Exploring Discretization Techniques for Breast Cancer Detection with Bayesian Networks

Bachelor Thesis

Saskia Robben (0316717)
saskiarobben@student.ru.nl

April 27, 2010

Supervisors:
Dr. P.J.F. Lucas
Dr. M. Velikova
Dr. N de Carvalho Ferreira
Dr. I.G. Sprinkhuizen-Kuyper

Radboud University Nijmegen
Artificial Intelligence: Cognitive Science

Contents

1 Introduction **3**
- 1.1 Problem Description . . . 3
- 1.2 Research Question . . . 4
- 1.3 Outline . . . 4

2 Background **4**
- 2.1 Mammography and CAD . . . 4
- 2.2 Related Work on CAD . . . 6
- 2.3 Bayesian Network Theory . . . 6
- 2.4 Causal Model . . . 7
- 2.5 Discretization . . . 8

3 Methods **9**
- 3.1 Discretization Algorithms . . . 9
- 3.2 Data Description . . . 10
- 3.3 Experimental Setup . . . 11
- 3.4 Analysis . . . 11

4 Results **12**

5 Conclusion and Discussion **14**

References **17**

Abstract

Cumputer Aided Detection systems (CAD) are assisting radiologists with deciding whether a detected anomaly is malignant. The current CAD systems are detecting most cancers, but false positives are the biggest problem. In a collaboration with the Radboud University Nijmegen Medical Centre (UMCN) and the computer science department of the Radboud University Nijmegen a multi-stage CAD system has been developed. The final stage is a Bayesian Network which elaborates on features like 'contrast' or 'size' of a suspicious region. In this stage two views of the same region are regarded simultaneously. In this thesis we improve this causal model by discretizing the variables. Both to capture the underlying probability distribution and to aid usability of the CAD system since radiologist would typically annotate a region in a categorical fashion (e.g. 'high' or 'very low' contrast). Classification performance is determined using ROC curves. A few algorithms perform better than continuous baseline, best was the entropy based method of Fayyad and Irani, but also simpler algorithms can outperform continuous baseline. Two simpler methods with only 3 bins per variable gave results similar to continuous baseline. This indicates that usablity can be improved without decline in performance.

1 Introduction

1.1 Problem Description

Breast cancer is a common disease among women, and also a major cause of death. The success of therapy depends on an early detection of breast cancer. Therefore screening programs exist in many countries. A very effective technique for detection is found in radiology. The projections of the breast obtained with X-rays are called mammograms. Radiologists are trained to detect anomalies on mammograms and distinguish between malignant and benign ones. This is a hard task given the nature of breast cancer. Sometimes malignancies are overlooked (False Negative) or on further investigation anomalies turn out to be benign (False Positive). Both scenarios are very disturbing for the women involved.

To aid the radiologists in this decision process and to improve performance, computer systems are being developed. These Computed-Aided Detection (CAD) systems help find small irregularities on a mammogram and help to decide whether these are benign or cancerous. These systems still can use a lot of improvement, on detection and classification itself and also in interaction with the radiologist. We will focus mostly on the improvement of classification, the decision process whether a found anomaly is cancerous or harmless. This is a hard task, and whether an anomaly is cancerous or not is inherently uncertain (D'Orsi, Bassett, & Berg, 2003).

The domain thus demands a method which can handle uncertainty. Bayesian network technology is a good candidate for this. And it surpasses other machine learning algorithms on additional aspects. Most machine learning algorithms learn well from examples, but are not able to explicitly incorporate background knowledge, Bayesian networks allow both. By means of the structure of the network it can for example display causality. Also most common algorithms are black-box methods, like for example artificial neural networks. Despite its general good performance, it is not clear to the radiologist how such a system came to a certain conclusion. The radiologist can not receive feedback to enhance his own performance. Besides working as a method for classification, Bayesian networks may give the radiologist new insights on how certain aspects of a found irregularity contribute to the decision process.

Fortunately some work has been done before. In a collaboration of the radiology department of the Radboud University Nijmegen Medical Centre (UMCN) and the computer science department of the Radboud University Nijmegen a multi-stage CAD system has been developed. As a final stage a causal model of breast cancer was constructed which includes features like 'contrast' or 'size' of a suspicious region. Also in this stage it is chosen to consider multiple views of the same breast simultaneously. A visual representation of the model is shown in Figure 2.

A limitation of this existing model is that most of its features are continuous whereas a radiologist would annotate the regional features in a categorical fashion (with only a few categories, eg. low/medium/high contrast compared to a number on a continuous scale). Since radiologists must be able to work with the model, emphasis must be placed on the usability. Having a limited number of categories to annotate the features aids to this; thus, the solution is to discretize the data. Discretization is the process of converting features (either continuous or categorical) into features with a limited number of categories. (e.g. so any sort of data can be converted to a feature with a 3 category maximum). Apart from this intuitively directed reason to discretize the data, there are other reasons to believe that discretization can aid the model. When concerning classification performance it seems somehow counter-

intuitive to lose information in favor of usability. But it turns out that under some conditions performance could even improve.

Actual probability distributions are generally unknown, so people have sought for solutions to approximate the real (unknown) distribution. The better the approximation, the better the model. Though there are endless ways to represent a probability distribution, a common solution is to find the one Gaussian which best explains the data. This is a relatively simple method since only two parameters have to be learned, and still it is quite successful. One drawback of this method is that a Gaussian distribution is always symmetric, which is often inaccurate. Discretization does not make this assumption, even simple algorithms with enough bins can approximate the underlying distribution quite well.

1.2 Research Question

In this thesis we are going to investigate whether discretization techniques can improve the performance of Bayesian networks, specifically focussing on the previously constructed causal model of breast cancer. For this the model is further developed and classification performance on breast cancer data is analyzed.

