discrete features with clearly defined values, which makes the network modelling step easier.

In [28], the authors compare different algorithms for structure learning of BNs in order to build a model for facilitating an emergency hospital service. The study was based on a real dataset containing manually collected patient data, where some of the variables were manually discretised based on meaningful context. Another method using discretisation as a data-preprocessing step to structure learning of a BN is proposed in [29] where the authors build a model to predict local failure in lung cancer. The discretisation boundaries are determined using a three-bin strategy based on mutual information. Three-valued unsupervised discretisation using minimum, maximum and mean value of image functional MRI features is used in [30] to learn the connectivity between brain regions as dynamic Bayesian networks. In [31] another recent approach is described where binary threshold-based discretisation of continuous electrocardiogram features is used prior to learning Bayesian network classifiers for distinguishing between various patient age risk groups.

Some commonalities can be noticed in these previous works. First, they all apply discretisation as a data pre-processing step based on one simple unsupervised strategy but they do not investigate the effect of various discretisation schemes on the classification performance and data fitting capabilities of the networks. Second, the approaches mostly aim to learn and compare various Bayesian network structures in terms of prediction accuracy of a particular variable of interest and not in terms of dependence relationships between all features in general. In contrast, in this study we first compare automated supervised and unsupervised discretisation methods to reexamine the probabilistic parameters of the manual network for the problem at hand. Based on the data obtained from the best-performing discretisation method, we next examine the (in)direct dependencies between the image features, assumed by expert knowledge, by learning various network structures.

3. Mammography, the Bayesian network model and data

We start by reviewing the problem domain of screening mammography and describe in detail the Bayesian network that was constructed for the purpose of mammogram interpretation.

3.1. Mammographic analysis

The early detection of breast cancer is crucial for the effective management of the disease as it increases the survival chance and improves the quality of life for the patient. Currently, the most cost-effective early detection method is screening mammography, which is based on regular X-ray examinations of asymptomatic women. Every patient's examination is analysed independently by two radiologists, yielding cancer detection rates up to 15% higher than for examinations done by one radiologists only [32].

The radiologists judge the presence of cancer on the basis of two projections, or *views*, of the same breast: mediolateral oblique (MLO), taken under 45° angle and showing part of the pectoral muscles, and craniocaudal (CC), taken head to toe; see Fig. 1. If cancer is present, then it is expected to be observed in both views. Furthermore, abnormalities can appear on the mammograms as microcalcifications (tiny deposits of calcium) and masses, defined as space occupying lesions seen in two different projections [33]. Masses are the typical presentations of breast cancer and they are difficult to detect due to their similarity with normal breast tissue.

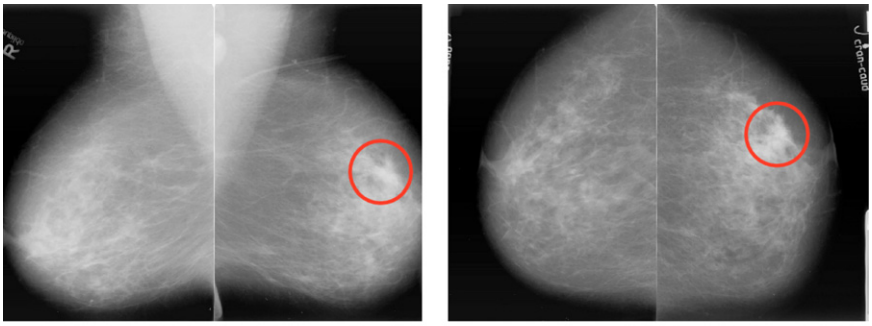
In the past few years, CAD systems have been introduced into screening to assist radiologists in their interpretation task [34]. The majority of current CAD systems are mainly meant for analysing only a single image with the aim to localise and classify an abnormality as being cancerous or not, using imaging techniques and classifiers such as neural networks. The CAD system used in this study [35], executes the following four steps: (1) segmentation of the mammogram into several components, such as breast tissue, background, and the pectoral muscles; (2) initial detection of suspicious pixel-based locations; (3) extraction of regions and region-based features, and (4) classification of the extracted regions as cancerous or normal using a neural network classifier. The system is more intended as a prompt system to focus attention of the radiologists, rather than as a clinical decision-support system. In screening, breast cancer appears mostly as one or two cancerous regions in the two views. The CAD system, however, often yields false positive regions. The reason is that the CAD system uses only local information to determine whether a region is suspicious and ignores the basic working principles of the radiologists, such as examining complementary information from the other view, or mammograms made at previous screening examinations. Integrating such domain knowledge in the modelling scheme of the CAD system can considerably improve its performance in terms of detection rate and interpretation of the results and Bayesian networks (see below) are very suitable for this purpose.

We adopt the following terminology from the breast cancer domain throughout this paper. We call a contoured area on a mammogram a *region*, marked, for example, manually by a human or detected automatically by a CAD system. By *lesion* or *finding* we refer to a cancerous region detected in the patient; see Fig. 1. A region detected by a CAD system is described by a set of continuous (real-valued) *single-view features*, e.g. size, location, contrast. By *link* we denote established correspondence, between two regions in MLO and CC views, respectively. The term *case* refers to a woman who has undergone a mammographic exam. Below we shall frequently refer to experimental results as corresponding to the *link level*, i.e. with respect to two corresponding regions in the CC and MLO views, and to the *patient level*, i.e. the patient case with or without breast cancer.

Previous research has already demonstrated the potential of exploring multi-view dependencies to improve the automatic detection of breast cancer on mammograms. The approaches in [36–39] focused on improving the lesion-based results, based on the distinction between true and false positive links of regions in MLO and CC views. In most of the previous work mentioned, neural networks or linear discriminant analysis were explored for the detection of breast-cancer lesions.

In recent studies of ours we have opted for Bayesian network methodology as it has the advantage of providing not only strong predictive power but also explicit modelling of expert knowledge and insight in the results obtained—properties desired especially by medical specialists. In particular, we built Bayesian network models using multi-view information to increase both lesion-based and case-based performance, i.e. fraction of true positive exams where an exam is true positive if cancer is found in the MLO *or* CC views, in comparison to a single-view CAD system [1,2]. The building process of these models was guided by domain knowledge of how radiologists interpret mammograms with the main goal of improving the classification performance of the CAD system. The model structures were manually constructed based on the continuous mammographic features extracted from the images, and the parameters were learnt from data.

While our previous research showed good results in terms of better cancer detection, it did not explicitly and thoroughly study the various modelling principles of automated image interpretation in terms of graph structure (manually built or learnt from data) and type of features (discrete or continuous) as done in this study. Therefore the current research contributes to getting more insight



(a) MLO (b) CC

Fig. 1. MLO and CC views of both breasts of a patient. A cancerous lesion is marked by the circle.

in the choices to be made when modelling the complex problem of medical image interpretation.

3.2. A Bayesian network model for mammographic analysis

The baseline BN model used in this study was developed using expert knowledge and it was first proposed in [40]; it is reproduced in Fig. 2. The BN incorporates MLO and CC features, represented by the white rectangles on the figure, which can be interpreted at the same time, allowing the integration of information from two views. These features are continuous (real-valued) and computed by the single-view CAD system independently per view. Below we describe the most important features, used in the BN model, which determine whether or not two regions linked between both views represent a finding:

A. *Observed features* extracted from the image in step (3) mentioned above:
   - `LocX` and `LocY`: The relative location of the region in terms of *x*- and *y*-coordinates on the mammogram. Some areas of the breast are more likely to contain cancer, e.g. the upper outer part, than others.
   - `D2Skin`: The shortest distance of the region to the skin.
   - `Contrast`: High contrast on the mammogram is often associated with a malignancy: tumour tissue absorbs more X-rays than fat and glandular tissue.
   - `Spic`: Indication whether the region margin has a spiky pattern towards the centre of a lesion, called 'spiculation'; the higher the degree of spiculation, the higher the likelihood for malignancy.
   - `FM`: The presence of a circumscribed (well-defined) lesion, the so-called 'focal mass'.
   - `LinText`: Linear texture, which represents normal breast tissue—the higher the linearity, the lower the likelihood of being malignant.
   - `Size`: Size of the region–very large regions are usually benign abnormalities.

