Critiquing Knowledge Representation in Medical Image Interpretation Using Structure Learning

Niels Radstake[1], Peter J.F. Lucas[1], Marina Velikova[1], and Maurice Samulski[2]

[1] Radboud University Nijmegen, Institute for Computing and Information Sciences
{peterl,marinav}@cs.ru.nl
[2] Radboud University Nijmegen Medical Centre, Department of Radiology
m.samulski@rad.umcn.nl

Abstract. Medical image interpretation is a difficult problem for which human interpreters, radiologists in this case, are normally better equipped than computers. However, there are many clinical situations where radiologist's performance is suboptimal, yielding a need for exploitation of computer-based interpretation for assistance. A typical example of such a problem is the interpretation of mammograms for breast-cancer detection. For this paper, we investigated the use of Bayesian networks as a knowledge-representation formalism, where the structure was drafted by hand and the probabilistic parameters learnt from image data. Although this method allowed for explicitly taking into account expert knowledge from radiologists, the performance was suboptimal. We subsequently carried out extensive experiments with Bayesian-network structure learning, for critiquing the Bayesian network. Through these experiments we have gained much insight into the problem of knowledge representation and concluded that structure learning results can be conceptually clear and of help in designing a Bayesian network for medical image interpretation.

1 Introduction

The past decade has seen a transition in radiology from film-based storage of images to digitised, computer-based storage. The digitisation of medical images has offered a unique opportunity for adding computer-aided support to the traditional, human interpretation of medical images by radiologists. Although radiologists are well-trained for the task of image interpretation, there is room for improvement as misinterpretation of medical images is far from rare. Computers are successfully used in many areas of health care; however, it has been hard to match, and certainly surpass, the capabilities of expert radiologists in interpreting medical images. Medical images are noisy and patient specific, and, thus, computers have difficulty in coping with them.

The research described in this paper focuses on one of such hard medical image interpretation tasks: the interpretation of X-ray images of the breasts, usually called mammograms, for breast-cancer detection. Although there has been considerable progress in the last decade in computer-aided interpretation of mammograms, most of the improvement have come from new pattern recognition techniques which detect potentially suspicious breast regions. The proper

interpretation of mammograms, however, requires an approach similar to the one used by radiologists, who 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.

Bayesian networks have been used in our research as they permit integrating knowledge and information from different sources. In the early stages of our research, we therefore decided to design a Bayesian network that incorporated the most important image features of the two mammograms available for each breast; construction of its graph structure was guided by expert knowledge. This Bayesian network can, thus, be looked upon as a *knowledge representation* of mammogram interpretation: it offers a compact representation of how the features extracted from an image are interpreted in terms of breast tissue architecture and presence of masses (mammographic abnormality signs).

Extensive experimentation with the Bayesian network using image data, however, yielded disappointing results, which we did not fully understand. To gain more insight into the structure of the Bayesian network, we subsequently carried out work on learning Bayesian network structures, both restricted and unrestricted, from image data with the aim to improve the structure of the network. In this paper we discuss at length how structure learning succeeded in achieving this goal. The results can be seen as an application of learning for the purpose of knowledge critiquing.

The structure of the paper is as follows. In the next section, the issue of mammogram interpretation is reviewed. In Section 3 we briefly summarise the principles underlying Bayesian network representation and learning, and present the Bayesian network for mammogram interpretation. In Section 4, the different experimental methods and associated results are discussed. Lessons learnt are given attention too in Section 5.

2   Background

2.1   Mammographic Analysis

Mammography is the diagnostic procedure to detect breast cancer in the breasts using low-dose X-rays. The resulting mammograms are made using different projections, also called *views*. The most common views of the breast are *mediolateral oblique* (MLO) and *craniocaudal* (CC); see Figure 1. The MLO view is a 45° angled side view, showing a part of the pectoral muscles. The CC view is an projection of the breast from above with the nipple centered in the image.

Because a mammogram is a projection of the breast, its layers of breast tissue are superimposed. The X-ray attenuation, which is due to absorption and scattering of photons, describes the density of a region. This results in contrast, or whiteness, of a region, on the mammogram. The darker areas of the breast are non-dense and consist mainly of fatty tissue. The lighter areas are dense and contain lobules, ducts, and possibly masses.