1.3 Outline

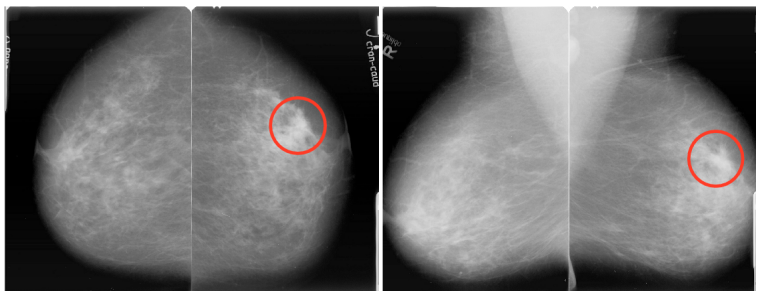
The setup of this thesis is as follows. First some background information is provided. Section 2.1 discusses the workings of mammography and the single view CAD system, its limitations, the promises of multi-view design and its problems. Then some related work concerning screening mammography and classification is discussed in section 2.2. A short introduction on Bayesian network technology is provided in section 2.3. A description of the model follows in section 2.4. Relevant aspects of discretization are discussed in section 2.5, and specific discretization algorithms which have been used in this experiment are explained in section 3.1. Also in the method section are a description of the data, more details of the experimental setup (section 3.2 and 3.3) and some more details on the analysis and validation of the results (section 3.4). In section 4 follow the results and at last section 5 provides a short discussion and future directions for research.

2 Background

2.1 Mammography and CAD

It is common screening practice to make multiple mammograms of one breast. This is important in case an abnormality is hidden behind other tissue in one view or just to gain more information of one region. Mostly one mammogram is taken from above (CC view, craniocaudal positioning), and another under an 45° angle (MLO view, mediolateral oblique positioning). See also Figure 1 for an example of mammograms with an annotated malignant region.

A suspicious area on a mammogram is called a *region of interest* (ROI), it could be a malignant *lesion* (or *abnormality*) or it can be a benign lesion or nothing special. Each region can be described by *features* such as contrast, size. Once a (malignant) lesion is observed in one view, we want to match it with the corresponding region in the other view. Since it is hard to identify the same region in two different projections, each of the ROIs in one



(a) (b)

Figure 1: CC (1a) and MLO (1b) view of both breasts of a patient. A cancerous lesion is seen within the circle.

view is coupled to each of the ROIs in the other view. A coupling between two regions is called a *link*. There are two types of links, one where none of the regions is malignant, and the other where at least one of the two regions is malignant. The latter is called a *finding*. Sometimes a malignancy is not visible in one of the projections, so with this definition we capture also those cases. When two coupled regions represent the same lesion it is called a *true* link, otherwise it is called a *false* link.

The causal model shown in Figure 2 is built on top of the UMCN single view CAD system, therefore it is useful to explain roughly the stages the CAD system goes through. For a more detailed description of the CAD system we refer to Engeland (2006). A short summary of the workings is as follows: The first step of the system is to segment the image into breast, background, and pectoral muscle (only in MLO view). After that some pixel-based locations of interest are detected and features are calculated and fed to a neural network classifier to calculate a pixel-based likelihood (of being malignant). Subsequently of these pixel-based locations the ones with a high likelihood (threshold used) are selected to extract regions. Adjacent pixels are merged to create regions of interest (ROIs), from which continuous features are calculated. The final stage is for a neural network classifier to decide whether the region is probably cancerous or not, based on these features. Based on this a region-based mass likelihood is computed to indicate the suspiciousness. The CAD system thus far provides regional features and a regional likelihood score separately for the regions in the MLO and CC view.

It appears that most lesions are detected by the single view CAD-system thus far. Not detecting regions at all does not seem to be the problem. However in the classification stage some regions (thus patients) are falsely classified as being benign (False Negatives). A new means has to be developed to reconsider the diagnosis of the regions. A possible solution is to combine information of two views of the breast. That allows to gain more information of a suspicious region. However, the coupling of regions is not trivial, the two mammograms are projections of a three-dimensional breast and to identify the same lesion on the two projections is not straightforward. To work around this each of the ROIs in the CC-view are connected to each of the ROIs in the MLO-view. In this linking process a region is also linked to other regions (which don't represent the same lesion/region). In the false link situation the second region does not contribute extra information (because the two region views are not of the same lesion). When a true link is established the second region does contribute

and may give more information. The latter case is where improvement is expected. A region might be (partly) obscured by healthy tissue. Therefore the system must be able to classify regions correctly also if no additional information is available. So single view classification is maintained.

In this set-up the working environment of the radiologist is also approached. A radiologist would not look only at a single region. He would compare it with the rest of the breast tissue and look for the same irregularity in another view. He might even compare it to the other breast or to older mammograms of the same breast. This all adds to his certainty of the diagnosis.

2.2 Related Work on CAD

In developing better CAD systems a lot of research branches exist, Oliver et al. (2009) provide a broad overview of research related to mass detection. Many focus on better region detection and segmentation of the mammogram. There are several others which also consider multiple views, some make a comparison of the left and right breasts while others use temporal information on the same breast. Few combine information of two different views (MLO and CC) as we do, but most exploit also a single-view CAD system for basic steps.

The research of Paquerault, Petrick, Chan, Sahiner, and Helvie (2002) is a nice example of a focus on matching the same area in the two views, using geometrical and nipple information to locate masses and textural and morphological features for further comparison. LDA is used for classification and the information is combined with the results from the single view classifier to come to a definitive conclusion.

