Learning about Related Tasks with Hierarchical Models

Andrew L. Maas
Advisors: J. Andrew Bagnell and Charles Kemp
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
amaas@andrew.cmu.edu

Abstract

One of the hallmarks of human learning is the ability to apply knowledge learned in previous situations to novel scenarios. In contrast, many machine learning algorithms are explicitly designed to train and test on a single, unchanging distribution of examples. My work is aimed at better understanding how we can extend machine learning algorithms as to enhance their ability to use knowledge from previous experience in novel situations. My project explores two domains where the ability to leverage knowledge learned in previous situations enables performance improvements and new capabilities within a task domain. First, I focus on the domain of signal classification, specifically for detecting seizure events in EEG data recorded from patients with epilepsy. Second, I explore the domain of learning real-world concepts with Bayesian networks. In both of these domains, my approach relies on storing knowledge in a hierarchy with multiple levels of abstraction.

1 Classification of Epileptiform Signals

This section presents experiments which employ a hierarchical knowledge framework for signal classification. The problem domain of interest is electroencephalograph (EEG) brain activity signals from patients suffering from epilepsy. This problem domain is one where the advantages of a hierarchical approach can have significant real-world impact.

1.1 Neurology of Epilepsy

Epilepsy is a neurological disorder which affects approximately 1% of the population. The disorder is characterized by episodes of abnormally synchronous activity in the brain. Figure 2 shows EEG recordings of normal and epileptiform brain activity. There are a range of clinical symptoms of such episodes, but the disorder has the overall effect of significantly reducing the quality of life for those who suffer from it. Several treatments exist to reduce the severity and frequency of epileptic seizures. These methods include medication, surgery to remove parts of the brain most responsible for causing the episodes, and implantation of stimulation devices to disrupt overly-synchronous brain activity. Medication is insufficient for about a third of patients because it either produces intolerable side-effects, or does not completely control seizure episodes (Ellis & Stevens, 2008). Removal of brain tissue is not always possible, and has sometimes severe side-effects. For example, the amnesia of the famous patient HM was caused by brain tissue removal in an effort to combat severe epilepsy.

1.1.1 Treatment with Neural Stimulation

Electrical stimulation of the brain is a promising direction for the treatment of epilepsy. The treatment relies on the implantation of a battery-powered device into the skull of the patient. The device has electrodes which disrupt overly-synchronous brain activity by delivering electric pulses to the brain tissue. 1 shows one such device following surgical implantation into the skull of a patient.

The proper stimulations to administer to patients is a topic of current research. There are several parameters to consider when using a stimulator. These include the duration of the electric pulse, its amplitude, and frequency. Some studies use continuous stimulation of the brain, while others administer pulses intermittently (Ellis & Stevens, 2008) For patient comfort, and to maximize battery life of the devices, we would like to maximally disrupt epileptic episodes while minimizing the amount of electrical stimulation delivered to the patient. This suggests a need for stimulators which are *reactive* to the brain signals of a patient. Such reactive neural stimulation (RNS) devices exist, but their signal classification capabilities are limited. Current RNS devices use signal classification routines which are not tailored to individual patients (Anderson, Kossoff, Bergey, & Jallo, 2008). The EEG signal manifestation of an epileptic event varies across patients, so any classifier which is not tailored to individuals will likely have sub-optimal accuracy.