The interpretation of mammograms by radiologists produces *regions of interest*, or *regions* for short. A region is also referred to as a *lesion* or an *abnormality*.
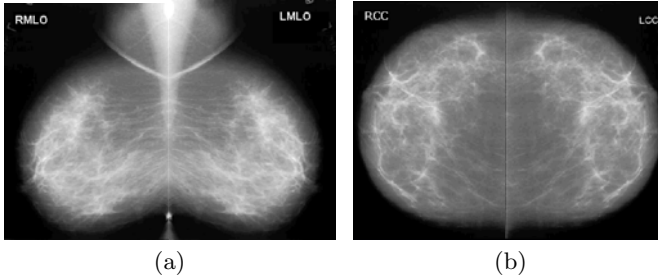
Fig. 1. Right and left breasts in (a) MLO view and (b) CC view

2.2   Feature Extraction and Computer-Aided Detection

Studies have shown that radiologists fail to identify a significant number of cases with breast cancer, i.e., false negatives, due to misinterpretation. The reasons for these misses are unclear [3]. Audits have shown that abnormalities that are clearly visible in retrospect must have been overlooked or its signs were misinterpreted.

To increase the detection rate, *computer-aided detection* (CAD) systems are being developed. These systems use pattern recognition techniques to extract *features* in a mammogram, which are subsequently used to identify regions that are possibly suspicious. With such markings, the CAD system can assist the radiologist while analysing mammograms with the detection of breast abnormalities.

The CAD system [3] we have used employs four steps to classify regions: (1) mammogram segmentation into 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. Note that in practice most patients with breast cancer have one or two cancerous regions at most. The CAD system, however, often finds more, which then are false positives. There reason for these false positives is that the CAD system uses local information only to determine whether a region is suspicious. Complementary information from the other view, or previous mammograms, would allow concluding whether a positive region is true or false positive, and this is the way radiologists work.

The region features used in this study can be categorised into two groups:

A. *Observed features* extracted from the image in step (3) mentioned above:

- The relative location of the region (LocX and LocY);
- The shortest distance of the region to the skin (d2skin);
- Contrast;
- The presence of radiodensity similar to that of adjacent tissue (IsoDens);
- Spiculation, indicating whether the region margin has a spiky pattern;
- The presence of a circumscribed lesion (FocalMass);
- Linear texture (LinTex), which is typical for normal breast tissue;
- Size of the region (RegSize).

B. *Calculated features*, computed from classifiers based on pixel- or region-based features:

- The malignancy pixel-based likelihood (MassLik);
- The false-positive level of a region (FPLevel), indicating the average number of normal regions in an image with the same or higher likelihood scores.

3 Bayesian Network Principles

3.1 Bayesian Networks for Knowledge Representation

Consider a finite set of random variables X, where each variable X_i in X takes on values from a finite domain $dom(X_i)$ and let P be a joint probability distribution of X. A *Bayesian network* $\mathcal{B} = (G, P)$, BN for short, is a probabilistic graphical model that represents conditional independence assumptions in the form of an acyclic directed graph, ADG for short, G; it is assumed that those conditional independences are obeyed by the associated joint probability distribution P, and P is then called Markov over G [9]. The graph $G = (\mathbf{V}, \mathbf{A})$ is represented by a set of nodes $\mathbf{V}$ corresponding one to one to the random variables in X and a set of arcs $\mathbf{A} \subseteq (\mathbf{V} \times \mathbf{V})$ corresponding to direct causal relationships between the variables. Independence information is modelled in an ADG by blockage of paths between nodes in the graph by other nodes. BNs have the virtue that they can be both manually constructed and learnt from data. Manual construction is usually guided by interpreting arcs in Bayesian networks as causal relationships.

Initially in the research we constructed a Bayesian network based on available domain knowledge, shown in Figure 2 [4]. The BN incorporates the features described above and it is capable of interpreting MLO and CC features at the same time, allowing the integration of information from two views. The simultaneous interpretation of the MLO and CC features is modelled by the corresponding *hidden* variables (in light grey in the figure), which are not directly observed or measured in the CAD system, but represent the way 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 AbDensity and AbStruct, indicating the presence of abnormal density and structure and they have two states: "present" and "absent". Furthermore, since the two calculated features MassLik and FPLevel are extra overall indicative measures for suspicious regions, we use them as conditional variables to determine a priori the probability of having a finding, modelled by their incoming arcs to the variable Finding. On the other hand, every observed feature partially characteristizes a finding and this is represented by the causal arcs outgoing from Finding.