There exists also some validation work whether CAD systems actually aid the radiologists, as a second reader in a double reading setting. Oliver et al. (2009) describe a few articles on this. If CAD systems succeed to detect otherwise overlooked malignancies (which they often do), the biggest problem now is that there are still too many false positives. Once the performance improves some say that the radiologists become lazy or rely too much on the system.

Little research has been done which exploits Bayesian networks. Kahn, Roberts, Shaffer, and Haddawy (1997) made a Bayesian network which besides some mammographic features concerning both masses and calcifications, includes features about patient history and physical findings such as pain. More recently Burnside et al. (2009) used structure learning to make a Bayesian network, also including both calcification information (mammographic sign of malignancy) and a few parameters concerning masses.

2.3 Bayesian Network Theory

In this section Bayesian network theory is introduced, for this Jensen and Nielsen (2007) is followed, which can also be a source for a more detailed explanation. We have a set of random variables $\mathbf{X}$, every variable $X_i \in \mathbf{X}$ can have values from a limited domain or in the discrete case they can be in a limited number of states. A Bayesian network can be defined as $BN = (G, P(\mathbf{X}))$ where G is an directed acyclic graph (DAG) and $P(\mathbf{X})$ is the joint probability distribution of $\mathbf{X}$. The DAG is defined $G = (N, E)$ where the set of nodes N correspond with the set of random variables $\mathbf{X}$ and the set of arcs between nodes $E \subseteq (N \times N)$ represent dependencies between variables. A dependency between two nodes can also be interpreted as a causality. A cause (parent-node) can lead to an effect (child-node).

Especially graphically this is easy to comprehend, e.g. a presence of a 'abnormal density' will lead to an higher 'contrast' value (Figure 2).

If $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are disjoint subsets of $\mathbf{X}$, then $\mathbf{U}$ and $\mathbf{V}$ are independent if $P(\mathbf{U} \mid \mathbf{V}) = P(\mathbf{U})$. $\mathbf{U}$ and $\mathbf{V}$ are conditionally independent given $\mathbf{W}$ if $P(\mathbf{U} \mid \mathbf{V}, \mathbf{W}) = P(\mathbf{U} \mid \mathbf{W})$. So $\mathbf{U}$ and $\mathbf{V}$ are independent only in the context of $\mathbf{W}$. We denote this conditional independence as $\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$. This independence is also reflected in the structure of the network. The Markov blanket of a node in a Bayesian network consists of the children of the node, its parents and the children of its parents. For every node it holds that given its Markov blanket it is conditionally independent of every other node in the graph.

The dependencies between two related nodes are specified by a conditional probability distribution. In this case all the variables are discrete, so conditional probability tables (CPTs) are used. Each of the nodes have an associated conditional probability distribution $P(X_i \mid parents(X_i))$. The joint probability distribution can be calculated with the chain rule for Bayesian networks

$$P(\mathbf{X}) = \prod_{X_i \in \mathbf{X}} P(X_i \mid parents(X_i))$$

Since a classification task is performed a special variable is included in the model. The class variable is a variable which denotes the class of an instance.

2.4 Causal Model

In this section the previously constructed model is described. All the variables are listed and discussed below. Some examples of causality and independence are given. The unique handling of these issues is after all one the main characteristics of this method. In figure 2 the visualization of our model is shown. All the white squares represent observable variables. Ovals represent hidden variables. The node 'Finding' represents the class node. Now follows a short description of the variables (Ferreira, Velikova, & Lucas, 2008).

Finding The class node, the associated probability distribution reflects the probability of a cancerous finding as verified by an expert. Recall that finding is 'true' if the region in at least one of the two views is malignant.

Pixel-based Malign Likelihood Reflects the likelihood of being malignant as calculated on pixel level in the single view CAD system.

False-positive Level Regional feature which reflects the number of false positives when the ROI is classified as cancerous. Hence the lower the score on this variable, the more certainty that the ROI is in fact cancerous.

Distance to Skin The shortest distance to the skin.

Location X and Location Y Some areas of the breast are more likely to contain cancerous regions than others (e.g. more in the upper outer quadrant, less in retromammillary area.) This measure is normalized for the breast size.

Abnormal Density An abnormal density can either be a focal mass or an abnormal structure. Presence of an abnormal density will probably also cause a higher contrast value.

Abnormal Structure Either a spiculated structure or a distortion in linear texture.