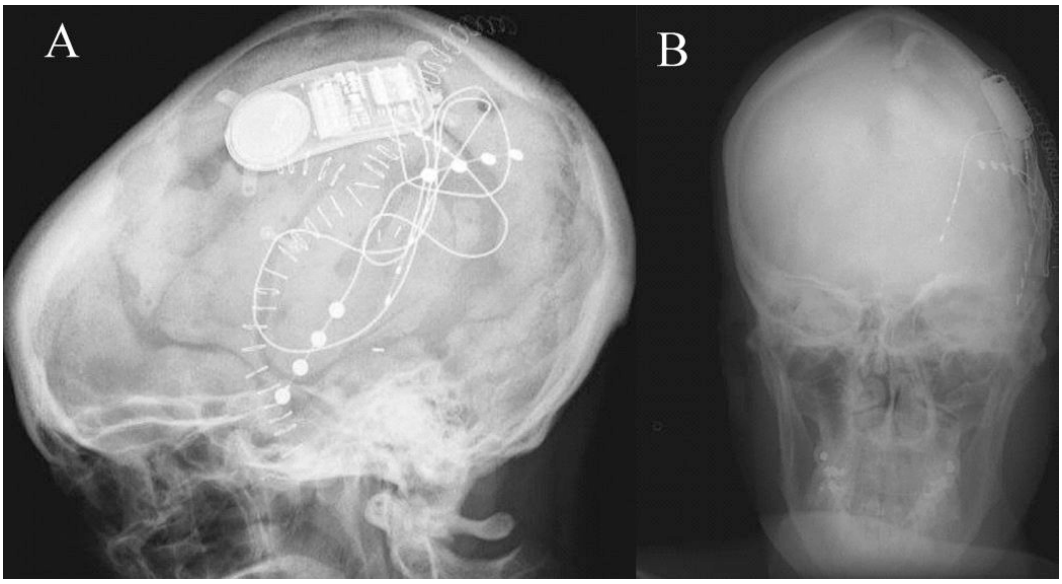

Figure 1: Radiographs showing a reactive neural stimulation (RNS) device implanted in the skull of a patient. Such stimulators hold promise to become the most effective treatment of severe epilepsy. These images are from a clinical study on the effectiveness of RNS devices by (Anderson et al., 2008).

1.2 Approach

Reviewing the state of current RNS devices suggests that there are several factors to consider when building classifiers of epileptiform signals:

1. We desire a large corpus of training data to make accurate, robust classifiers. This is especially important because the accuracy of an RNS device directly correlates with patient comfort
2. Individual differences exist in EEG signals. Additionally, RNS devices are implanted in different parts of the brain based on which regions are most responsible for causing epileptiform activity. As a result, it might not be reasonable to expect a single classifier to work well for all patients and brain regions.
3. Training data for an individual patient is limited because of the cost and risk involved in data collection. Collecting data requires insertion of recording electrodes into the skull of

a patient. In practice, this data collection only occurs before a surgical procedure, and may capture only minutes of epileptiform signal.

There are at least three approaches we might pursue which tradeoff the above three factors in varying ways. First, we can create an *aggregate* classifier which uses data from many patients to increase the size of the training corpus. This approach may sacrifice accuracy because it is not adapted to particular patients. Second, we can make *individualized* classifiers – one for each patient. This approach should learn classifiers which adapt to individual differences, but data sparsity may become an issue. Third, we can develop a *hierarchical* approach to classification. Such an approach will adapt to individuals by building a separate classifier for each patient, and combat data sparsity by sharing abstract knowledge among individual classifiers. After introducing these three approaches in more detail, their performance is compared on a data set of EEG signals recorded from several patients who suffer from epilepsy.

1.2.1 Classifier Notation

Following the standard supervised learning framework, we assume we are given (X, Y) where x_i is a data vector and y_i is a label corresponding to that data vector. In this work, we have m different patients, so we denote the data for patient h as (X^h, Y^h) The classifier used is a linear support vector machine(SVM) where the output is computed by

$$f(x) = w^T \Phi(x) \tag{1}$$

Where w is a learned weight vector, and $\Phi(x)$ is a function to map the data vector into a feature vector. Training for a linear SVM attempts to minimize the expected loss by optimizing w

$$\underset{w}{\text{argmin}}\ \mathrm{E}[L(X, Y, w)] + \|\lambda w\|_2 \tag{2}$$

We use the L^2 loss function for L. λ is a parameter controlling the amount of regularization. This work uses `LibLinear` (Fan, Chang, Hsieh, Wang, & Lin, 2008) to solve the above problem with a trust region Newton optimization method.

We can build m individualized classifiers, one for each of the m patients. To build an individualized classifier for patient h we use the training technique above applied to only the data for that patient – (X^h, Y^h). This approach assumes we have m independent data distributions, and data samples for patient h are i.i.d. samples from $P(X^h)$. This independence assumption ignores similarities among patients.