B. *Calculated features*, computed from classifiers based on pixel- or region-based features:
   - `DLik`: The malignancy pixel-based likelihood computed by a neural-network classifier using pixel-based features
   - `FPlevel`: The false-positive (FP) level of a region computed by a neural-network classifier using region-based features; it indicates the average number of normal regions in an image with the same or higher likelihood scores, so the lower its value, the higher the likelihood that the region is cancerous.

The simultaneous interpretation of the MLO and CC features is modelled by the corresponding *hidden* variables (in light grey ovals in the figure), which are not directly observed or measured in the CAD system, but represent the way as radiologists would evaluate the mammographic characteristics of a finding. The variable `Finding` represents the conclusion whether or not there is cancer in the breast, i.e. whether or not two linked regions in MLO and CC views represent a lesion. Central to the BN model are also the hidden variables `Abnormal Density` and `Abnormal Structure`, indicating the presence of abnormal density and structure and they have two states: 'present' and 'absent'. The model was developed for the purpose of two-view mammographic interpretation where the main variable of interest is `Finding`.

In the following sections, we explore various assumptions concerning both graphical structure and parameters of the BN in depth in order to get more insight in the network modelling by means of experiments with mammographic data. First, in the original model with continuous variables, the assumption was that all view features can be described by Gaussian distributions. We discretised the MLO and CC features to see what changes in the performance and knowledge representation of the model are achieved. Although previous research suggests that performance increase can be expected, there is no literature telling how the improved performance can be interpreted in the context of available domain knowledge, and this is what we intend to investigate. Second, the difficulty in the mammographic analysis task implies also complex interactions between the mammographic features, which might not be easily determined a priori. Hence, by learning graphical structures from data we aimed to discover typical relationships between the mammographic features per view and per breast, again in the context of available knowledge. We next describe the characteristics of the data used in this study.

3.3. Data description

Data were obtained from the Dutch breast cancer screening practice and includes the mammographic examinations of 795 patients, of which 344 were cancerous. All cancerous breasts had one visible lesion in at least one view, which was verified by pathology reports to be cancerous. Lesion contours were marked by a mammogram reader.

For each image (mammogram) we have a number of regions detected by the single-view CAD system. We selected the three most suspicious regions per image (view). Every region is described by continuous features (see Section 3.2). Based on the ground-truth data, for each region we assign a class value of 'cancerous' if the detected region hits a cancerous abnormality and 'normal' otherwise. Since a region in one view cannot always be coupled to the corresponding area in the other view due to the compression and the rotation of the view, for every breast we linked every region from MLO view with every region in the corresponding CC view.

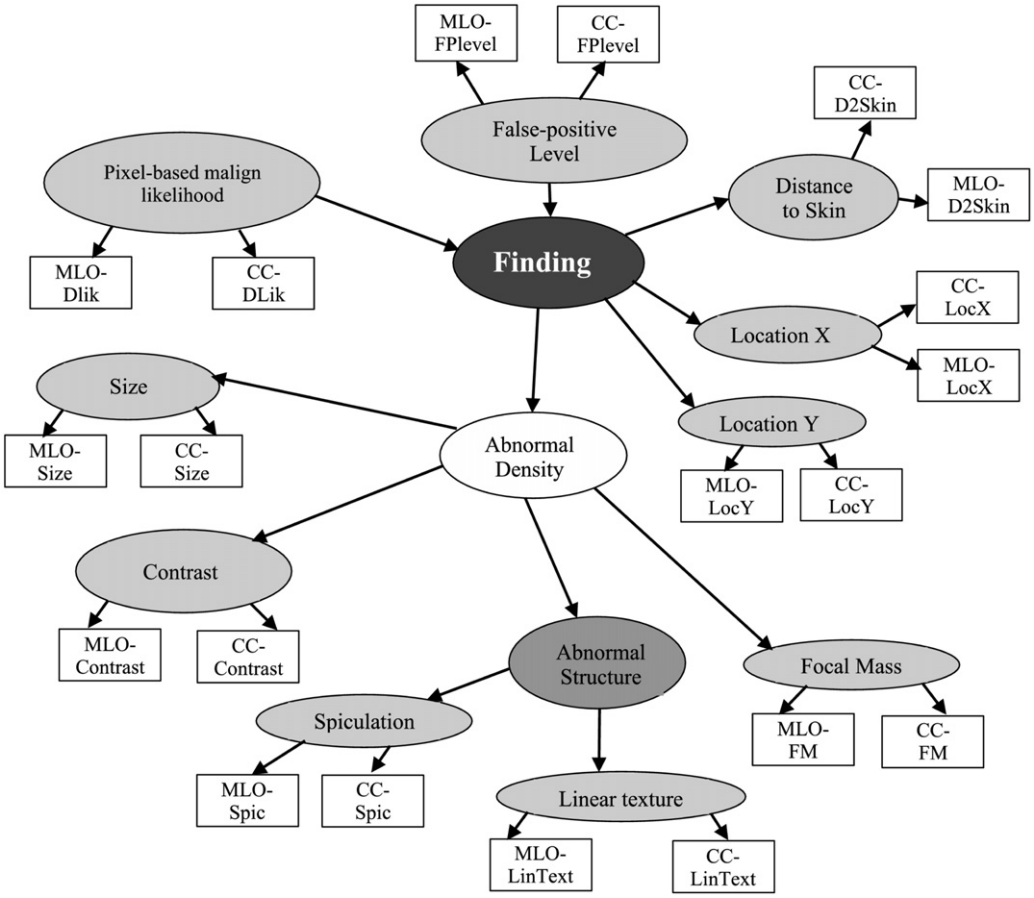


Fig. 2. Bayesian network model for two-view mammographic analysis.

For every link we added the class values of 'cancerous' ('true') if at least one of the linked regions is cancerous; otherwise the class is 'non-cancerous' ('false'). This forms the data for the variable `Finding` in the BN model. We assign analogous classes for the patient based on the ground-truth information. This results in a database where for each breast multiple instances are added, and each instance reflects a link between a CC and a MLO region. The final dataset consists of 14,129 links. For the experiments with structure learning algorithms, we also created two datasets containing only MLO regions (4707 in total) and CC regions (4710 in total), where the class variable is based on the ground-truth of the respective region.

Although this dataset is based on the mammographic examinations from a specific national breast cancer screening program, we note that it represents typical mammographic data obtained from automated image analysis including standard image processing steps such as breast segmentation, region segmentation and feature extraction. In this sense, we believe that the scope of the results and conclusions from this study are to be considered more broadly and applicable generally to the problem of modelling complex medical tasks.

4. Reappraisal of the probabilistic parameters

As most of the variables modelled by the manual Bayesian network were continuous features, they were represented using conditional Gaussian distributions. A limitation of Gaussian distributions is that they are symmetric, which will not allow capturing asymmetries available in the data. Rather than using other continuous probability distributions, that would allow representing asymmetries, however again with particular assumptions, discretisation of the continuous data offers a way to fit the probability distribution to the data with no assumptions about the shape of the distributions. Discretisation is studied in two different ways, namely with regard to (i) classification performance and (ii) goodness of fit of the resulting probability distribution to the data.