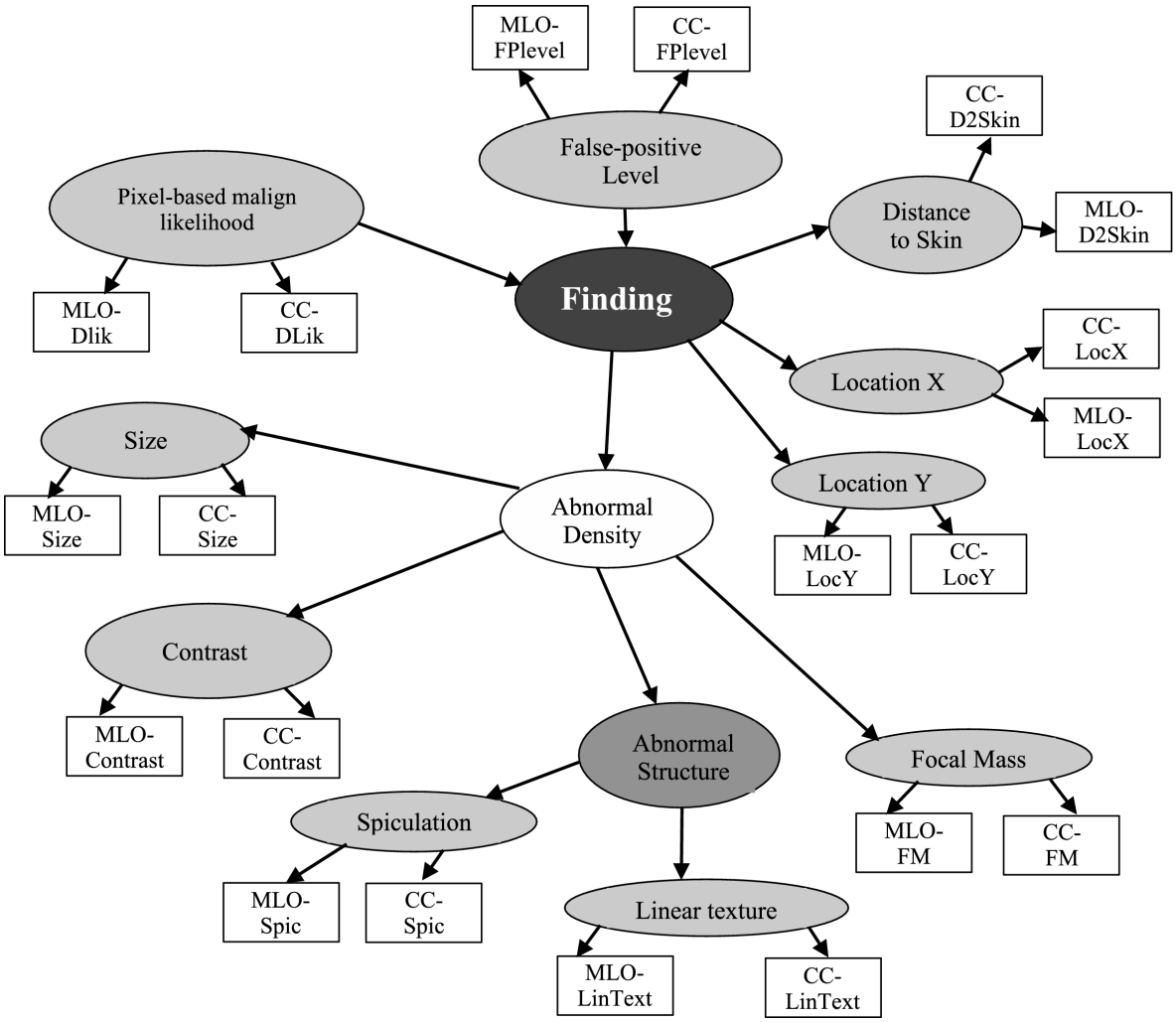


Figure 2: Causal Model

Size This reflects the size of a ROI.

Contrast High contrast on the mammogram is often associated with a cancerous area.

Spiculation Spicules are spikes of straight lines towards the center of a lesion. This is a measure for how stellate (star-like) a region is. The higher the degree of spiculation, the higher the likelihood for malignancy.

Linear Texture This measure reflects how normal the breast tissue is. The higher the linearity, the lower the likelihood of being cancerous.

Focal Mass Neat circumscribed lesions have higher values on this one.

2.5 Discretization

Discretization is the conversion of any sort of data into nominal (categorical) variables. While the algorithms we use keep the data ordinal, Bayesian network technology does not exploit

that and handles the categories as independent. The actual real-world probability density of the data is often unknown, therefore estimates are used. Often a Gaussian distribution is chosen, or some other common distribution, but this can be an incorrect representation of the real (unknown) distribution. Some argue that when one performs discretization, you can better estimate the probabilities (based on counting/tallying). Thus it is better to use the discretized data than assuming a flawed density, although there is some information loss (Yang & Webb, 2003).

Another reason to discretize the data is to make the CAD-system more usable for the radiologist. Numbers are harder to interpret than categories. For example a radiologist will typically annotate a region with 'high contrast' or 'very low contrast', in stead of '19.48-like' or '0.42-ish'.

In searching for an optimal discretization from a classification point of view we will utilize some common discretization techniques. These will be explained below in the method section 3.1. Apart from that we will test if enhancing performance is possible while maintaining usability for the radiologist. Discretization with 10 or more bins will perhaps do its works on enhancing performance but probably not appeal to the radiologist. Therefore we control the number of categories made by some of the algorithms.

3 Methods

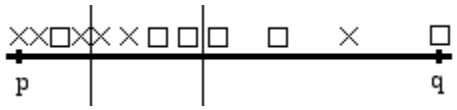
3.1 Discretization Algorithms

In this section a short description of the chosen algorithms for discretization is given. Some promising methods are selected, both supervised and unsupervised. Also some simple techniques where the number of bins can be easily limited (Dougherty, Kohavi, & Sahami, 1995). All methods categorize the data based on the same set of training data and the characteristics are saved for new data. All methods handle the discretization of one variable at the time, thus not taking dependencies between variables into account. For the implementation/execution of the techniques Weka (Witten & Frank, 2005) is used. For easy comparison and clarification of the algorithms a visualization is given in Figure 3.

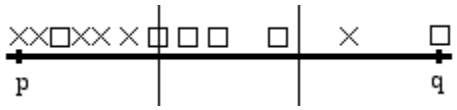
Equal Frequency Binning (EFB) This algorithm determines the bin boundaries by first sorting the data on ascending values and subsequently divide the data in equally sized bins. This algorithm is executed twice. Once with 10 bins, for high performance. And once with 3 bins, for checking performance while maintaining usability. It does not use class information and thus is an unsupervised algorithm. A visualization of this method with 3 bins is shown in figure 3a.

Equal Width Binning (EWB) This unsupervised algorithm works similar to EFB, but the sorted data is now divided in equally ranged bins. Also this algorithm is performed both with 3 or 10 bins. An additional run is done with an optimization to determine the number of bins. A visualization of this method with 3 bins is shown in figure 3b.

