Bayesian Networks Seminar Paper: Applying Bayesian Networks to diagnosis and image recognition

Allard Hendriksen
s1078666
hendriksen.a.a@gmail.com

Simon Szatmari
s1416685
artabalt@gmail.com

Two papers are presented. The first treats the problem of image classification. In particular the authors succeed in fusing two tools of the machine learning community: the framework of ontologies and the framework of bayesian networks. An ontology is a structure that represents human knowledge in terms of concepts and relationships between the concepts. Bayesian networks are successful at handling uncertainty. The two perspectives come together to successfully classify images. The second paper treats the problem of deciding whether a person is affected by the Alzheimer's disease. The authors combine the idea of using Newton interpolation on missing data which is then used in the K2 algorithm. Remember that an ordering of the factors is necessary for the K2 algorithm, this is done via a mutual information criterion of the factors with MCI, which is a Mild Cognitive Impairment that has been associated to Alzheimer's. In particular the authors do not use an expert system. Noting that an ontology is similar to an expert system, these two articles are prototypical examples of one team that uses an expert system and of one that does not. In section 1, we present the image classification paper and in section 2 we present the MCI paper.

1 Evidence-Driven Image Interpretation by Combining Implicit and Explicit Knowledge in a Bayesian Network [1]

This paper joins the structure of ontologies with the structure of Bayesian networks into a coherent whole with applications in computer vision, object recognition and concept mining. More specifically they use this combination for

- Image categorization,
- Localized Region Labeling,
- Weak annotation of video shot keyframes.

We will explain in detail the steps and results for image categorization. The other two are similar.

1.1 Method

In very broad strokes, the framework works as follows: with the goal of categorizing an image,

- we have a knowledge base Kb representing an ontology (relations between concepts),
- we have a training set that is annotated with the concepts present in Kb,
- we map that knowledge base Kb to a causal graph whose structure represents the structure of Kb,
- we get the CPTs of the BN from the training set using the EM algorithm,
- we split the concept set into two: a set of *hypothesis concepts* and a set of *evidence concepts*,
- we use a greedy search method by alternating on the set of hypotheses and the set of evidences in order to find the concept that matches best a given image.

1.2 Ontologies and Bayesian Networks

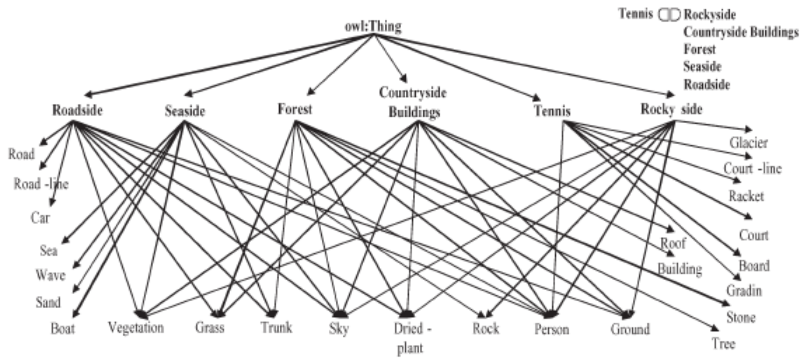
We present a minimal working example of the aforementioned framework, borrowed from the paper. One can find a formal description of ontologies in [2], a detailed explanation of the migration of ontology to bayesian networks in [3].

An ontology is a system of concept objects, members of a set C, linked together via some relations. These relations are mainly the usual set theoretic ones, such as "is instance of", "is subset of". Effectively an ontology is an expert system, built either by hand or via concept mining algorithms that we will not cover here.

To take a knowledge base Kb into a BN, we must map ontological elements, such as concepts and relations, to graph elements such as nodes and arcs. The general principle underlying the translation rules is that all classes are translated into nodes in BN, arcs are drawn between two nodes in BN if the corresponding two classes are related by a "predicate" in the knowledge base. The direction of the arrows is from the superclass to the subclass if it can be determined. Control nodes are created during the translation to facilitate modeling.