4.1. Experimental set-up

The following discretisation methods, discussed in Section 2.2, were investigated and compared, as implemented in the software package WEKA [41]:

- Equal frequency binning with ten bins (EFB-10)
- Equal width binning with ten bins (EWB-10)
- The method of Fayyad and Irani (FI)

To build and evaluate the data-driven BN models with discrete data, we used a two-fold cross validation procedure: the dataset is randomly split into two subsets with approximately equal number of observations and proportion of cancerous cases. The data for a whole case belonged to only one of the folds. Each fold was used as a training set and as a test set. We built, trained and tested the networks by using the Bayesian Network Toolbox in Matlab [42]. The learning of the probability parameters was done using the expectation-maximisation algorithm, which is typically used to approximate a probability function given incomplete samples, as the network contains hidden variables [43]. The performance of the BN models learnt with discretised data were compared with the benchmark model, described in Section 3.2, learnt from the

Table 1
AUC and log-likelihood test results obtained from the continuous baseline and the discretisation methods.

Method	AUC		*LogLik*	
	Link	Patient	Link	Patient
Continuous baseline	0.707	0.628	0.466	0.760
FI	0.790	0.755	0.382	0.633
EFB-10	0.754	0.733	0.394	0.617
EWB-10	0.720	0.654	0.407	0.661

continuous data, for short called the *continuous baseline*. The comparison analysis is done using the receiver operating characteristic (ROC) curve and the area under the curve (AUC), a standard performance measure in the medical image research [44]. We also evaluated the data fitting capabilities of the models learnt by means of the log-likelihood, *LogLik* for short, of the class predictions C given the dataset D, $L(C|D)$:

$$L(C|D) = \frac{1}{N}\sum_{i=1}^{N} -\log P(C_i|\mathcal{E}_i), \quad (1)$$

where $N = |D|$ is the number of observations, C_i and $\mathcal{E}_i$ are the class value and the feature vector of the ith observation in D, respectively. Thus, the value of $L(C|D)$ indicates how close the posterior probability distribution is to reality: when $P(C_i|\mathcal{E}_i) = 1$ then $\log P(C_i|\mathcal{E}_i) = 0$ (no extra information); otherwise $-\log P(C_i|\mathcal{E}_i) > 0$.

4.2. Results

The discretisation obtained from the supervised FI method led to various number of bins for each variable. They range from one bin for the variables `MLO-FM`, `MLO-LinText` and `CC-Contrast` to five bins for `MLO-LocY`—clearly less than the ten bins obtained from the alternative discretisation methods. Table 1 presents the AUC and log-likelihood test results at a link and patient level for the three discretisation methods and the continuous baseline. In terms of accuracy at a link level, the FI method performs best, followed by the EFB-10 and EWB-10 methods. Although the results at a link level are an indicator for the model performance, from a clinical point it is interesting to consider the results at the patient level. The FI method again achieves the best discrimination between cancerous and normal cases, followed by EFB and EWB with ten bins.

To obtain better insight into the improvement of the classification performance, we plotted the ROC curves for the best performing methods at both link and patient level, as shown in Fig. 3.

It is interesting to observe that for the supervised FI method the bigger improvement in the model's performance is in the lower FP range (<0.5)—a desired result in the screening practice where the number of normal cases is considerably larger than those of the cancerous ones. Furthermore, we note that the curves (and respective AUCs) for all methods are lower at the patient level than at the link level, as for the former the number of false positives is much smaller leading to a bigger penalty for a misclassified cancerous case.

We further evaluated the data fitting capabilities of the models with the discrete and continuous data using the log-likelihood measures reported in Table 1. Clearly the FI method fits best to the data at a link level as it achieves the lowest *LogLik* value, and it is followed by EFB-10. At a patient level, the performance of both methods has the opposite pattern. The BN model with continuous features fits considerably worse to the data in comparison to the models with discretised data, indicating a mismatch between the model and the data. These results confirm our expectation that discretisation can facilitate the knowledge representation and modelling of the problem of automated mammographic analysis.

Finally, Fig. 4 illustrates the behaviour of the BN model with discrete MLO and CC features obtained from the FI method for one cancerous link (true finding) from the data. The evidence has been set on the observable nodes, and thereafter the posterior probability of the finding being cancerous has been updated. The model clearly succeeds in the correct classification of the link. Furthermore this model is easier to work with and interpret in comparison to the model with continuous features as the former yields a representation that is closer to the knowledge of the radiologist.

5. Network structure reappraisal by learning

In this section we investigate the structures learnt from various algorithms and compare them with the continuous baseline

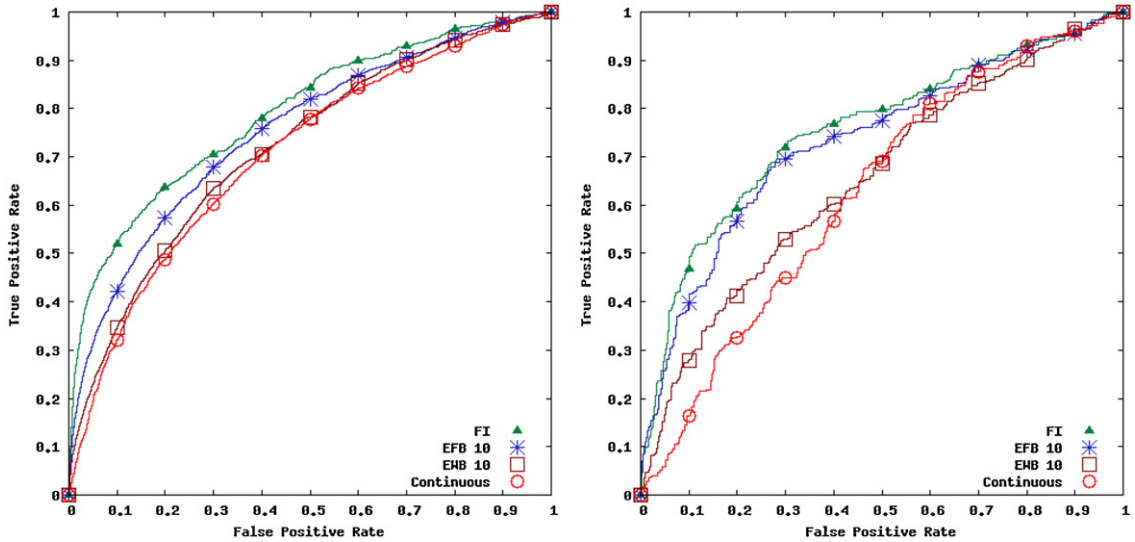


Fig. 3. ROC curves for the best performing discretisation methods against the continuous baseline at (a) *link* level and (b) *patient* level.