To build an aggregate classifier, we group the data for all patients into a single set. Thus we have the aggregate training set $(X^a, Y^a) = \bigcup_h (X^h, Y^h)$. Training for this classifier then proceeds using the standard linear SVM technique described above. Note that by forming the set (X^a, Y^a) we assume data for all patients are drawn from the same distribution, $P(X^a)$. The standard supervised learning framework assumes that any data point x is an i.i.d. sample from $P(X^a)$, regardless of which patient generated x. This assumption contradicts our domain knowledge about the existence of individual differences.

1.2.2 Hierarchical Classifier

Both the individualized and aggregate classifiers make assumptions about the structure of the classification problem which are incorrect. Viewing epilepsy classification as an instance of a *multi-task learning* problems allows assumptions about the data which are better suited to the domain. The general premise of multi-task learning is to explicitly model tasks such as epilepsy classification as separate but related.

We can build a hierarchical approach to a multi-task learning problem using a method proposed by (Ando & Zhang, 2005). For each task, we build a separate classifier. However, we assume that each task shares an underlying structure We can learn about this underlying structure by observing any of the tasks. Furthermore, modeling or understanding this shared underlying structure should facilitate building classifiers for individual tasks. For epilepsy classification, we consider each patient as a classification task, and we wish to discover a shared, underlying structure upon which we can build classifiers. The multi-task learning framework allows us to consider each patient as separate while

leveraging the similarities between patients. Furthermore, the shared structure approach to multi-task learning results in an efficient algorithm for linear SVM classifiers

A central question in the structure learning approach to multi-task learning is how to model the shared problem structure. Because the linear SVM computes its output using Equation 1, one method for creating a shared structure is to assume

$$w = v\Theta \tag{3}$$

Here Θ is a matrix shared across all classifiers. The purpose of this matrix is to map the feature vector $\Phi(x)$ into a shared, low-dimensional subspace. Θ is learned in addition to the other terms which compose w. Learning Θ corresponds to learning a function which maps the input data x into a low-dimensional subspace which has (hopefully) high predictive power. There are several candidate methods for learning Θ. Following the method proposed by (Ando & Zhang, 2005), we construct Θ using singular value decomposition(SVD) applied to the weight matrix S where each row of S is a weight vector s_i produced by training a linear SVM on X^i. Note that using SVD in this way is qualitatively different than performing SVD or some other type of dimensionality reduction on the input data itself.

Algorithm 1 Train Hierarchical Classifiers

Require: Input Data $X^1, X^2, ...X^m$
  for $i = 1$ to m **do**
    $s_i \leftarrow$ TrainSVM(X^i)
  end for
  $S \leftarrow [s_1; s_2; ...s_m]$
  $\Omega \leftarrow$ SVD(S)
  $\Theta \leftarrow \Omega(1 : n, :)$
  for $i = 1$ to m **do**
    $w_i \leftarrow \Theta s_i$
  end for
  return all w_i

1.3 Experiments

The data set contains invasive (intracranial) EEG recordings from 21 epilepsy patients. For each patient, there is approximately 24 hours of recording which took place during pre-surgical monitoring. This data set was provided by the Epilepsy Center of the University Hospital of Freiburg, Germany. Signals representing epileptiform activity, called *ictal* events, were labeled by an expert. The remainder of the data is called *interictal* because it either represents normal activity, or pre-ictal warning signs of an oncoming seizure.

1.3.1 Signal Processing

Each data trace was segmented into non-overlapping frames of 1 second in duration. Note that the sequential information is not used. Instead, each frame is treated as an isolated sample. This process yielded approximately 88,000 samples per patient. Because ictal periods are fairly infrequent, less than 1% of samples are of ictal events.