For example, consider the ontology whose concepts are made up of $C = \{$ Countryside buildings, seaside, rockyside, forest, tennis, roadside, building, roof, tree, stone, grass, ground, dried-plant, trunk, vegetation, rock, sky, person, boat, sand, sea,wave,road, road-line, car, court, court-line, board, grading, racket$\}$. The training set that was be used here was created by the authors to use in a competition. This set of concepts is mapped using the structural relations of the knowledge base into the corresponding causal graph.

Figure 1: Ontology that encodes domain knowledge, graphical representation

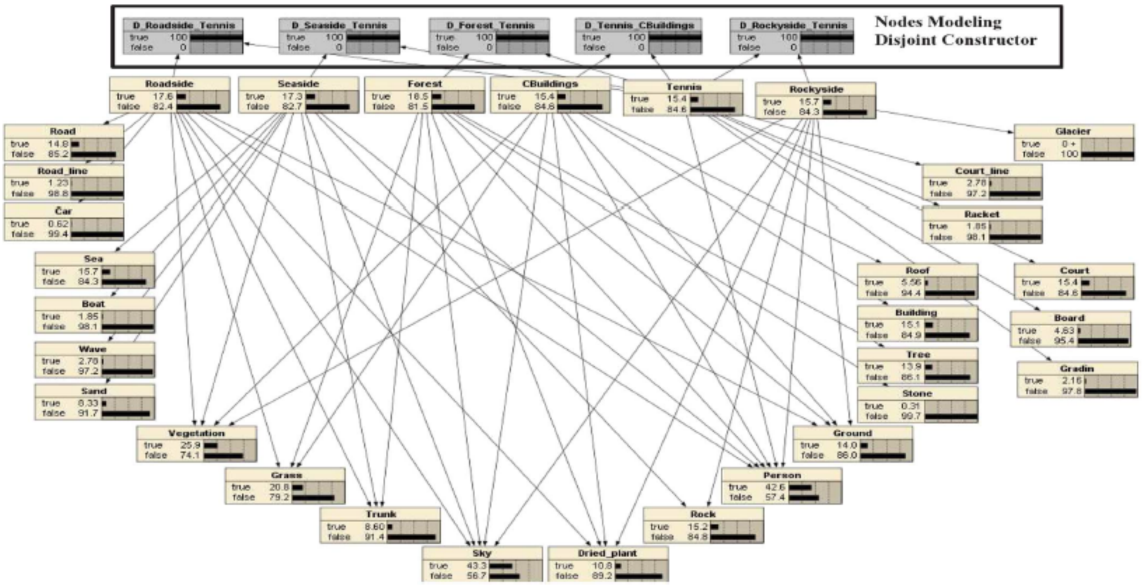


Note that the structure of the ontology is split into two: a set of concepts on the top and a set of concepts at the bottom. The set of concepts at the top form the set of *categorical (global) concepts* $C_G = \{$Countryside-buildings, seaside, rockyside, forest,tennis, roadside$\}$ and the set of concepts at the bottom form the set of *regional (local) concepts* $C_L = \{$Building, roof, tree, stone, grass, ground, dried-plant,trunk, vegetation, rock, sky, person, boat, sand, sea,wave,road, road-line, car, court, court-line, board, gradin,racket$\}$.

1.3 Greedy Search and Focus-of-Attention

Finally in order to categorize an image, a greedy search is used on the space of concepts. The reason why we do not try to fit all concepts to a given image is because that would suffer from an exponential explosion of checks. The advantage of using a knowledge base Kb is not only to ease the construction of the bayesian network, but also to greatly diminish the number of checks necessary before finding a concept that matches the image. The greedy search uses what the authors call a focus of attention method (FoA). It makes up a big portion of the paper and so we proceed to expose all the details. Recall that the concepts are of two types: hypothesis concepts and evidence concepts. The search steps as follows: pick the h-concept with the highest prior, if the prior is high enough then stop, otherwise update the hypothesis concept with the evidence concept having the highest prior, using the BN. If the posterior of the hypothesis concept is high enough then stop, otherwise go on to the next hypothesis concept having the highest prior. Do this until you meet a hypothesis concept with high enough probability of being in the image, otherwise if the search has ended and no suitable hypothesis concept has been found, then pick the hypothesis concept with the highest probability.