3.2 Structure Learning

Structure learning is basically finding the Bayesian network graph, or *structure* as is often said, that fits the data best. The number of acyclic directed graphs
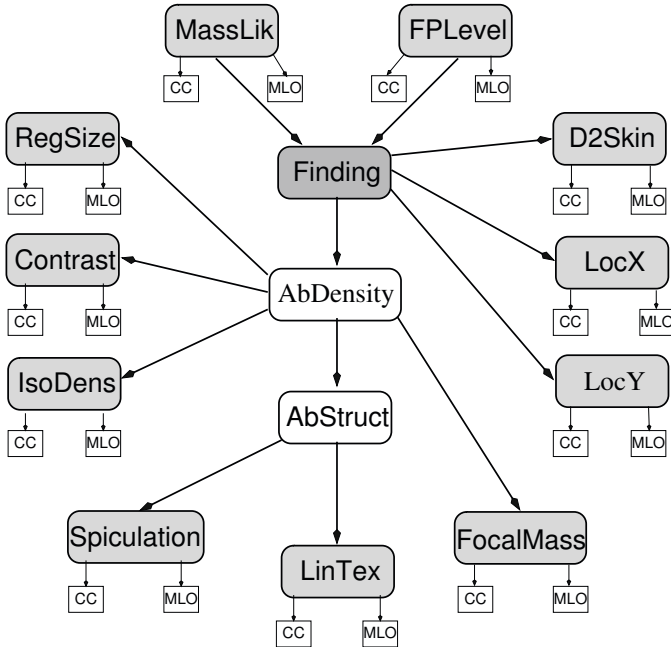
Fig. 2. Feature-based BN model for the interpretation of mammograms; indicated is that each feature node is linked to a corresponding CC and MLO feature [4]

is, however, more than exponential in the number vertices of the graph [11]. Exhaustively search for the best graph is, therefore, infeasible for most problems. However, for this problem, which concerns 11 variables, we are on the edge of what is still possible. Removing only 2 variables would make it feasible to carry out exhaustive search, although such would, of course, be very time consuming.

Structure learning is an optimisation problem, where a *score measure* is used to judge the fitness of the model, and a search method allows exploring the search space of acyclic directed graphs. The score measure always includes some measure of the likelihood of the data given the graph and its probabilistic parameters, $\Pr(D \mid G, P)$, where D are the data, or the marginalised likelihood $\Pr(D \mid G)$, where the parameters are marginalised out. In addition, they typically include the possibility to include a prior on the structure, $\Pr(G)$, and a penalty for unwanted complexity of graph structure. The two score measures used in this research are the *Bayesian score*, which is based on marginalised likelihood, and the *Bayesian Information Criterion* (BIC), which is likelihood based [6]. These measures take into account that graphs, though different in structure, may encode the same conditional independence assumptions, i.e., are *Markov equivalent* as is said [5].

As was argued above, use of exhaustive search is uncommon; more common is the use of greedy search, which searches the space of ADGs, or the space of equivalent classes of structures, called *essential graphs* (EG), i.e., greedy search

in EG space. One of the first, and still popular, structure learning methods is the K2 algorithm introduced in [2]. It is a special case of a greedy ADG search algorithm; it minimises the search space by having an initial order on the nodes and by restricting the number of parents a node can have.

In addition to structure learning, we also explored three simpler BN structures for comparison: a fully disconnected graph (all variables are independent), naïve Bayes (NB), and tree-augmented network (TAN). In this order, the network structures are expected to give an increasingly better fitting model, although not as good as the BNs obtained through structure learning.

4 Structure Learning from Mammographic Data

Here, we describe the following set of experiments with actual mammographic data in order to explore various knowledge representation schemes: (i) learning different structures based on a hand-constructed expert sub-model, (ii) comparing the structures learnt based on the observed and calculated features, (iii) modifying the greedy search algorithm for mammographic structure learning, and (iv) studying the robustness of the structures learnt based on different data subsets.

4.1 Data and Experimental Set-Up

The image dataset used here was obtained from the Dutch breast cancer screening programme and contained data of 1063 cases, of which 383 were cancerous as confirmed by pathological reports. For each case, both the CC and MLO views were present. For each mammogram, the 5 most suspicious regions were selected. In the experiments, structures were learnt using separately the CC data and MLO data. The dataset is divided into a training set, used for learning the models, and a test set, used for scoring the models afterwards. These sets have an equal distribution of cancerous regions. The experiments described in this section were performed using the Bayes Net Toolboxes (BNT) [8], [7].