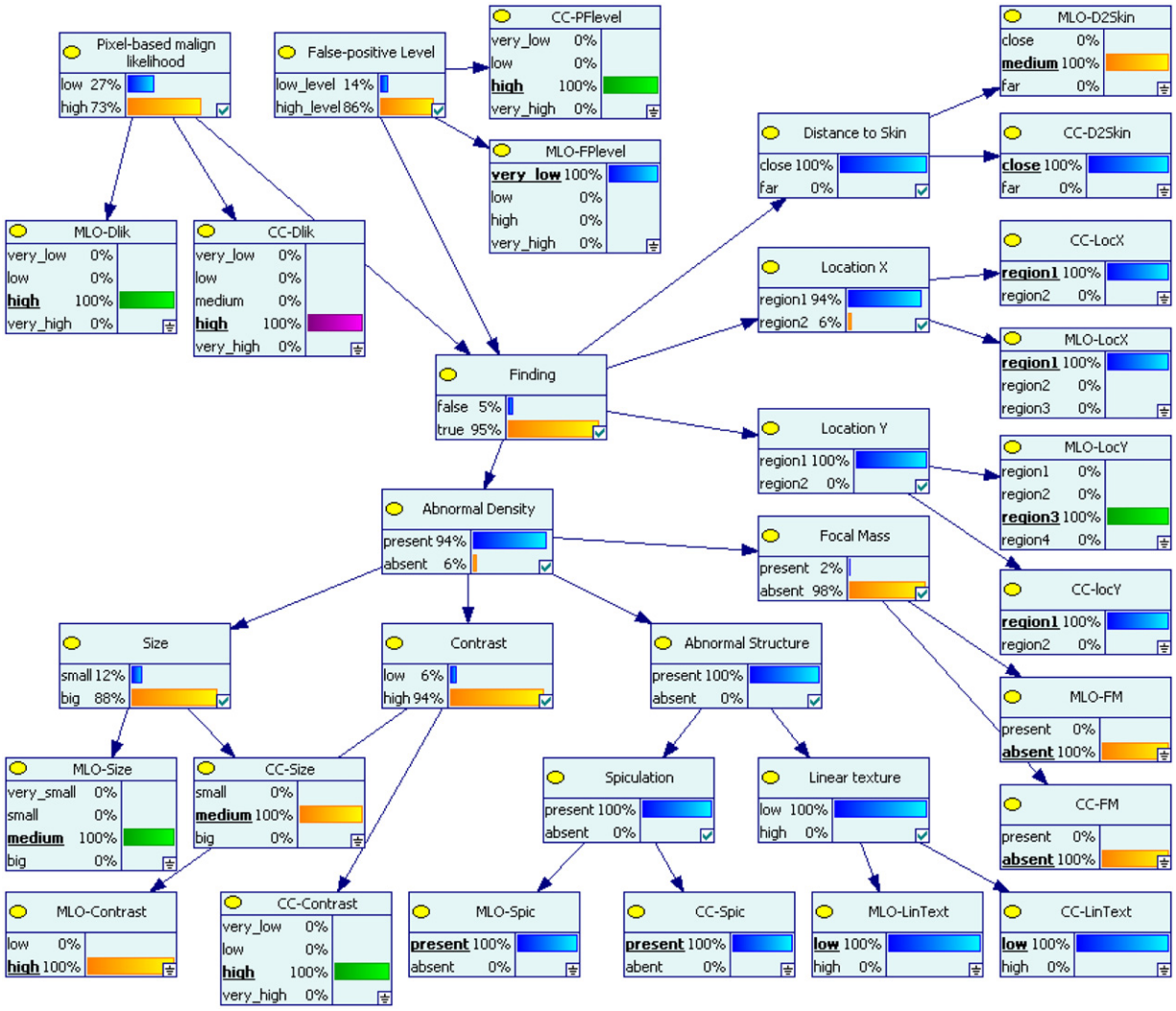


Fig. 4. Bayesian network with evidence set (represented by bold and underlined names of the states) for one cancerous link and posterior probabilities with discretised data using the FI method.

network. For this purpose, we used the discretised data obtained from the best performing FI method as reported in the previous section.

5.1. Experimental set-up

In the following experiments, we use two main structure learning algorithms: the *Max-Min Hill Climbing* (MMHC) algorithm [45]—a recent hybrid algorithm that has been extensively tested over a wide variety of data sets and showed good performance, and the constrained-based *Grow-Shrink* (GS) algorithm [46]. In addition to structure learning, we also explored simpler reference BN structures for comparison:

Naïve Bayes: This model consists of only one parent node—the fixed class node `Finding`—and assumes that all the feature variables are conditionally independent given the class.

TAN: Tree-Augmented Bayesian Network (TAN) algorithm—a simple classifier and an extended version of Naïve Bayes, allowing for more dependencies between the features.

Independent: All variables are considered *independent*, i.e. there are no arcs in the model.

Fully connected: All variables are considered *dependent*, i.e. the network is fully connected.

The aim of the experiments was to investigate the dependence relationships between the observed and calculated features, and the class variable `Finding`. In particular,

(i) We investigated the assumptions for dependencies and independencies between `Finding`, the calculated and observed features modelled in the manual network based on a subset of the combined-view data containing 4 MLO and 4 CC features (`FPLevel`, `DLik`, `LocX` and `D2Skin`), and the `Finding` variable. We compared the manual network with the structures learnt from MMHC, GS and the reference BN models.

(ii) We studied the presence of persistent direct dependencies between the mammographic features across structures obtained from MMHC and GS applied to datasets *with* and *without* the calculated features.

(iii) Using the same two-fold cross-validation setup as described in Section 4.1, we compared the classification performance in terms of AUC and fitting capabilities of MMHC and TAN, applied to the discrete data, and MMHC applied to the original continuous data (MMHC-Gaussian). The comparison was done for the MLO, CC and combined view (breast-image) data. We also

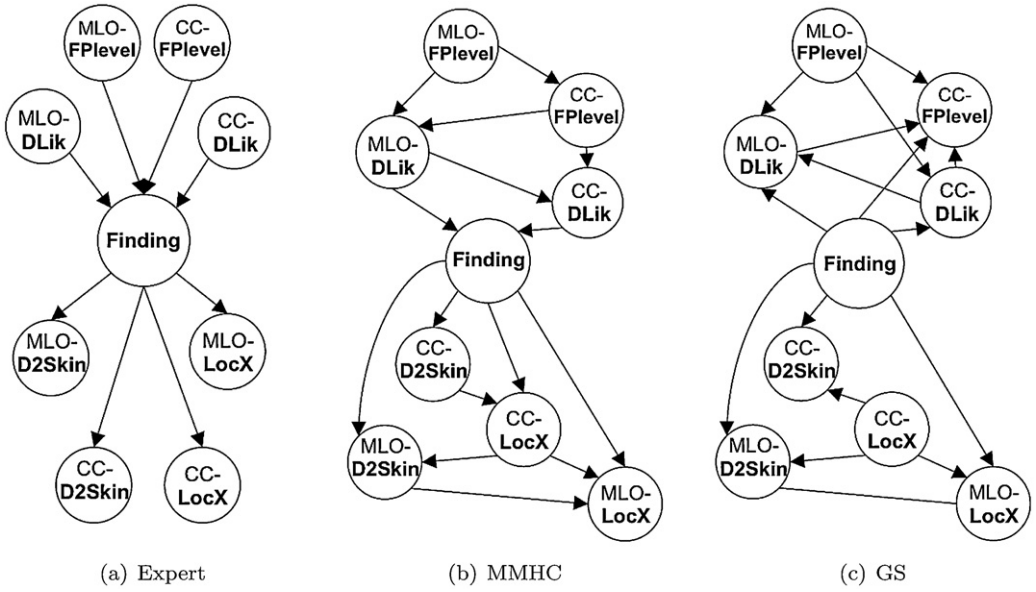


Fig. 5. Structures based on the expert sub-model, and learnt by the MMHC and GS methods.

compared the structures learnt from each of the two folds in order to see whether or not they differ considerably in terms of feature dependencies.

For the structure learning experiments we used the algorithms implemented in the freely available *bnlearn* package in R [47].

5.2. Results

5.2.1. Learning structures based on an expert sub-model

The resulting expert sub-model and the structures learnt using MMHC and GS based on the data subset are shown in Fig. 5.