Each signal frame was transformed into a feature vector using a method similar to that in (Vincent, Pineau, Guzman, & Avoli, 2007). Each frame was normalized by subtracting the mean and dividing by the full range of the entire frame. The per-frame mean and range are used as features Each frame was then apodized with a Hann window and converted to a power spectrum using the discrete fast Fourier transform. 2 shows sample ictal and interictal signals, and their power spectra. The power spectrum from 1-256 Hz is then divided into 128 frequency bands. The real and imaginary components of each band of the FFT were combined into a single magnitude. This results in 130 features per frame (mean, range, 128 power spectrum bands).

1.3.2 Training

The signal processing technique gives for each patient a data set X already transformed with Φ. Using these datasets, we can use the training techniques already described to develop individualized,

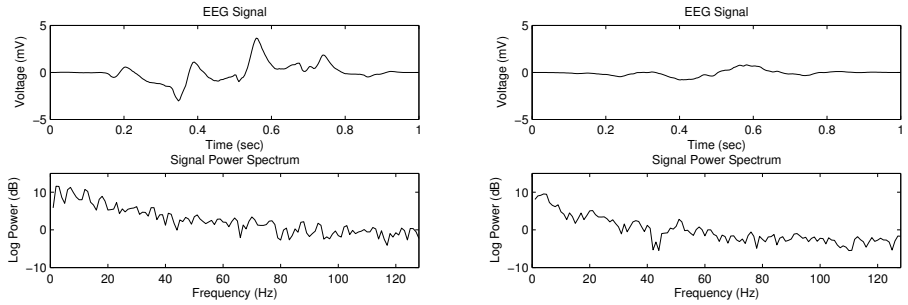

Figure 2: Sample Signals and their Power Spectra. A one second window of EEG ictal (left) and interictal (right) signals. The log power spectrum for each is also shown. Power at different frequency bands are used as features in the classifiers.

aggregate, and hierarchical classifiers. The data was randomly split with 10% used for training, and 90% for testing. To counteract the extreme inbalance between the amount of positive and negative examples, each positive example was repeated 1000 times in the training set.

1.3.3 Results

Table 1 shows the results for each type of classifier on the epilepsy classification task. The individualized classifiers all learned a sufficient model to nearly always correctly identify the interictal examples, but did so at a high false-negative rate. The individual classifiers' low accuracy for ictal samples is most likely due to a lack of training data. Because only 10% of the data was used for training, the number of ictal samples in the training set was extremely low. For example, patient 1 had only 4 ictal examples in the training set. With such data scarcity, it is not surprising that the individual classifiers did not learn sufficiently robust models. The aggregate classifier overcame some of the data scarcity issues because it observed all the training data at once in order to create its single classifier. This resulted in more robust ictal detection, but at the cost of an increased false-positive rate. The hierarchical classifiers show the best ictal classification accuracy, but perform comparatively poorly on interictal samples. The high ictal accuracy suggests that the hierarchical approach is effectively leveraging all of the data and that learning about the underlying domain structure has a positive effect.

Table 1: Accuracy for the Epilepsy Classification Task. Individual models were trained using only data from the test patient. The aggregate approach trained a single model using the data from all patients. The hierarchical approach builds an individualized model for each patient, but learns a low-dimensional subspace with high predictive power by leveraging the data from all patients

Patient	**Individual**		**Aggregate**		**Hierarchical**	
	Ictal	**Inter**	**Ictal**	**Inter**	**Ictal**	**Inter**
1	2.12	99.9	51.1	79.1	80.9	73.0
2	38.3	99.6	50.5	87.0	40.1	83.0
3	12.1	99.3	24.2	83.4	55.0	61.5
4	1.6	99.8	47.9	81.0	40.2	69.2
5	28.4	99.0	33.7	89.5	26.7	99.1
Average	16.5	99.5	41.5	84.0	48.6	77.1

1.4 Conclusion

This work presented an approach to classification of epileptiform signals which employs a hierarchical knowledge framework. The hierarchical approach learns at the high level an underlying domain structure, and maintains a more specific knowledge level which is a classifier for each patient. This approach was compared with two simple models of the domain. While the performance gain of the hierarchical model was not overwhelming, this might be attributed to features of the problem domain

such as training data scarcity. This work contributes to a growing body of evidence that explicitly modeling problems as multi-task learning is advantageous when the problem domain suggest such a model is appropriate.