Most structure learning implementations work with discrete values for the variables. The features in the dataset were real-valued, so we discretised them using a histogram algorithm built in [7], which finds an optimal number of bins according to a cost function based on Akaike's criterion [1]. Since for some of the variables, the obtained number of bins was too high (up to 33), we conducted additional structure testing experiments with the resulting discretised data in order to obtain reasonable discrete ranges. To see the influence of discretisation, we learnt TAN and GS structures and estimated the probability distributions based on various datasets for which the maximum number of bins was varied from 2 to 20. For every learnt structure, we computed the Bayesian score as a measure for fitting the data and the area under the receiver operating characteristic curve (AUC) as a measure for classification performance. The results indicated that the data fitting and accuracy capabilities of the structures learnt worsen, especially when the number of bins per variable was larger than 10. Considering the most optimal results obtained from both TAN and GS algorithms, we have

restricted the final number of values between 2 and 7. We want to emphasise that discretisation algorithms were not the main topic of the reported research, which is why we do not go into detail (cf. [10] for more detail); discretisation was rather used as a preprocessing step to facilitate structure learning.

For the K2 algorithm, the node ordering was derived from the expert model and other orderings were studied using the BIC score. For the greedy search algorithm an empty network was used as an initial structure.

4.2   Results

Learning structures based on an expert model. A sub-model containing 5 MLO variables was selected from the expert model (cf. Figure 3(a)), so that it was possible to perform exhaustive search. The Bayesian score of this sub-model is −48332. The fitness of the models learnt was compared to the two reference models–fully disconnected and naïve Bayes–which do not consider any of the knowledge incorporated in the expert model. Next we present the structures learnt by various algorithms using the 5 variables from the expert sub-model.

The *TAN algorithm* learnt the structures using the fixed class node Finding. All possible TAN structures were learnt and one of them using FPLevel as an ancestor to the remaining variables is shown in Figure 3(b), with a Bayesian score of −44915.

For the *K2 algortihm* all possible node orderings were investigated. The model with the highest score is shown in Figure 3(c). For the initial structure of the *greedy search algorithm* (GS), different network structures have been used: (i) a fully disconnected structure, (ii) naive Bayes, and (iii) a structure learnt using K2. They all resulted in a network structure equivalent to the structure learnt by K2. Finally, we learnt the optimal network structure using *exhaustive search*. All possible network structures with 5 variables (29281) were scored and the best performing model found with exhaustive search is chosen, which is the same model found using K2 and GS. The Bayesian score for this model is −43941. These results imply that K2, GS, and exhaustive search algorithms have found the structure that fits better to the data than the TAN algorithm. Since the search space of all possible ADGs is relatively small and the model is very restricted, this is not surprising.

The reference models–fully disconnected and naïve Bayes–yielded Bayesian scores of −49006 and −48077, respectively, indicating a fit to the data worse than for the structure-learnt models.

When considering the resulting structures in comparison to the expert model in Figure 3(a), it can be observed that in the model learnt by K2, GS and exhaustive search Finding is conditioned on MassLik and FPLevel, as in the expert model, but the location variables LocX and d2skin are not conditioned on Finding and they have a direct causal relationship with the two calculated features. For the TAN model, fixing the class node to be Finding results in a structure where all the remaining features were conditioned on this node.

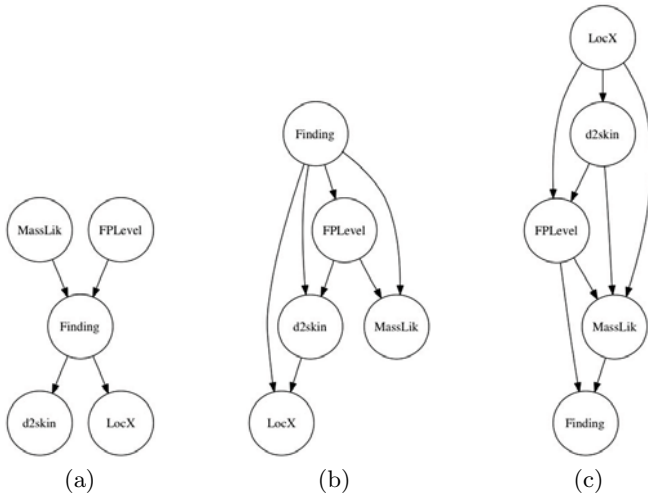**Fig. 3.** Structures based on (a) expert sub-model, learnt by (b) TAN, and (c) K2, GS and exhaustive search