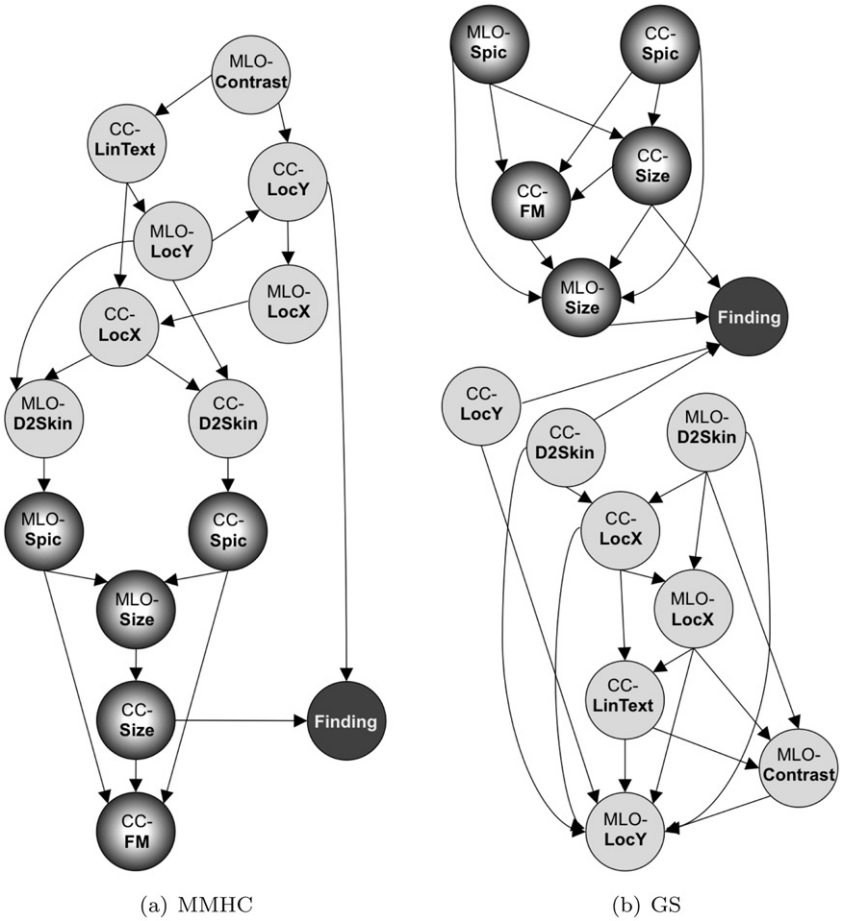
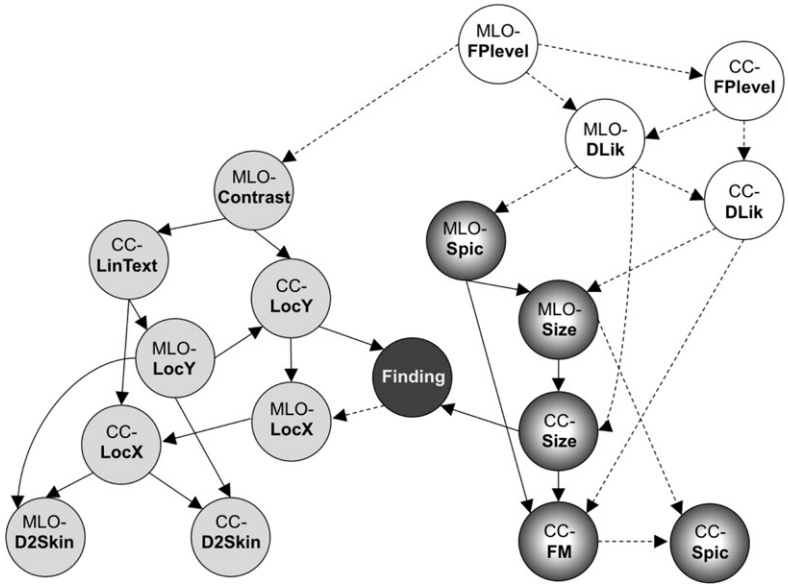


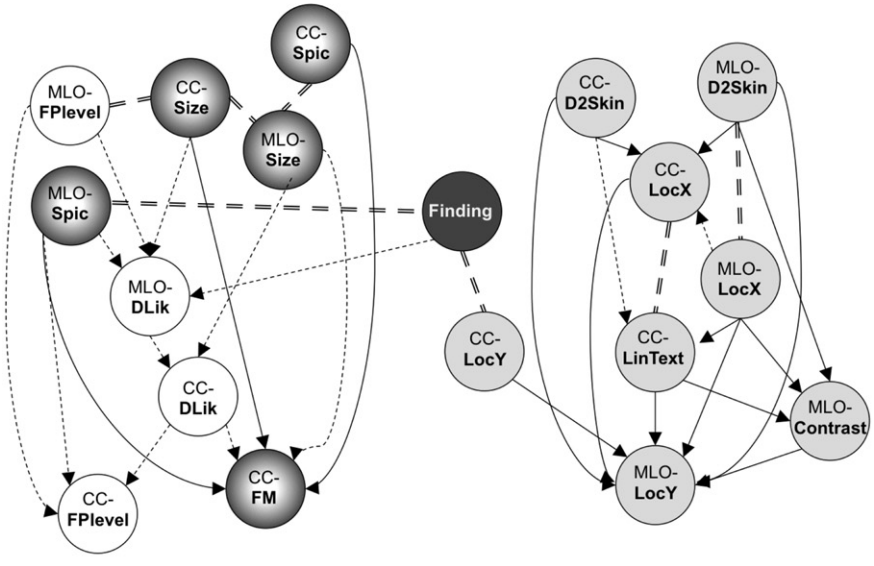
Fig. 6. Structures learnt from the MMHC and GS methods applied to the discrete combined-view data *without* the calculated features.

The structures learnt by MMHC and GS are very similar in terms of dependence relationships: the graph learnt by MMHC includes a bit more independence information with respect to the variable `Finding` and the calculated features `MLO-FPlevel`, `CC-FPlevel`, `MLO-DLik` and `CC-DLik`, whereas there is an even smaller difference in the dependence information of the observed features below the variable `Finding` in the networks. For example, where there is a *un*conditional dependence `Finding` $\not\perp\!\!\!\perp_{G_b}$ `CC-LocX` $|\varnothing$ in the graph of Fig. 5(b), this dependence is conditional, e.g. `Finding` $\not\perp\!\!\!\perp_{G_c}$ `CC-LocX`| `MLO-LocX`, in the graph of Fig. 5(c). Furthermore, the model obtained by the constrained-based algorithm contains one undirected arc between `D2Skin` and `LocX` in the MLO view, implying that the direction does not matter. Another major observation for the learnt structures is that the observed features are conditionally independent of the calculated features given knowledge of `Finding`. This property also holds for the expert sub-model. However, the models learnt clearly contain a larger number of dependence relationships between the observed and the calculated features than the manually constructed model. While the direct dependencies between the respective MLO-CC features are partially represented via the hidden variables in the original network in Fig. 2, the dependencies between the distance to skin and size, as well as between both calculated features are clearly missing in the manual network.

The lacking dependence relationships are also confirmed by the fitness BDe scores for the manual and learnt structures reported in Table 2, where we also report the scores for the reference BN models, which do not consider any knowledge incorporated in the expert model:



(a) MMHC



(b) GS

Fig. 7. Structures learnt from the MMHC and GS methods applied to the discrete combined-view data *with* the calculated features. The dashed arcs present different dependencies in comparison to the graphs in Fig. 6, while the double dashed edges are undirected dependencies yielded by the GS algorithm.

Table 2
BDe scores obtained from the expert model, the structure learning algorithms, and the reference models applied to the combined MLO and CC subset data.

Model	BDe score ($\times 10^4$)
Expert	−7.6579
MMHC	−7.3173
GS	−7.3318
Naïve Bayes	−7.6543
Independent	−7.7068
Fully connected	−7.5731

The models with the highest scores are the ones obtained from MMHC and GS, followed by the fully connected model. Note that the expert model fits to the data comparably to the Naïve Bayes and the independent model. This implies that the expert sub-model fails to capture dependences between the observed and the calculated features.