Figure 2: Bayesian network derived from ontology given above

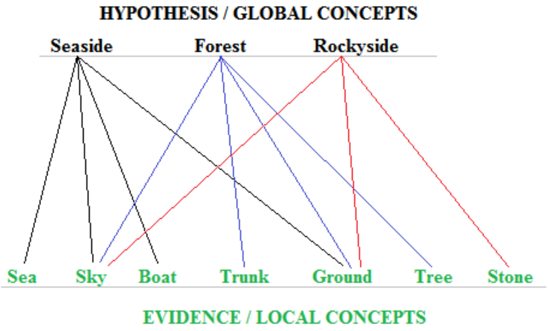


Formally:

- Define $h(I_q, c_i) = P(c_i|I_q)$ the degree of confidence that concept c_i is in image I_q, this is done via a cheap classifier.
- Set the hypotheses to be the categorical concepts: $C^H = C_G$; setting the hypotheses to be the categorical concepts is for image categorization. For regional image labeling you would set the hypotheses to be the regional concepts.
- Define $H(I_q) = \{h(I_q, c_i) : c_i \in C^H\}$ the set of confidences for the hypothesis concept set C^H of being in image I_q.
- First order the hypothesis set $H(I_q, C^H)$, this is not a set concepts ordered by a decreasing sequence of prior probabilities.
- Greedy Search:
  1. Pick the concept c_i with the highest degree of confidence $P(c_i|H(I_q), E(I_q), Kb, W_{ij})$. The W_{ij} is a correlation matrix of concepts that goes into the BN.
  2. If $P(c_i|H(I_q), E(I_q), Kb, W_{ij})$ is high enough, done.
  3. Else, order the evidence set $H(I_q, C^{E_i})$, where C^{E_i} is the set of regional local concepts under c_i.
  4. Pick an arbitrary number of evidence concepts that ranks highest, and plug into the BN,
  5. If now the updated value, posterior probability, of the node of c_i is big enough, then return that concept as the categorization.
  6. Else repeat steps 1-5 with the next highest ranked hypothesis concept.
  7. If by the end none of the hypothesis concepts have a high enough degree of confidence, then pick the one with the highest confidence degree.

Confusion might arise from step 3. Suppose you want to test the next likely hypothesis concept c_i in C^H, but that the degree of confidence of c_i is not high enough for an image. As an example, look at the following sub-ontology: Now if the hypotheses are ordered as $H(I_q|C^H) = \{Sea, Rockyside, Forest\}$ (decreasing), but the concept *sea* does not have a high enough degree of confidence, then we will look at the concepts under sea, namely $\{sea, sky, boat, ground\}$ and order that set as well via $H(I_q|C^{E_i}) = \{boat, sea, sky, ground\}$ (note that *sea* might look a bit like *sky*, so *boat* is more revealing). The next step will be to first plug in the hypothesis concept in the BN and then to update the BN with the evidence concept having the highest prior. If now the posterior probability of the node *sea* is high enough we stop, otherwise we go to the hypothesis concept c_{i+1} *rockside* with the next highest prior probability, that has evidence concepts $\{sky, gruond, stone\}$ under it, in the order $H(I_q|C^{E_{i+1}}) =$

Figure 3: a sub-ontology

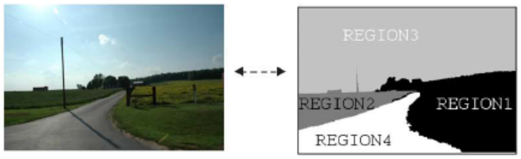


$\{stone, ground, sky\}$. If by the end of this process we find nothing with a high enough degree of confidence, we pick the combination that has the highest degree of confidence.

1.4 Images and Regions

This is the general outline, however one can refine this idea. Suppose you have the following image: which is split into 4 regions by the help of some exterior salient point detector. First we would see

Figure 4: An image and its regions



whether the hypothesis $c_i = sea$ is good enough, whereupon finding that it is not, we then proceed to look at evidence $c^{E_i} = \{sea, boat, sky, ground\}$. We can check this set of evidence against each four regions separately. For each region, pick the evidence concept with the highest classification score, plug that value into the BN to the the posterior of the hypothesis concept. In this fashion, we would get a set of four posterior probabilities for *sea*, one per region. We can play around with these values, depending of the classification of interest: we can stop the search if the maximum of these four values is high enough, which would correspond to the statement "there is a *sea* concept in the image". We could also only stop the search when the minimum of these four values if high enough, which would correspond to the statement "every region in this image corresponding to the concept *sea*".