Table 1. Bayesian scores ($\times 10^4$) using all and observed variables

Method	All variables		Observed variables	
	CC	MLO	CC	MLO
NB	-9.9787	-10.1710	-8.0157	-8.1824
TAN	-9.5442	-9.7866	-7.7521	-7.9304
K2	-9.4059	-9.6489	-7.5682	-7.7859
GS	-9.3385	-9.6023	-7.5612	-7.7728

Influence of calculated features. In this experiment we investigated the influence of the calculated features, MassLik and FPLevel, on the network structures when learning models from data. Models were learnt using all 11 variables (observed and calculated) and using only observed variables. The results are shown in Table 1. The best performing algorithm is GS whose resulting structures from both the observed and calculated features are shown in Figure 4.

A closer look at the structures learnt revealed that Finding is conditioned only on FPLevel and does not have any children. This means that Finding is conditionally independent of the remaining features given the false-positive level of the region. Hence, the entire structure could be replaced by the very simple model: FPLevel → Finding when the false-positive level is known. This is not a surprising result as FPLevel is the outcome of the neural network classifier of the CAD system to predict the likelihood for cancer and one would expect a strong dependence with Finding.

Another observation is that the features LocX, LocY and d2skin, describing the location of the region in the breast, are related in all learnt models. In some models, especially those learnt using CC data, these variables are independent
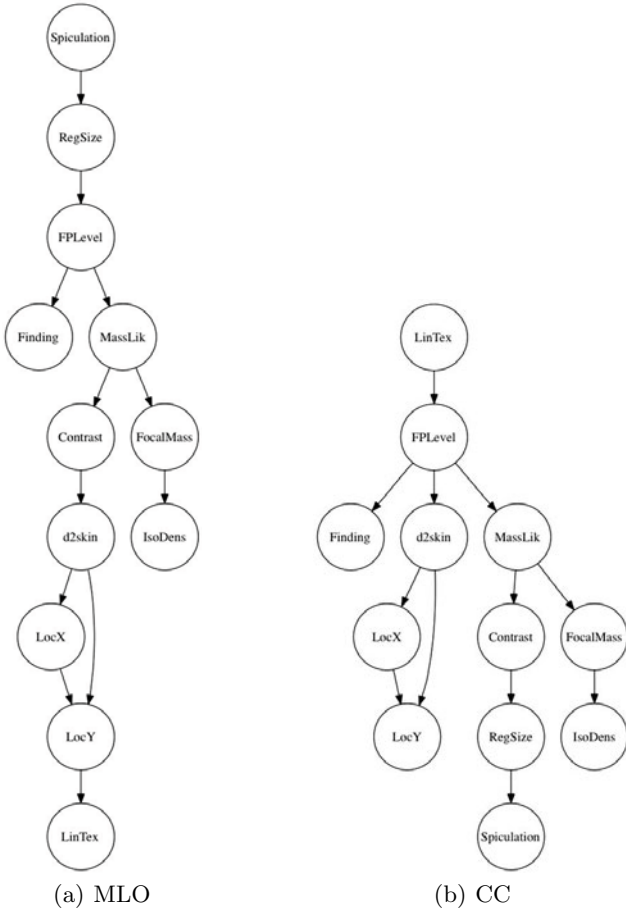
(a) MLO   (b) CC

Fig. 4. ADG structures learnt by the GS algorithm with the observed and calculated features

of the other variables. It was also expected that the structures learnt would reveal causal relationships between Spiculation and LinTex as these features are relatively complementary to each other: if linear texture is present, the region is not spiculated and vice versa. However, in only 25% of the learnt models this relation was present.