5.2.2. Direct dependencies between the mammographic features

Fig. 6 presents the structures learnt from MMHC and GS algorithm applied to the discrete combined-view data *without* the calculated features.

We observe that the structure learning finds the direct dependencies between the respective observed features in MLO and CC views, showing that they are strongly present even after the feature discretisation. This confirms our modelling assumptions used in the manual network, where these dependencies are modelled via the hidden nodes for each feature. Furthermore, in the structures learnt by both methods `Finding` appears as a direct effect of `CC-Size` and `CC-LocY`—but the constrained-based method finds two more parent nodes of the class variable—`MLO-Size` and `CC-D2Skin`. This finding is opposite to the relationships modelled in the manual network, where `Finding` is the parent node to the observed features.

Another interesting result is that both structure learning methods uncover two clusters of features that exhibit strong direct within-cluster relationships and relatively little between-cluster

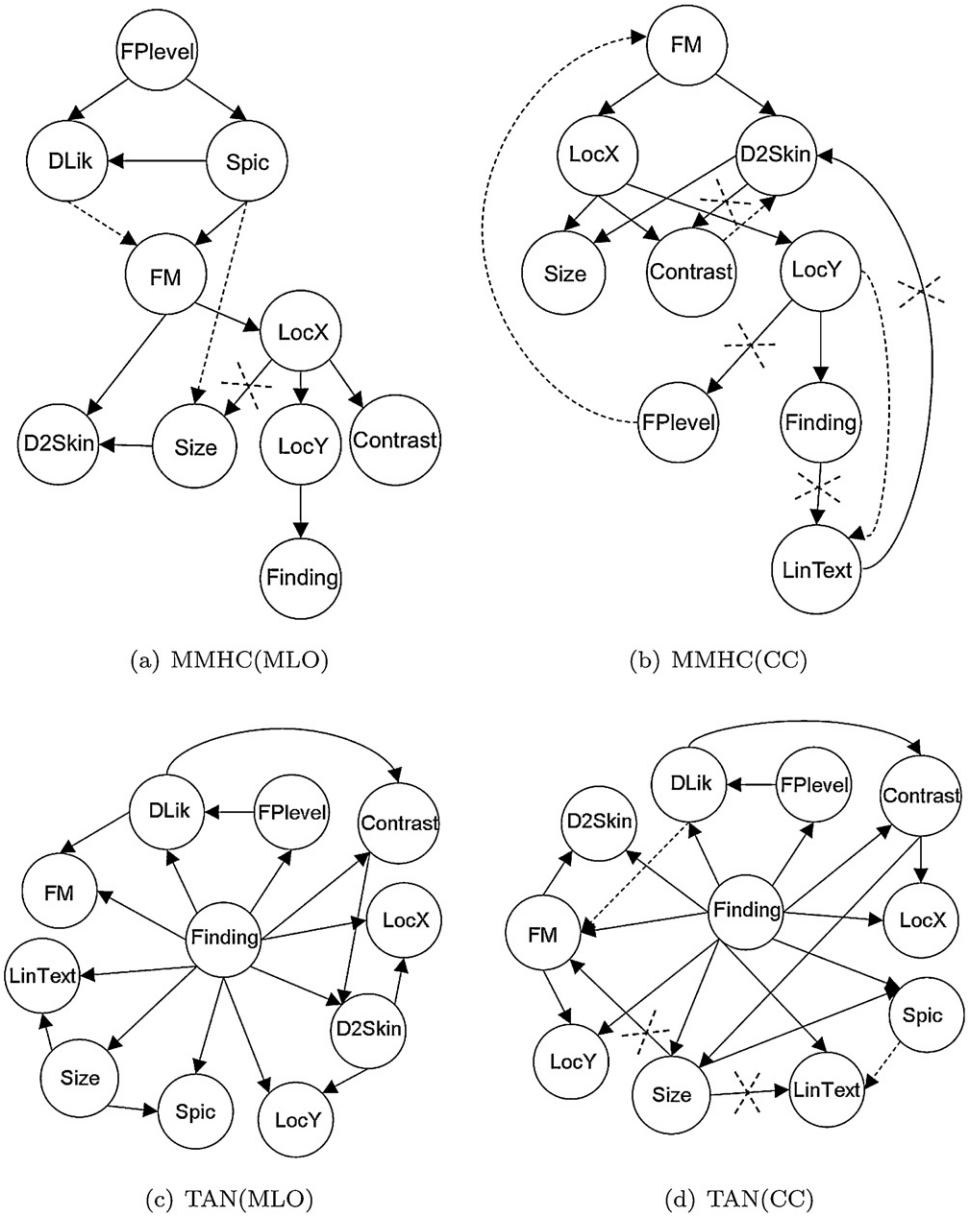


Fig. 8. Structures learnt by MMHC and TAN for MLO and CC views from different data subsets. Dashed graph elements correspond to changes: arcs and nodes indicating addition, and crosses indication deletion in the structure.

Table 3
BDe scores ($\times 10^4$) and AUC obtained from structure learning applied to the discretised and the continuous MLO, CC and combined-view data.

Method	MLO		CC		MLO &CC	
	BDe score	AUC	BDe score	AUC	BDe score	AUC
MMHC-Gaussian	–	0.680	–	0.827	–	0.748
MMHC	−1.326	0.845	−1.143	0.863	−6.857	0.827
TAN	−1.462	0.829	−1.364	0.844	−8.187	0.819

relationships (in the figure this is represented by the grey colouring of the nodes). The first cluster reveals that spiculation and focal mass are directly dependent, which is expected given that these features are pixel-based. The strong dependence of these features with the region size, however, is novel. One explanation may be that spiculated cancerous regions are relatively smaller than non-cancerous abnormalities or false positive detections made by the CAD system. The second cluster contains the region-based features related to the region location, linear texture and contrast. While in the manual network we model some expected dependencies between the spiculation of a region and its linear structure via the hidden variable `Abnormal Structure`, in the structures learnt such dependence is not observed given the other features.

We next checked to what extent these direct dependencies are still present when adding the calculated features to the data. The structures learnt are shown in Fig. 7. We observe that the two clusters of dependencies between the observed features are overall preserved, which indicates their strength. Furthermore, the calculated features tend to have higher dependence with the cluster where the pixel-based features and the region size are present. For `DLik` this finding is not surprising as this feature is computed by the CAD system using only the pixel-based features. However, for `FPlevel` it is interesting to see that most of the dependencies are mostly based on the pixel-based suspiciousness, while this feature is computed by the CAD system using also the other region-based features. While in the MMHC structure `FPlevel` and `DLik` are parent nodes for the observed features, in the GS structure the former features appear also as effect variables.

One of the most striking finding, however, is that in these structures `FPlevel` and `DLik` are not parent nodes of `Finding`, or directly dependent, as modelled in the original manual network and as we observed in the experiments with the expert sub-model. Similarly to the structure in Fig. 6a, the MMHC network in Fig. 7a preserves the dependence of `Finding` on the size and *Y*-location of the region in the CC view, but it also adds the former as a parent of the region *X*-location in the MLO view. For the GS network with added calculated features, the direct dependencies of `Finding` changed to the spiculation and pixel-based suspiciousness of the region in the MLO view.

In summary, these results show that there existed more direct dependencies between the observed mammographic features than were originally represented in the manual network. Furthermore, the dependences between the calculated features and `Finding` are not revealed in the structure learning, indicating that the assumptions made in the manual modelling needed reconsideration.