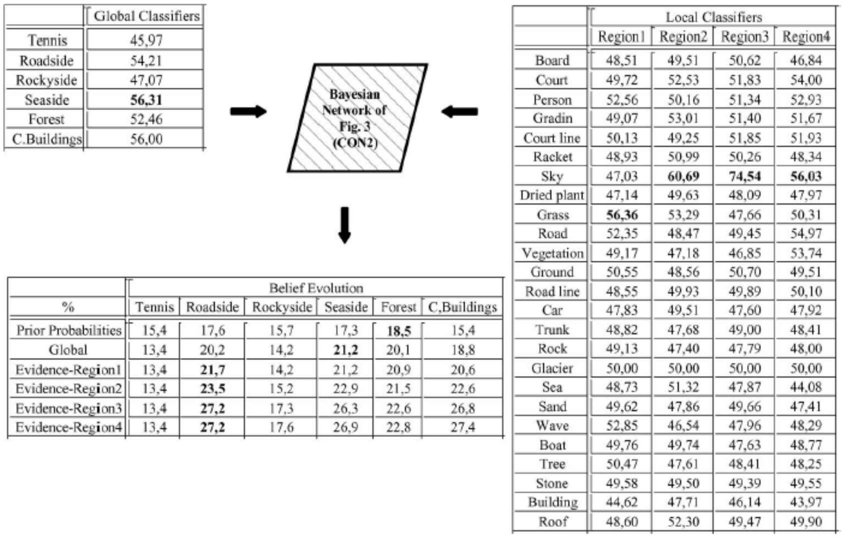
1.5 Results

The authors present the results for a collection of 648 images I^{PS} comprised the data set for the PS domain, which is the Kb introduced earlier. All images in I^{PS} are annotated at the global and region details using the set of category concepts C_G and the set of regional concepts C_L.

I^{PS} was split in half to formulate the test I^{PS} test and training I^{PS} train sets, each set containing 324 images. I^{PS} train was used for training the classifiers F_c and learning the parameters of the BN.

They examine the efficiency of categorizing the images of I^{PS} test to one of the categories in C_G using three configurations. These configurations vary in the amount of utilized context and knowledge. In the baseline configuration $CON1$, they assess the performance of image categorization based solely on visual stimulus. Images are categorized based on the maximum value of the global concept classifiers. The second configuration $CON2$ uses context and knowledge base to extract the existing evidence

Figure 5: Posteriors corresponding to the ontology, for image categorization


Bayesian
Network of
Fig. 3
(CON2)

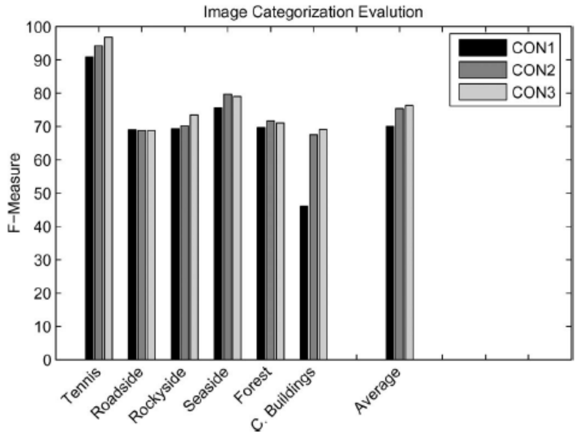
	Global Classifiers
Tennis	45,97
Roadside	54,21
Rockyside	47,07
Seaside	**56,31**
Forest	52,46
C.Buildings	56,00

	Belief Evolution					
%	Tennis	Roadside	Rockyside	Seaside	Forest	C,Buildings
Prior Probabilities	15,4	17,6	15,7	17,3	**18,5**	15,4
Global	13,4	20,2	14,2	**21,2**	20,1	18,8
Evidence-Region1	13,4	**21,7**	14,2	21,2	20,9	20,6
Evidence-Region2	13,4	**23,5**	15,2	22,9	21,5	22,6
Evidence-Region3	13,4	**27,2**	17,3	26,3	22,6	26,8
Evidence-Region4	13,4	**27,2**	17,6	26,9	22,8	27,4