Modifying greedy search. In the previous experiments we observed that in all cases where MassLik and FPLevel variables are present, Finding becomes conditioned on FPLevel. Here we learnt structures using greedy search based on data without including FPLevel and MassLik in the learning process, but only in the scoring step of the algorithm. The greedy search algorithm started with a initial network structure G, which consisted of the nodes Finding, LocX, LocY, d2skin, Contrast, Isodense, Spiculation, FocalMass, LinTexN, and RegSize without
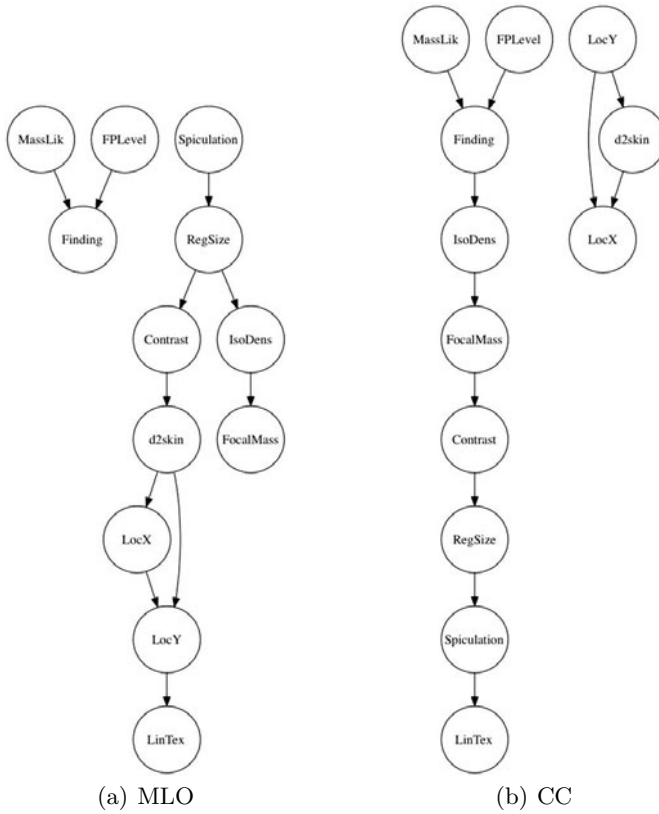
(a) MLO   (b) CC

Fig. 5. ADG structures learnt by the modified GS algorithm

arcs. For each step, it defined a set of neighborhood graphs **NG**. A copy of this set **NG′** is made and each DAG in **NG′** is modified by adding FPLevel and MassLik as conditions on Finding: FPLevel → Finding ← MassLik. For each modified DAG, the score is computed. The (modified) graph with the highest score was selected and its (unmodified) original version was used for the next iteration. The search was stopped when there was no neighborhood network graph with a higher score than the current structure. For the CC data the Bayesian score was -9.5461 ($\times 10^4$), whereas for the MLO data the Bayesian score was -9.7830 ($\times 10^4$). The final networks are depicted in Figure 5.

For the three location variables we observed again a strong causal relationship, discovered for both views and for CC they were independent from the remaining features. The relationship IsoDens → FocalMass persisted in both view structures, indicating that the presence of focal mass is dependent on the presence of isodensity. Furthermore, Spiculation and Contrast appeared to be conditionally independent given RegSize for both view structures as indicated by the paths between these variables. This is an interesting result revealing that knowledge
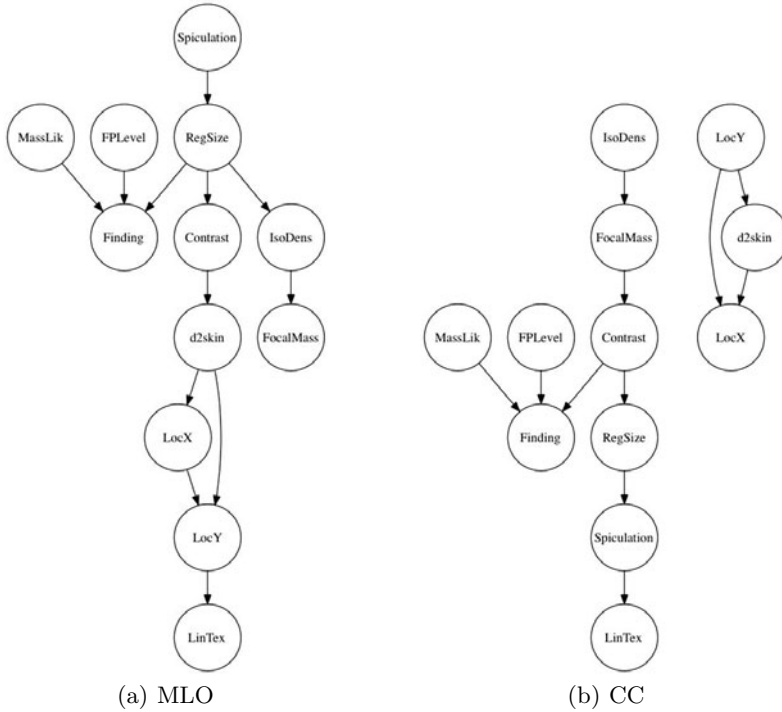
(a) MLO   (b) CC