5.2.3. Classification performance and structure sensitivity on data size

We next apply structure learning on the two folds of each dataset (discrete and continuous), where each fold was used once as a training set to fit the model structure and the parameters and once as a test to compute the fitting BDe score (for the discrete data) and the classification performance in terms of AUC. The average results from both test folds are presented in Table 3 and they show that the hybrid algorithm is capable of finding structures that better fit to the data and tend to distinguish better between cancerous and non-cancerous regions/links than the TAN algorithm. The Gaussian models perform worse in terms of classification accuracy and contain nearly twice as many arcs than the MMHC models obtained from the discrete data (21 vs. 12 for MLO, 18 vs. 11 for CC, and 56 vs. 27 for the combined views), showing that the latter leads to more parsimonious and interpretable models.

We also compare the structures learnt from each of the non-overlapping data subsets used as a training (A) and test (B) set in the evaluation in order to see whether there is a large difference. Fig. 8 shows the resulting structures obtained from MMHC and TAN for the MLO and CC views, where the independent nodes have been omitted. Clearly the structures learnt from the two data subsets are similar with only a few differences. This result is not surprising given the relatively large size of the data, consisting of thousands of observations. To check whether these relationships are preserved for different samples from the data, we performed bootstrapping together with structure learning using MMHC as algorithm. Two hundred samples from the subsets A and B of the MLO and CC data were generated, respectively. The results summarised in Table 4 present the arcs per breast view with average strength and direction larger than 0.5. These results clearly agree with the structures depicted in Fig. 8(a) and (b). Analogous results were obtained from the combined view data.

Table 4
Arcs per breast view with average strength and direction larger than 0.5 obtained from MMHC learning with bootstrapping on 200 samples of each subset A and B. The '*' sign indicates an opposite arc direction.

From	To	Strength		Direction	
		A	B	A	B
(a) MMHC(MLO)					
LocX	Contrast	1.00	1.00	0.87	0.90
FM	LocX	1.00	1.00	0.76	0.88
LocY	Finding	1.00	1.00	0.74	0.89
FPlevel	Spic	1.00	1.00	0.63	0.70
D2Skin	Size	1.00	1.00	0.58	0.73*
Spic	FM	0.98	1.00	0.73	0.87
LocX	LocY	0.95	1.00	0.75	0.88
Spic	DLik	0.91	1.00	0.61	0.60
FM	D2Skin	0.80	0.87	0.92	0.97
LocX	Size	0.74	–	0.67	–
FPlevel	DLik	0.69	1.00	0.81	0.80
D2Skin	Contrast	0.68	–	0.73	–
Spic	Size	–	0.89	–	0.77
DLik	FM	–	0.57	–	0.99
(b) MMHC(CC)					
D2Skin	Size	1.00	1.00	1.00	0.98
LocX	Contrast	1.00	1.00	0.92	0.71
FM	D2Skin	1.00	1.00	0.85	0.83
D2Skin	Contrast	1.00	1.00	0.83	0.58*
LocY	Finding	1.00	1.00	0.68	0.52
FM	LocX	1.00	1.00	0.50	0.51
LinText	D2Skin	0.90	–	0.64	–
FPlevel	LocY	0.90	–	0.61	–
LocX	Size	0.90	–	1.00	–
LocX	LocY	0.81	0.73	0.73	0.63
Finding	LinText	0.70	–	0.82	–
FM	FPlevel	0.56	0.81	0.57	0.58
LocY	LinText	–	0.80	–	0.74

Table 5
Markov blankets of the variable `Finding` for subsets A, B of the dataset and for both subsets.

Method	Subset	MarkovBanket(`Finding`)
MMHC-Gaussian	In both	`CC-FPlevel`, `MLO-DLik`, `CC-DLik`, `MLO-Spic`, `CC-Spic`, `CC-FM` `MLO-Size`, `CC-Size`, `CC-LinText`, `CC-D2Skin`
	Only in A	`MLO-LocY`, `MLO-FPlevel`
	Only in B	`MLO-LocX` `CC-LocX`
MMHC	In both	`CC-Size`, `MLO-LocY`

The network structures obtained from MMHC-Gaussian applied to both data subsets tend to vary more than MMHC on the discrete data, as observed, for example, by the different Markov blankets of the `Finding` variable for the combined data given in Table 5. We observe again that the calculated features are parents of `Finding` in the Gaussian networks, whereas in the discrete network these features do not have direct dependencies with the class variable.

6. Discussion and conclusions

Our aim was to obtain insight into the validity of the modelling assumptions made when developing a BN for complex medical image interpretation problems based on expert knowledge, with the interpretation of mammograms as a real-world example. Where in other problem domains it might be easier to construct such manual models using knowledge engineering methods, in the domain of image interpretation it is not unlikely that modelling mistakes are made. We carried out this study to find out whether data discretisation and structure learning can be used to scrutinise the modelling assumptions to improve the quality of a manually developed BN model.

The decision whether or not to discretise data is not straightforward and it highly depends on the nature of the data and the problem at hand. As mentioned in the introduction, based on the AI journal paper by Pradhan et al. the general wisdom in at least a significant part of the field is that the probabilistic parameters are only of secondary importance [3]. However, our research results show otherwise, namely that discretisation can improve the representation and the accuracy of the models in comparison to the model with continuous variables. First, the discrete data better capture the way radiologists analyse mammograms and evaluate abnormalities. This allows for easy interpretation and usability of the Bayesian network model. Second, appropriate discretisation provides better approximation of the true probability distribution of the data used and avoids the strong Gaussian assumption imposed on the continuous variables, leading to better accuracy and data fitting capabilities of the models, as shown in this study.

The purpose of the structure learning used in this study was to see whether it could be effectively used as a source for critiquing a manually constructed BN and as a means to complement knowledge representation by hand. Such an approach may not always be useful, for example in cases where there is an easy conceptualisation of the problem domain available, or when data are not available. In addition, often representations obtained by machine learning are hard to understand, and structure learning of Bayesian networks is no exception to this general rule. However, this makes the combination of techniques from manual and automatic construction of Bayesian networks even more interesting. The results we achieved clearly show that structure learning results can be conceptually clear and of help in designing a medical BN. First, local interactions between variables in the structures learnt were revealed; some of them were as expected according to the domain knowledge, whereas others were novel and not obvious a priori. Second, the results indicate that manual construction based on expert knowledge offers a good start to build a medical BN, as it guides the selection of important domain factors and such a model will act as point of reference in structure learning.

Finally, we investigated whether learning structures from the discrete data can have an added value to improve the performance of the mammographic analysis modelling. Our results turned out to be positive here as well, strongly indicating the necessity to restructure and re-evaluate the parameters of the originally built manual Bayesian network. The newly revealed graph structure and variables' values provide the basis for the next step in the design process.

In summary, the lessons learnt from this study are that developing a BN for a complex medical problem requires a well-balanced exploration of expert knowledge and data. The development process that results is much more complicated than suggested in previous research, but given the improvement in performance and insight that comes with it, it is also worth the effort.

Acknowledgements

We would like to thank Saskia Robben and Niels Radstake for conducting the initial experiments and providing the preliminary results on discretisation and structure learning. We also thank the reviewers for their useful comments that help improve this paper. This work has been funded by the Netherlands Organization for Scientific Research under BRICKS/FOCUS Grant Number 642.066.605.