For the future design of RNS devices, a hierarchical approach to classification should provide a principled way of customizing to an individual while still benefitting from the robustness obtained by training on a large data set. The classifiers presented in this work are not expected to be immediately useful for an RNS device for two reasons. First, the classifiers presented here did not reason about time. Knowing that there is an ictal event in the current time frame should heavily bias what a classifier predicts in the next time step. Second, the classifiers here did not consider the task of *prediction*. To avoid any discomfort of the patient, it is best to prevent synchronous brain activity from building into an ictal event.

2 Hierarchical Knowledge for One-Shot Learning in Bayes Nets

Humans are able to discover and exploit relationships between attributes (e.g. nationality and language) and between attribute values (e.g. Brazilian and Portuguese) (Davies & Russell, 1987). Some relationships are near-deterministic, including the relationship between birth country and native language. We know, for example, that two individuals born in the same country are very likely to have the same mother tongue, and we know in particular that individuals born in Brazil are very likely to speak Portuguese. Other relationships are probabilistic, including the relationship between hair color and eye color. We know that these attributes tend to be related, and we know about specific relationships between values of these attributes (blondes often have blue eyes).

Suppose, for example, that after meeting several people from various countries, you meet a single person from Randeria, a country that is completely new to you. You observe that the person has blonde hair and speaks Randerian. Based on this single example, you may be very confident that the next Randerian you meet will speak the same language, but less confident that this second Randerian will also have blonde hair. Figure 3(a) shows a schematic representation of the observed data, and Figure 3(b) shows conditional distributions that capture our expectations about the language and hair color of the second Randerian. The Randeria problem just introduced is a special case of the more general problem of *one-shot learning* (Fei-Fei, Fergus, & Perona, 2003). Here we describe and evaluate a probabilistic model that can handle one-shot learning problems similar to the Randeria problem.

One-shot learning has been previously considered in the psychological literature. One prominent line of work has focused on "fast mapping" in word learning (Carey & Bartlett, 1978; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Empirical studies of word learning have documented that children are able to learn the meaning of some new words given a single training example and researchers have developed formal models (Colunga & Smith, 2005; Kemp, Perfors, & Tenenbaum, 2007) that help to explain this ability. Our approach grows out of this literature, and the work we describe builds on the hierarchical Bayesian model presented by Kemp et al. (2007). Hierarchical Bayesian models (Gelman, Carlin, Stern, & Rubin, 2003) can include representations at multiple levels of abstraction, and help to explain how humans acquire abstract knowledge that supports rapid or one-shot learning given exposure to a novel situation.

Our hierarchical Bayesian approach is built on top of a standard method for learning Bayesian networks, also known as Bayes nets. A Bayes net captures relationships between attributes using probability distributions that specify how the value of a given attribute is generated given the values of its parents. Our approach allows for two kinds of relationships: relationships where an attribute value is a soft probabilistic function of the values of its parent attributes, and relationships where an attribute value is generated in a near-deterministic way given the values of its parents (Figure 4b). By learning which relationships are probabilistic and which are near-deterministic, a Bayes net approach can account for one-shot learning while preserving the ability to handle probabilistic relationships.

After reviewing related work and introducing our approach, we apply it to an everyday problem that requires one-shot inferences—learning about people and their characteristics. Using demographic data for immigrants who arrived at Ellis island in the early twentieth century, we introduce two one-shot learning scenarios which correspond to real-world versions of the Randeria problem. We show

(a)

Nationality	Language	Hair
British	English	Red
British	English	Brown
Brazilian	Portuguese	Blonde
Brazilian	Portuguese	Brown
Spanish	Spanish	Brown
Spanish	Basque	Black
Randerian	Randerian	Blonde
Randerian	?	?

(b)

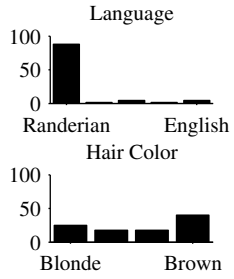

Figure 3: Randeria one-shot learning problem. (a) After meeting people from several different countries, you might discover that people from the same country tend to speak the same language. (b) Discovering the pattern in (a) supports one-shot learning about people from a new country. After observing a single Randerian, you might have strong expectations about the language spoken by a subsequent Randerian, but weak expectations about her hair color.