Fig. 6. ADG structures learnt by the modified GS algorithm by adding the calculated features after learning

about the size of a region would determine the region spiculation and contrast features. We also note that for the MLO view the calculated features are the only determinants whether or not a finding is present, whereas the remaining features are independent of this node.

Comparable results were obtained using a modification of the greedy search algorithm by not including MassLik and FPLevel in the learning process but adding them only afterwards to the learnt structure; see Figure 6. The resulting structure for MLO, with a Bayesian score of -9.5819 ($\times 10^4$), differed from the one in Figure 5(a) by having the additional arc RegSize $\rightarrow$ Finding. For CC the structure included, in comparison with the one in Figure 5(b), the arc Contrast $\rightarrow$ Finding and excluded the arc Finding $\rightarrow$ IsoDens, and its Bayesian score is -9.9670 ($\times 10^4$). These results demonstrated that using MassLik and FPLevel in the learning or scoring step of the structure building process make them the only causes of Finding, confirming the strong impact of the calculated features.

Comparing structures learnt from different datasets. Given the limited sample of data and the split of training and testing data, we next explore to what extent the structures learnt from various data subsets differ. We perform TAN structure learning from non-overlapping subsets of the MLO data with 3 different
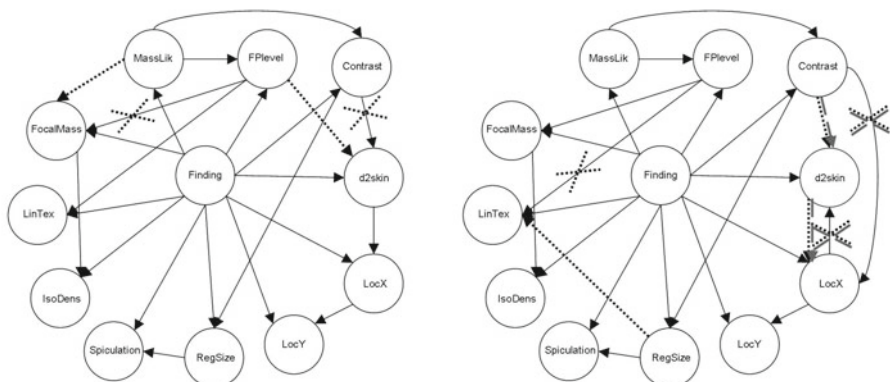
(a) ∼ 10500 observations per dataset   (b) ∼ 2600 observations per dataset
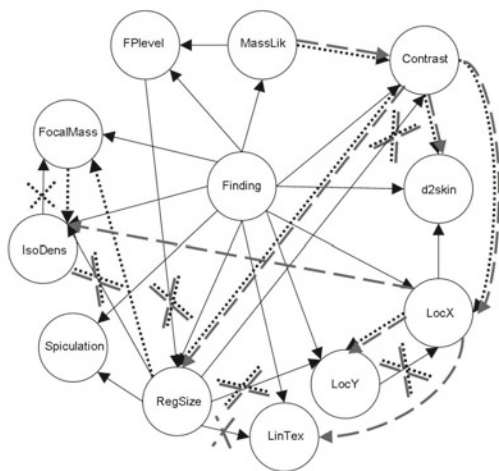
(c) ∼ 650 observations per dataset

Fig. 7. TAN structures learnt from MLO view data with different sizes of data subsets: (a) two structures based on the two subsets (∼ 10500 observations per set) obtained from the split of the whole data, (b) three structures based on four subsets containing ∼ 2600 observations per set and (c) three structures based on three subsets containing ∼ 650 observations per set. The differences in the structures are shown with the dotted and dashed lines: arcs (addition) and crosses (deletion).

sizes: 2 sets each containing 50% of the whole data (∼ 10500 observations), 4 sets with 12% of the data (∼ 2600 observations) and 8 sets with 3% of the data (∼ 650 observations). The region data for a particular case was contained in only one subset and the proportion of cancerous and normal cases in the subsets was the same as in the whole dataset. We naturally expect that the larger the data samples the more robust and similar the structures learnt from them would be

in comparison with smaller datasets where the data variations are to be larger. The results, presented in Figure 7, are in line with our hypothesis.