	Local Classifiers			
	Region1	Region2	Region3	Region4
Board	48,51	49,51	50,62	46,84
Court	49,72	52,53	51,83	54,00
Person	52,56	50,16	51,34	52,93
Gradin	49,07	53,01	51,40	51,67
Court line	50,13	49,25	51,85	51,93
Racket	48,93	50,99	50,26	48,34
Sky	47,03	**60,69**	**74,54**	**56,03**
Dried plant	47,14	49,63	48,09	47,97
Grass	**56,36**	53,29	47,66	50,31
Road	52,35	48,47	49,45	54,97
Vegetation	49,17	47,18	46,85	53,74
Ground	50,55	48,56	50,70	49,51
Road line	48,55	49,93	49,89	50,10
Car	47,83	49,51	47,60	47,92
Trunk	48,82	47,68	49,00	48,41
Rock	49,13	47,40	47,79	48,00
Glacier	50,00	50,00	50,00	50,00
Sea	48,73	51,32	47,87	44,08
Sand	49,62	47,86	49,66	47,41
Wave	52,85	46,54	47,96	48,29
Boat	49,76	49,74	47,63	48,77
Tree	50,47	47,61	48,41	48,25
Stone	49,58	49,50	49,39	49,55
Building	44,62	47,71	46,14	43,97
Roof	48,60	52,30	49,47	49,90

and facilitate the process of evidence-driven probabilistic inference. In this case, information from the image regions is incorporated into the analysis process, but no semantic constraints are taken into account, no FoA is used. The third configuration $CON3$ takes into account the semantic constraints of the domain using the methodology of FoA. In both CON2 and CON3 configurations, the analysis unfolds as follows. Initially, they formulate the hypotheses set using all category concepts. Then, they search for the presence of all possible regional concepts determined in Kb before deciding which of these concepts should be used as evidence. Then, the network is updated to propagate evidence impact, and the concept that corresponds to the node with the highest resulting posterior probability among the nodes that represent category concepts is selected to categorize the image.

Figure 6: F corresponds to the cut-off for the confidence degrees

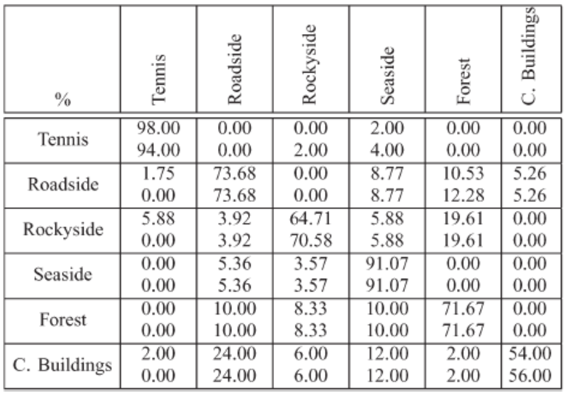


F-Measure scores for the task of image categorization using CON1, where the output of the global concept classifiers is used to categorize the image. CON2 uses knowledge and application context for categorizing the image, and CON3 also takes into account the semantic constraints expressed in an ontology.

We see that the FoA mechanism is as successful as a full brute force classification search. Using ontologies thus reduces the amount of computation while giving good results.

Figure 7:

Confusion Matrix for Image Categorization. $CON2$: Lower Part of the Cells. $CON3$: Upper Part of the Cells



%	Tennis	Roadside	Rockyside	Seaside	Forest	C. Buildings
Tennis	98.00 94.00	0.00 0.00	0.00 2.00	2.00 4.00	0.00 0.00	0.00 0.00
Roadside	1.75 0.00	73.68 73.68	0.00 0.00	8.77 8.77	10.53 12.28	5.26 5.26
Rockyside	5.88 0.00	3.92 3.92	64.71 70.58	5.88 5.88	19.61 19.61	0.00 0.00
Seaside	0.00 0.00	5.36 5.36	3.57 3.57	91.07 91.07	0.00 0.00	0.00 0.00
Forest	0.00 0.00	10.00 10.00	8.33 8.33	10.00 10.00	71.67 71.67	0.00 0.00
C. Buildings	2.00 0.00	24.00 24.00	6.00 6.00	12.00 12.00	2.00 2.00	54.00 56.00