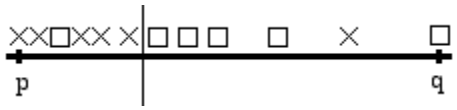
Method of Fayyad and Irani This method selects a bin boundary based on the minimization of the class information entropy. This can be done multiple times to create multiple bins (Fayyad & Irani, 1992). It is the only supervised method in this list.



(a) EFB with 3 bins



(b) EWB with 3 bins



(c) FI one split

Figure 3: Visualization of different discretization algorithms on same (artificial) attribute. The data is sorted from lowest (p) to highest (q) value first. The class only plays a role in the algorithm of Fayyad and Irani (3c). See text for more explanation.

In detail: The class entropy of a (sub)set S is defined as

$$Ent(S) = -\sum_{i=1}^{k} P(C_i, S) log_2(P(C_i, S))$$

$P(C_i, S)$ represents the proportion of instances in S with class C_i, k is the number of classes, in our case 2, cancerous or non-cancerous. 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:

$$E(A, T; S) = (|S_1|/|S|)Ent(S_1) + (|S_2|/|S|)Ent(S_2)$$

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 (MDL) criterion is used as stop criterion to avoid ending up with too many bins. In Figure 3c can be seen how the algorithm succeeds in splitting the attribute into two bins, effectively having the cut point on a class boundary and creating homogeneous bins. For a more detailed explanation and justification see Fayyad and Irani (1993).

PKI-discretize This is EFB where the number of bins is equal to the square root of the number of instances. For the justification of this algorithm see Yang and Webb (2001).

None To interpret the performance with the different discretization techniques we use a set-up with continuous data as a baseline for comparison.

3.2 Data Description

Data is obtained from the Dutch breast screening practice. After it is processed by the single view CAD system, features are selected. Feature selection is necessary because multiple

algorithms for feature extraction are present, e.g. there are multiple contrast features yet only one is needed for this model. The features with the most distinctive power have been chosen. The dataset include entries of 795 patients, of which 344 are cancerous. A problem is that a ROI in one view can not always be coupled to the corresponding area in the other view. Sometimes simply because such an area is not visible at all. To overcome this all links of which at least one region is cancerous, are labeled as cancerous. From both the MLO-view regions as the CC-regions the three most suspicious regions are used and coupled to each other. This results in a database where for each breast multiple instances are added, where each instance reflects a link between a CC and a MLO ROI. The database consists of 14129 links.

3.3 Experimental Setup

In our model the class variable is the binary variable 'Finding', which represents a finding of a cancerous lesion in at least one of the two views. To use the model for diagnostics we have to set the probability distributions in the model and then we can infer the posterior probability of a region being cancerous given some evidence. The real probability distributions are unknown. But the model can use the limited available data to estimate the underlying distribution. For this learning process we use the EM algorithm (Expectation Maximization). It is a frequently used approximation algorithm to learn the CPTs when data is incomplete. Some of the nodes in the model are unobservable, the hidden nodes, hence the incompleteness. When the model has learned the CPTs, it is ready for classification. The observed values of the regional features (evidence) are set on the corresponding nodes and passed through the model. The posterior probability of being cancerous as reflected in the 'Finding' node can easily be calculated. Classification of a region can be accomplished by choosing a cut-off point on the posterior probability distribution of being cancerous. e.g. refer all the patients with a ROI with a posterior probability of being cancerous of more than 60% to a specialist. More on how we tackled classification and validated the model is discussed in the method section (3.4). The Bayesian Network Toolbox in Matlab (Murphy, 2007) is used to build, train and test the model.

As can be expected, all other aspects of the procedure besides different discretization algorithms are kept alike. The EM algorithm is used for learning the probabilities in the Bayesian network. All hidden nodes are binary. Some priors have been set on the hidden nodes, based on expert knowledge, this is based on previous work on the same model. We used 2-fold cross validation to make more use of the data, so each instance is used once for training the model and once for testing. The outcome on the testdata of both folds are put together afterwards.

3.4 Analysis

Instead of choosing a cut-off point for classification, in each experiment a ROC graph is constructed (Figure 5). This has some advantages. If a certain class has a high occurrence in a dataset (or real world), classification algorithms can have a good score just by always selecting that certain class. So there must be an higher punishment for errors on the minority class. Usually there is a trade-off between classifying as much correct of one class (True Positives, Cancerous), while minimizing the number of erroneously classified instances of the other class (False Positives). In a ROC graph it is visible how much errors are made when a

certain cut-point for classification is chosen.

The cut-off point eventually chosen is primarily a political choice. If a society want to take no risk in missing cancerous spots it is possible to direct all patients for further investigation (e.g. biopsy, other scans). This will cost a fortune and also for other reasons it might not be optimal to examine healthy people (e.g. stress or health risks of the methods). On the other hand if no people at all are directed, no unnecessary mistakes are made but also no cancer could be treated, this makes the existence of the CAD system pointless. The optimal cut-off point lies somewhere in between, depending on resources and priorities of the society.

If two ROC-curves are compared the improvement can be seen for a specific cut-off point or area. If a curve is above another it means for the same false positive rate the classifier has a higher hit rate, thus classifies more of the cancerous cases correctly while not directing more health people (Fawcett, 2006). A similar comparison can be made if a curve lies left of another curve. This means that for the same hit rate there are less false positives.

The area under the ROC curve (AUC) is regarded as a general measure of classification performance. It misses certain nuances but generally the higher the AUC the better the classifier. More precisely, it is a measure for how well the instances are ranked (Bradley, 1997).