Figure 7(a) depicts the two structures learnt from the two halves of the original data and we observe differences in the incoming arcs to only two nodes: FocalMass and d2skin. Figure 7(b) depicts the three structures based on 4 subsets of the data with 12% of the observations. Note that for two of these subsets the structures learnt were identical but for the others more changes in the causal relationships are observed (indicated by the dotted and dashed arcs) than using half of the data. Finally using 8 random samples containing 3% of the data yielded structures with less robust causal relationships; Figure 7(c) depicts the three structures with the largest differences among them. Only for two of all 8 data samples the same structures were learnt.

It is also interesting to study the causal relationships learnt from the samples with different sizes. One strong direct dependence persistent among the structures, also discovered by the GS algorithm in both views as shown in Figures 5, and 6, is between the region size, spiculation and contrast feature. On the one hand, this general feature dependence revealed by the data is also partially represented in the expert causal model via the unknown hidden variables AbStruct and AbDensity. On the other hand, the structures learnt indicate direct dependence relationships between these variables, not explicitly captured in the manually constructed model. Similarly the causal dependences between the location features appeared very strong in data as indicated by the similar substructures learnt from different structure learning algorithms and data subsets. Furthermore, the substructure FPLevel ← MassLik→Contrast also shows that the region contrast in conditionally independent of the region likelihood for cancer given its pixel-based likelihood.

These results overall indicate that the expert model represents the main relationships between the mammographic features as guided by the domain knowledge but it also lacks certain direct dependences, which is to affect the knowledge representation and the model's performance.

5   Discussion and Conclusions

Although comparisons between manually constructed and learnt BNs are standard practice, in particular in studies where the performances of structure-learning algorithms are compared, the purpose of the present research was different, namely to see whether structure learning could be effectively used as a source for critiquing a manually constructed BN. Thus, here learning methods were used 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.

In this research we dealt with a problem domain that is very hard from a conceptual point of view: the interpretation of medical images. Whereas in other domains it might be easier to construct manual models using knowledge engineering methods, in the domain of image interpretation it is not unlikely that mistakes are made in the conceptualisation. We carried out this study to find out whether structure learning could be of any help in this case, and a positive answer would only be arrived at if the results obtained had a clear meaning.

The results we achieved clearly show that structure learning results can be conceptually clear and of help in designing a Bayesian network for image interpretation. First, local interactions between variables in the structures learnt were revealed, where some of them were expected based on the domain knowledge, whereas others were novel and not obvious a priori. Second, the results also indicate that manual construction based on expert knowledge is a good start to build a Bayesian network for medical image interpretation, guiding us in the selection of the important factors playing a role in the domain and providing a good basis for comparison with the structures learnt. Finally, we observed that the inclusion of calculated features diminished the explanatory power of the remaining features and obscured their meaning in the problem of mammogram interpretation.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
2. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347 (1992)
3. Van Engeland, S.: Detection of mass lesions in mammograms by using multiple views. PhD thesis, Radboud University Nijmegen, The Netherlands (2006)
4. Ferreira, N., Velikova, M., Lucas, P.J.F.: Bayesian modelling of multi-view mammography. In: Proc. of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications (2008)
5. Flesch, I., Lucas, P.J.F.: Markov equivalence in Bayesian networks. In: Lucas, P., Gomez, J., Sameron, A. (eds.) Advances in Probabilistic Graphical Models, pp. 3–38 (2007)
6. Heckerman, D.: A tutorial on learning with Bayesian networks. Technical report MSR-TR-95-06. Microsoft Research (1995)
7. Leray, P., Francois, O.: BNT structure learning package: documentation and experiments. Technical Report, Laboratoire PSI - INSA Rouen (2004)
8. Murphy, K.: Bayesian Network Toolbox for Matlab (2002)
9. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo (1988)
10. Robben, S., Velikova, M., Lucas, P.J.F., Samulski, M.: Discretisation does affect the performance of Bayesian networks. In: Proceedings AI 2010. Springer, London (2010)
11. Robinson, R.W.: Counting unlabeled acyclic digraphs. Combinatorial Mathematics V, vol. 622, pp. 28–43. Springer, Berlin (1977)