2 Diagnosis using Bayesian networks

The paper by Sun et al[4] describes how to apply Bayesian Networks to predict mild cognitive impairment (MCI), which is considered an early symptom of Alzheimer's disease. They have a dataset with eight input variables such as age, sex, education level, and various mental test results. The output variable is the level of cognitive impairment. Some of the data is missing and no causal graph structure is know beforehand.

The paper aims to assist MCI diagnosis in the hospital. Another goal is to allow for ordinary families to detect MCI at home using simple self-diagnosis.

For prediction of MCI, the paper uses a Bayesian network, for the construction of which the K2 algorithm is used. The K2 algorithm cannot be applied to a data set with missing values. Furthermore, the variables must be pre-ordered in order of dependence for the K2 algorithm to work. The authors solve the first problem by filling in missing data from similar rows and using Newton interpolation based on similar features. The order of dependence of the variables is decided by the mutual information of each variable with the MCI variable.

Results are reported on the MCI dataset, on multiple well known datasets, and a comparison of structure learning is made between different algorithms.

2.1 Method

Let the dataset $X = \{x_1, x_2, \ldots, x_n\}$ have n rows with d features each, i.e. $x_i \in \mathbb{R}^d$ for $1 \leq i \leq n$. Let $F = \{f_1, f_2, \ldots, f_d\}$ denote the set of features.

2.1.1 Missing values

We define the mutual information between features

$$I(f_i, f_j) := \sum_{\substack{v \in \{x_{1,i}, x_{2,i}, \ldots, x_{n,i}\} \\ w \in \{x_{1,j}, x_{2,j}, \ldots, x_{n,j}\}}} p(v, w) \log \frac{p(v, w)}{p(v)p(w)},$$

where $p(\cdot)$ is the probability mass function defined by the frequency of the symbols in the data set. The mutual information is a measure of how much information is shared between two variables. It equals zero, for instance, when the variables are independent and is equal to the entropy when two variables are equal.

When a value $x_{i,j}$ is missing, the authors propose to first find the feature that has the highest mutual information with feature f_j.

When this feature has been identified, a set of rows is selected. This is done using a measure of similarity between rows a and b

$$E_{ab} := \sum_{k=1}^{d} I(f_j, f_k)(x_{a,k} - x_{b,k})^2.$$

When a row b is similar enough to row i, i.e. E_{ib} is one of the σ lowest values of all row differences for some predetermined constant σ, then the row will be included in the estimation of the missing value.

Let f_k be the most similar feature to f_j. The missing value will be estimated using Newton interpolation based on the features f_k and f_j in the σ most similar rows indexed by $r_1, r_2, \ldots, r_\sigma$. Now the missing value is calculated as follows

$$x_{i,j} = \sum_{l=1}^{\sigma} a_l n_l(x_{i,k}),$$

where

$$a_l := [x_{r_1,j}, x_{r_2,j}, \ldots, x_{r_l,j}],$$

$$n_l(x) := \prod_{i=1}^{l} (x - x_{r_i,k}),$$

and $[y_0, \ldots, y_n]$ is the notation for the divided differences.

2.1.2 Feature ordering

In order to decrease the complexity of the K2 algorithm to polynomial time an ordering on the features is required such that if f_i precedes f_j, then no arc from f_j to f_i is allowed in the resultant structure.

Let f_1 be the feature to be predicted, MCI in our case. Then we order the features based on their mutual information with f_1: those with highest mutual information come first. Of course, the sequence starts with f_1. Furthermore, if the mutual information $I(f_1, f_k)$ of some feature f_k is below a cutoff parameter, then the feature is not included in the Bayesian network.

2.1.3 Constructing the Bayesian network

When the missing values have been replaced and a suitable ordering of the features is found, the Bayesian network is constructed using the K2 algorithm [5].

2.2 Results

The authors provide results from applying their algorithm on the MCI dataset and a couple of well known datasets of the University of California Irvine repository. Furthermore, they analyse how well their algorithm recovers a known structure from data.