The performance can be regarded on different levels. Each patient has two breasts, in both several suspicious regions are selected, which are coupled, resulting in several *links*. Since the model in both learning and classifying processes only one link at the time it is obvious to analyze the performance on link-level. But in practice when a patient is referred for further investigation she is examined more extensively, resulting in a better image of the breast. So if two systems seem to have the same performance, the system which has discovered cancer correctly in the most patients is better. Once a patient is referred by the radiologist to the oncologist for further examination, it does not matter much if the system classified other regions in the breast correctly as cancerous. In the screening situation it is worse to miss a patient than to miss another spot on the same patient. To analyse the results on patient-level, for each patient only the maximum probability of being cancerous is taken into account. With this data again a ROC analysis is performed.

4 Results

In this section we present the results. In Table 1 the name of the discretization technique is provided in the first column. The second column is the AUC (of the two folds) for every technique. In the third column the performance is the AUC of a ROC curve based not on the correct classification of each link (coupled region in CC view and in MLO view), but on the correct classification of a patient.

First we take a look at the results on link level, where each of the links is considered separately. The precise numbers can be found in Table 1, the best performing methods are visualized in Figure 5. With the majority of discretization techniques the results improve. The method of Fayyad and Irani performs best, Equal Frequency Binning follows, the alternative with 10 bins better than with 3 bins. Equal Width Binning with 10 bins also slightly improves, but the other two variants deteriorate. The performance with PKI Discretize deteriorates the most. Though this can be considered a measure of performance for the different instantiations of the model, more interesting is the performance on patient level.

On patient level the results are similar. The method of Fayyad and Irani performs best,

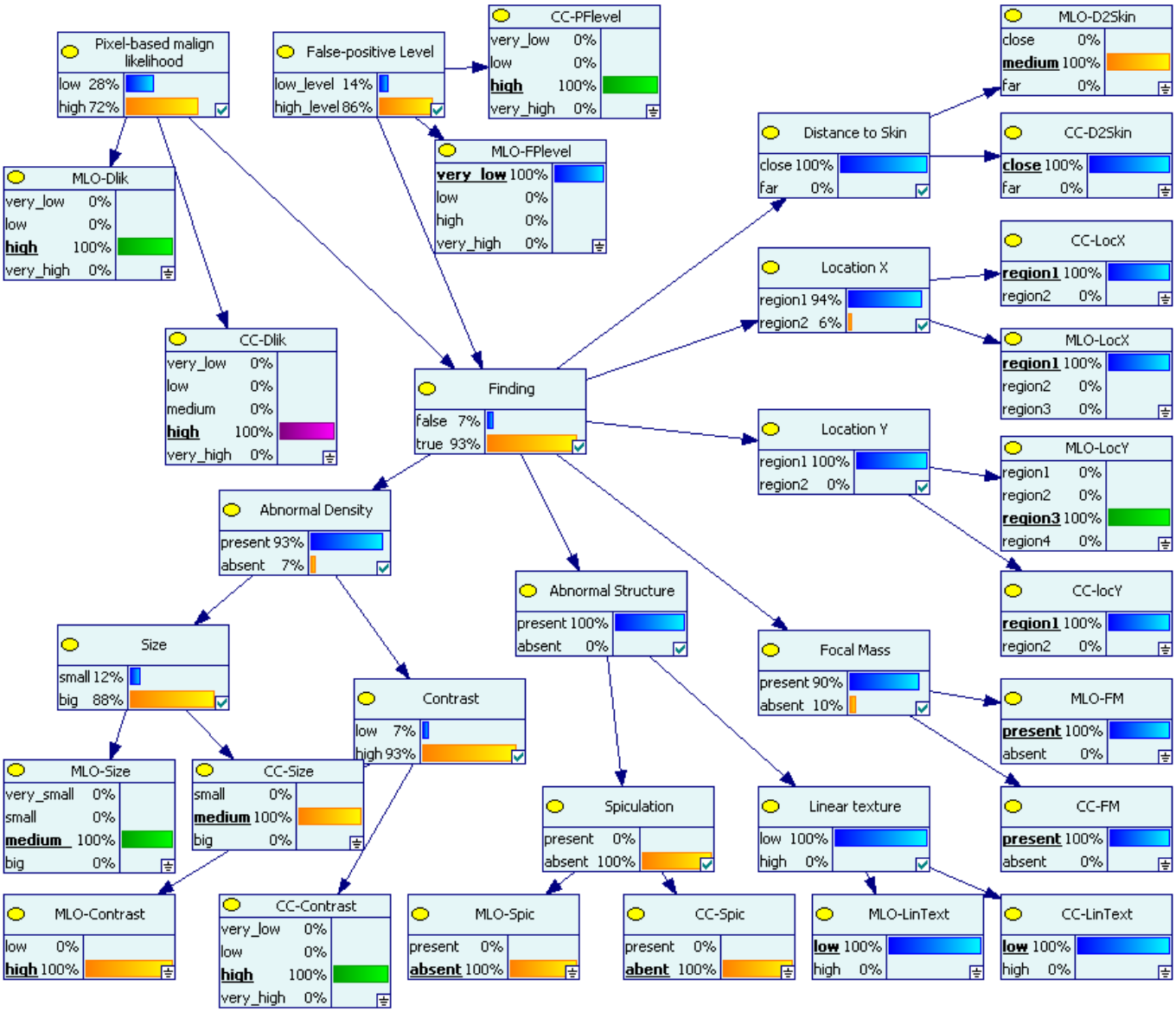


Figure 4: Causal model with evidence set (represented by bold and underlined names of the states) for a positive example and calculated posterior probabilities. This model is discretized with the method of Fayyad and Irani