References

[1] Velikova M, Samulski M, Lucas PJF, Karssemeijer N. Improved mammographic CAD performance using multi-view information: a Bayesian network framework. Physics in Medicine and Biology 2009;54:1131–47.
[2] Velikova M, Lucas PJF, Samulski M, Karssemeijer N. A probabilistic framework for image information fusion with an application to mammographic analysis. Medical Image Analysis 2012;16(4):857–65.
[3] Pradhan A, Henrion M, Provan G, del Favero B, Huang K. The sensitivity of belief networks to imprecise probabilities: an experimental investigation. Artificial Intelligence 1996;84(1-2):363–97.
[4] Druzdzel MJ, Onisko A. Are Bayesian networks sensitive to precision of their parameters? Proceedings of the international IIS08 conference, intelligent information systems XVI. Warsaw: Academic Publishing House EXIT; 2008. p. 35–44.
[5] Fischer EA, Lo JY, Markey MK. Bayesian networks of BI-RADS descriptors for breast lesion classification. In: Dumont G, Hudson DL, Liang Z-P, editors. Proceedings of the 26th annual international conference of the IEEE EMB Society, vol. 4. Electrical Engineering/Electronics, Computer, Communications and Information Technology Association; 2004. p. 3031–4.
[6] Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results. Journal of Radiology 2006;240(3):666–73.
[7] Robben S, Velikova M, Lucas PJF, Samulski M. Discretisation does affect the performance of Bayesian networks. In: Bramer M, Petridis M, Hopgood A, editors. Proceedings of the 30th SGAI international conference on artificial intelligence, research and development in intelligent systems XXVII. 2011. p. 237–50.
[8] Radstake N, Lucas PJF, Velikova M, Samulski M. Critiquing knowledge representation in medical image interpretation using structure learning. Lecture Notes in Artificial Intelligence 2011;6512:56–69.
[9] Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ. Probabilistic networks and expert systems. New York, USA: Springer-Verlag; 1999.
[10] Jensen FV, Nielsen TD. Bayesian networks and decision graphs. second ed. New York, USA: Springer-Verlag; 2007.
[11] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge, MA, USA: The MIT Press; 2009.
[12] Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajcsy R, editor. Proceedings of the 13th international joint conferences on artificial intelligence. Morgan Kaufmann; 1993. p. 1022–7.
[13] Geurts P, Wehenkel L. Investigation and reduction of discretization variance in decision tree induction. Lecture Notes in Computer Science 2000;1810:162–70.
[14] Flores JL, Inza I, Larrañaga P. Wrapper discretization by means of estimation of distribution algorithms. Intelligent Data Analysis 2007;11(5):525–45.
[15] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. In: Prieditis A, Russell S, editors. Proc. of the 12th international conference on machine learning. 1995. p. 194–202.
[16] Mizianty M, Kurgan L, Ogiela M. Comparative analysis of the impact of discretization on the classification with naïve Bayes and semi-naïve Bayes classifiers. In: Arif Wani M, Xue wen Chen D, Casasent L, Kurgan T, Hu K, Hafeez, editors.

Proc. of the 7th international conference on machine learning and applications. Los Alamitos, CA, USA: IEEE Computer Society; 2008. p. 823–8.
[17] Abraham R, Simha JB, Iyengar SS. A comparative analysis of discretization methods for medical datamining with naïve Bayesian classifier. In: Mohanty SP, Sahoo A, editors. Proc. of the 9th international conference on information technology. Los Alamitos, CA, USA: IEEE Computer Society; 2006. p. 235–6.
[18] Ismail MK, Ciesielski V. An empirical investigation of the impact of discretization on common data distributions. In: Abraham A, Köppen M, Franke K, editors. Proc. of the 3rd international conference on hybrid intelligent systems. 2003. p. 692–701.
[19] Robinson RW. Counting unlabeled acyclic digraphs. Combinatorial Mathematics V 1977;622:28–43.
[20] Andersson SA, Madigan D, Perlman MD. A characterization of Markov equivalence classes for acyclic digraphs. Annals of Statistics 1997;25:505–41.
[21] Pe na JM, Nilsson R, Björkegren J, Tegnér J. Towards scalable and data efficient learning of Markov boundaries. International Journal of Approximate Reasoning 2007;45(2):211–32.
[22] Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part I: algorithms and empirical evaluation. Journal of Machine Learning Research 2010;11:171–234.
[23] Rodrigues de Morais S, Aussem A. A novel Markov boundary based feature subset selection algorithm. Neurocomputing 2010;73(4–6):578–84.
[24] Heckerman D. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research; 1995.
[25] Meloni A, Ripoli A, Positano V, Landini L. Mutual information preconditioning improves structure learning of Bayesian networks from medical databases. IEEE Transactions on Information Technology in Biomedicine 2009;13(6): 984–9.
[26] Mansinghka VK, Kemp C, Tenenbaum JB. Structured priors for structure learning. In: Dechter R, Richardson TS, editors. Proc. of the 22nd conference on uncertainty in artificial intelligence. Cambridge, MA, USA: Uncertainty in Artificial Intelligence Press; 2006.
[27] Flores J, Nicholson AE, Brunskill A, Korb KB, Mascaro S. Incorporating expert knowledge when learning Bayesian network structure: a medical case study. Artificial Intelligence in Medicine 2011;53(3):181–204.
[28] Acid S, de Campos LM, Fernandez-Luna JM, Rodriguez S, Rodriguez JM, Salcedo JL. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. Artificial Intelligence in Medicine 2004;30(3):215–32.
[29] Oh JH, Craft J, Al Lozi R, Vaidya M, Meng Y, Deasy JO, et al. A Bayesian network approach for modeling local failure in lung cancer. Physics in Medicine and Biology 2011;56:1635–51.
[30] Rajapakse JC, Zhou J. Learning effective brain connectivity with dynamic Bayesian networks. Neuroimage 2007;37(3):749–60.
[31] Wiggins M, Saad A, Litt B, Vachtsevanos G. Evolving a Bayesian classifier for ECG-based age classification in medical applications. Applied Soft Computing 2008;8(1):599–608.
[32] Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191:241–4.
[33] Breast Imaging Reporting and Data System (BI-RADS). Reston, VA: American College of Radiology; 1993.
[34] Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. The New England Journal of Medicine 2008;359(16):1675–84.
[35] Timp S. Analysis of temporal mammogram pairs to detect and characterize mass lesions. PhD thesis, Radboud University Nijmegen; 2006.
[36] Paquerault S, Petrick N, Chan H, Sahiner B, Helvie MA. Improvement of computerized mass detection on mammograms: fusion of two-view information. Medical Physics 2002;29(2):238–47.
[37] Qian W, Song D, Lei M, Sankar R, Eikman E. Computer-aided mass detection based on ipsilateral multiview mammograms. Academic Radiology 2007;14(5):530–8.
[38] Good W, Zheng B, Chang Y, Wang X, Maitz G, Gur D. Multi-image CAD employing features derived from ipsilateral mammographic views. In: Hanson KM, editor. Proc. of SPIE. Medical imaging, vol. 3661. 1999. p. 474–85.
[39] van Engeland S, Karssemeijer N. Combining two mammographic projections in a computer aided mass detection method. Medical Physics 2007;34(3):898–905.
[40] Ferreira N, Velikova M, Lucas PJF. Bayesian modelling of multi-view mammography. In: Hauskrecht M, Schuurmans D, Szepesvari C, editors. Proc. of the international conference on machine learning workshop on machine learning for health-care applications. 2008.
[41] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. second ed. San Francisco, CA, USA: Morgan Kaufmann; 2005.
[42] Murphy K. Bayesian Network Toolbox (BNT). https://code.google.com/p/bnt/; 2007 [accessed August 2012].
[43] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 1977;39(1):1–38.
[44] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997;30(7):1145–59.
[45] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 2006;65(1): 31–78.
[46] Margaritis D. Learning Bayesian Network Model Structure from Data. PhD thesis, Carnegie-Mellon University; 2003.
[47] Scutari M. Learning Bayesian networks with the bnlearn R package. Journal of Statistical Software 2010;35(3):1–22.