2.2.1 MCI dataset

The authors apply their procedure to the MCI data set. The data set contains information about 87 people, of which 42 are MCI patients. In order estimate the mean squared error, they use five-fold cross-validation.

The resulting network does not contain the age variable, because its mutual information with the MCI variable was below the cut-off point. Furthermore, the paper discusses some of the ramifications of the resulting network. They find, for example, that three variables have the most predictive power.

2.2.2 UCI datasets

The procedure is tested on seven data sets from the University of California Irvine Machine Learning Repository. On all but one data set, the procedure achieves a very low mean squared error. Unfortunately, the algorithm takes more than 10 minutes to train on a moderately large data set (20,000 rows).

2.2.3 Structure learning on alarm dataset

Finally, the paper compares the performance of their procedure with other structure learning algorithms. This is done by applying their procedure and two other structure learning algorithms to the alarm data set[6] with an increasing number of missing values. Their algorithm is the slowest, but does produces networks with the lowest average structural difference. This is the number of arcs added to, reversed, or omitted from the original network structure by their learning procedure.

2.3 Remarks

The paper gives rise to some questions. First of all, a method for sorting the features in some causal order is given, but no explanation is given why this method is chosen. The authors provide no underlying principle or theory. Secondly, Newton interpolation is suggested in order to fill in missing values, yet no reason is given why this method is preferred over other interpolation methods such as Lagrange interpolation. Finally, although the paper extensively discusses how the Bayesian network structure is learned, no mention is made of how the conditional probability table parameters are learned.

We remark that the procedure developed in this paper cannot be used for general structure learning. It requires the user the supply the variable that is to be predicted.

It is unfortunate that the authors do not compare their method of feature ordering with the maximum weighted spanning tree algorithm of Chow and Liu[7]. This method is also based on mutual information.

The following criticism of the paper is not unwarranted. The paper uses Newton interpolation on data that is strictly categorical, such as education level and sex. They fail to explain why this is a good idea and how these categories should be mapped to the real numbers. Furthermore, the authors develop a structure learning algorithm to construct a network with nine nodes (the MCI network). Perhaps asking an expert to construct the causal structure would be less work. We believe that demonstrating their procedure on this small dataset fails to demonstrate the power of their approach.

3 A Comparison of two Bayesian network approaches to classification

From a high level perspective, Nikolopoulos et al use Bayesian networks and domain knowledge to extract more information from existing image classifiers. This is achieved by comparing the classification of regions of the image with their expected labeling. Domain knowledge is converted into a Bayesian network that “knows” which elements to expect together in the same picture. This enables it to improve the decisions of a global image classifier.

On the other hand, Sun et al use the Bayesian network itself as a classifier. They shun the aid of experts and construct a causal graph automatically based on the data. They solve the problem of missing data and are able to automate structure learning.

Both papers demonstrate the power of Bayesian networks in inference problems. Specifically, they are used wherever probabilistic reasoning is required.

References

[1] S. Nikolopoulos, G. T. Papadopoulos, I. Kompatsiaris, and I. Patras, “Evidence-driven image interpretation by combining implicit and explicit knowledge in a bayesian network,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 5, pp. 1366–1381, 2011.

[2] M. Krötzsch, F. Simancik, and I. Horrocks, “A description logic primer,” *arXiv preprint arXiv:1201.4089*, 2012.

[3] Z. Ding, Y. Peng, and R. Pan, “A bayesian approach to uncertainty modelling in owl ontology,” tech. rep., DTIC Document, 2006.

[4] Y. Sun, Y. Tang, S. Ding, S. Lv, and Y. Cui, “Diagnose the mild cognitive impairment by constructing bayesian network with missing data,” *Expert Systems with Applications*, vol. 38, no. 1, pp. 442–449, 2011.

[5] G. F. Cooper and E. Herskovits, “A bayesian method for the induction of probabilistic networks from data,” *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[6] I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper, *The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks.* Springer, 1989.

[7] C. Chow and C. Liu, “Approximating discrete probability distributions with dependence trees,” *Information Theory, IEEE Transactions on*, vol. 14, no. 3, pp. 462–467, 1968.