Discretization method	AUC Link-based	AUC Patient-based
Continuous variables (baseline)	0.7065	0.6276
Fayyad and Irani (supervised)	0.7898	0.7548
Equal Frequency Binning, 10 bins	0.7539	0.7331
Equal Frequency Binning, 3 bins	0.7090	0.6113
Equal Width Binning, 10 bins	0.7196	0.6543
Equal Width Binning, Optimized	0.7041	0.6322
Equal Width Binning, 3 bins	0.6775	0.5721
PKI Discretize	0.5922	0.5329

Table 1: AUC Results on model, different discretization methods compared to (continuous) baseline. Data of both folds is combined.

after that Equal Frequency Binning with 10 bins and Equal Width Binning with 10 bins. Equal Width Binning with an optimized number of bins now slightly improves compared to the continuous baseline, instead of deteriorating when considered on a link level. Equal Frequency Binning with 3 bins shows an opposite pattern. Equal Width Binning with 3 bins and PKI Discretize also show no improvement. The best performing methods are visualized in Figure 6.

To illustrate the behavior of the network with a positive example the causal model is showed once more in Figure 4. The evidence has been set on the observable nodes, whereafter beliefs have been updated. Now the posterior probability of the example being cancerous can easily be seen.

5 Conclusion and Discussion

We wanted to know if discretization techniques can improve the performance of Bayesian networks. For this we compared classification results on breast cancer data in a Bayesian network which functions as a causal model of breast cancer. The results of the experiments with the causal model make clear that discretization techniques can indeed improve the performance of Bayesian networks. The algorithm that gives the best results is the algorithm of Fayyad and Irani. This makes sense since bin boundaries are more likely on class boundaries, therefore there is less noise in each of the bins. For example the equal width binning algorithm with 10 bins sometimes causes very sparse bins, even empty ones, and on the contrary can split a group of data with the same characteristics into multiple bins. These problems hold for all of the unsupervised methods. The method of Fayyad and Irani avoids the creation of ambiguous bins, always create a majority as big as possible within a bin.

PKI discretization performs worst of all, yielding a considerable drop in performance compared to the continuous baseline. Perhaps this lies in the huge amount of bins created, over 80 in both folds. Perhaps the problem lies in that the number of bins created by PKI depends on the total amount of instances, that might work when class occurences are similar, but in screening practice (and thus in our dataset) healty cases are overrepresented.

The method of Fayyad and Irani did not always yield good results. For example, there were a few variables where the method produced one bin only, indicating that these variables were irrelevant. Perhaps these variables should have been discarded in the variable selection process. However, it may also be the case that the dataset we used was too small, or that

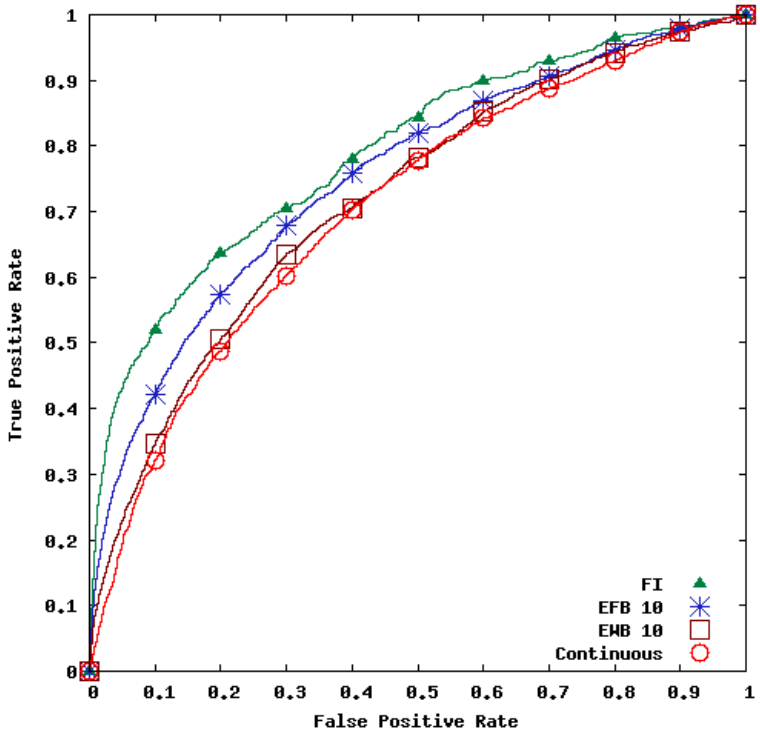


Figure 5: ROC curve of all methods on linklevel.

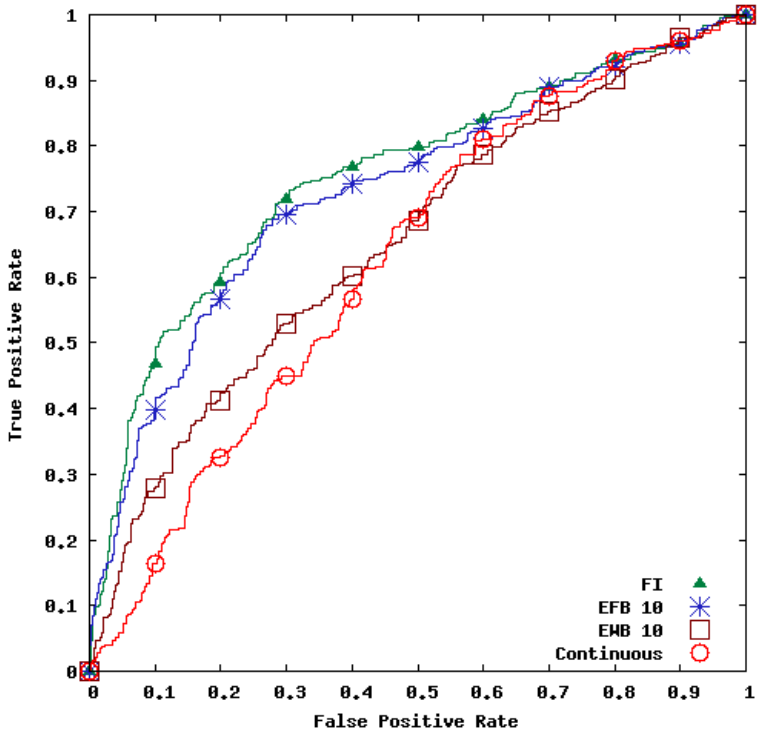


Figure 6: ROC curve of all methods on patient-based.

the chosen stopping criterion caused the algorithm to terminate too early. Finally, given that many of the variable are statistically dependent on each other, which is why the Bayesian-network representation makes sense, treating them as being independent in the discetization process also will give rise to errors.

This last issue is also a point of criticism in general, variables being only regarded independent of one another. Perhaps improvement can be made when variables are discretized simultaneously, thus also capturing dependencies. Bin boundaries may not lay on class boundaries of only one variable, but these might shift when others are taken into account. Not much is done here, which is not surprising, because it is a computationally expensive task, having to search through a large space of possible splits. Also it might be a problem that datasets are not big enough to find these things. This might change when in the future (it is happening already for example in the Netherlands) more and more mammograms are taken digitally, and thus a huge quantity of data might easily become available.

A possible direction for solutions might lie in better integration of the discretization process with the available knowledge. For example previous research has led to the knowledge that Contrast and Distance to Skin are two independent variables, when discretizing the Contrast variable it is not necessary to regard the Distance to Skin variable. Perhaps in Bayesian Network technology if a variable is discretized it is sufficient to only regard its Markov blanket. This process can be iterated for each variable, discretize it with this locally available extra information. In this way more information about the distribution is available, but the search space is limited compared to partitioning the variables all at once.

When more digital mammograms come available it might also direct to dynamically improving any method used for classification. Now certain dependencies might be not present in the data, some might even be consequences of noise (overfitting). The more data comes in, the better a CAD system can be, ideally would be to dynamically improving the CAD system. This is a whole new branch of research, not only use new data to fine-tune the existing models, but also elicit new knowledge dynamically.

Concerning usability, the results are also promising as even a very simple technique as equal frequency binning with 3 bins shows slight improvement. This guides to compatibility with the radiologist view. This research can be extended by optimizing the adjustment of the algorithms and its parameters. However, despite assuming the same amount of bins, it is not guaranteed that radiologists and discretization techniques choose similar bin boundaries. When eventually a superior model will be taken into practice, additional testing on usability is necessary. This might lead to adjustment of the CAD system and also to specific training for radiologists.

If the main reason that discretization works is that it is a nice estimate of the real probability distribution, discretization might become redundant when other ways to estimate that distribution emerge or improve (John & Langley, 1995). However, this is again a whole new branch of research while discretization can be relatively easy, and apparent advantages of radiologist compatibility disappear.

Despite all of the above comments, this thesis definitively did show that discretization can play an important role in Bayesian modelling.

References

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159.

Burnside, E., Davis, J., Chhatwal, J., Alagoz, O., Lindstrom, M., Geller, B., et al. (2009). Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*, *251*(3), 663-672.

D'Orsi, C., Bassett, L., & Berg, W. e. a. (2003). *Breast Imaging Reporting and Data System: ACR BI-RADS-Mammography (ed 4).* Reston, VA, American College of Radiology.

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the twelfth international conference on machine learning* (Vol. 202, p. 194-202). Morgan Kaufmann, Publishers, San Francisco, CA.

Engeland, S. (2006). *Detection of mass lesions in mammograms by using multiple views.* Unpublished doctoral dissertation, Radboud Universiteit Nijmegen.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fayyad, U., & Irani, K. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, *8*(1), 87–102.

Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the thirteenth international joint conference on artificial intelligence* (p. 1022-1027). San Francisco, CA: Morgan Kaufmann.

Ferreira, N., Velikova, M., & Lucas, P. (2008). Bayesian modelling of multi-view mammography. In *Proceedings of the icml workshop on machine learning for health-care applications.*

Jensen, F., & Nielsen, T. (2007). *Bayesian networks and decision graphs.* Springer Verlag.

John, G., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (Vol. 1, pp. 338–345).

Kahn, C., Roberts, L., Shaffer, K., & Haddawy, P. (1997). Construction of a Bayesian network for mammographic diagnosis of breast cancer. *Computers in biology and medicine*, *27*(1), 19–29.

Murphy, K. (2007). *Bayesian network toolbox (BNT).* (`http://people.cs.ubc.ca/~murphyk/Software/BNT/bnt.html`)

Oliver, A., Freixenet, J., Martí, J., Pont, J., Pérez, E., Denton, E., et al. (2009). A Review of Automatic Mass Detection and Segmentation in Mammographic Images. *Medical Image Analysis*.

Paquerault, S., Petrick, N., Chan, H., Sahiner, B., & Helvie, M. (2002). Improvement of computerized mass detection on mammograms: Fusion of two-view information. *Medical Physics*, *29*, 238.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems).* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yang, Y., & Webb, G. (2001). Proportional k-interval discretization for naive-Bayes classifiers. In *Machine learning: Ecml 2001* (pp. 564–575). Springer.

Yang, Y., & Webb, G. (2003). On why discretization works for naive-bayes classifiers. In *Ai 2003: Advances in artificial intelligence* (pp. 440–452). Springer.