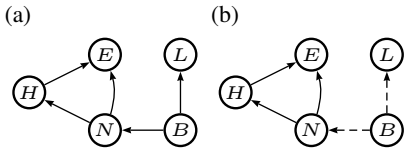

Figure 4: Models that capture relations among five attributes: birth country (B), language (L), nationality (N), eye color (E) and hair color (H). (a) A standard Bayes net can capture probabilistic relationships between attributes, shown here as solid arrows. (b) Our model learns Bayes nets that capture two kinds of relationships: near-deterministic relationships (dashed arrows) and probabilistic relationships (solid arrows).

that our model makes more intuitive inferences and predicts unobserved data better than a standard Bayesian network approach.

2.1 Logical Approaches To One-Shot Learning

One-shot learning has been previously considered by AI researchers, and the Randeria example introduced above is directly inspired by the work of Davies and Russell (1987). These researchers explore the role of determinations, or abstract logical statements that identify patterns of dependency between attributes. For example, the statement that "people of the same nationality speak the same language" is a determination that supports the conclusion that all citizens of Randeria are likely to speak the same language. Because this rule is defined over attributes, it is independent of any particular country and can be used to perform one-shot learning when exposed to a person from a new country. Russell (1989) discusses how determinations can be learned given a database such as the schematic example in Figure 3(a). The basic approach is to search through a hypothesis space of possible determinations and identify hypotheses that are consistent with the entries in the database.

A probabilistic approach to learning determinations can improve on existing work in several respects. First, a probabilistic approach can handle near-deterministic relations that are subject to noise and exceptions. Some citizens of Randeria may be English speakers who were born in the USA, and some countries (e.g. Spain) include different linguistic communities (e.g. Spanish speakers and Basque speakers). Second, a probabilistic approach can incorporate soft probabilistic relations, including the relationship between blonde hair and blue eyes. Russell (1989) allows for weighted determinations which can help to deal with uncertainty, but a probabilistic approach provides a principled treatment of reasoning under uncertainty. Finally, a probabilistic approach can provide a unified account of learning and using determinations. Logical approaches can rely on logical inference to explain how determinations are used, but must typically invoke some other principle to explain how these determinations are acquired.

There has traditionally been some tension between logical and probabilistic approaches to artificial intelligence, but several researchers have recently developed general-purpose frameworks that combine logic and probability (Milch et al., 2005; Richardson & Domingos, 2006). Some of these frameworks may be able to address the one-shot learning problems described earlier, but here we take a different approach. General-purpose frameworks are impressive in their scope, but the flexibility of these approaches often leads to very difficult learning problems. Here we describe a relatively simple probabilistic approach that relies on one of the best known formalisms for capturing relationships between attributes—Bayesian networks.

2.2 Learning Bayesian networks

A Bayesian network includes a graph and a set of distributions that specify probabilistic relationships between attributes. This section introduces a standard approach to learning and using these networks (Heckerman, Geiger, & Chickering, 1995).

A Bayes net can be represented as a pair (G, θ), where G is a directed acyclic graph over the attributes of interest and θ_i specifies the conditional probability distribution for attribute i, or the distribution over values of this attribute given the values of its parent attributes in graph G (Figure 5). Figure 4a shows a Bayes net graph structure over some of the attributes in the Randeria problem.

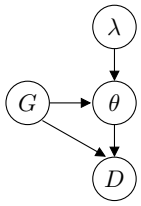

Figure 5: Graphical model for Bayes net structure learning. (G, θ) is a Bayes net, where G is a directed acyclic graph, and θ_i is a table that specifies the conditional probability distributions for node i in the graph. Each row in θ_i is drawn from a symmetric Dirichlet distribution with parameter λ_i.

We assume here that all attributes are categorical, and represent θ_i as a conditional probability table (CPT) with one row for each setting of the parent attributes. Each row in θ_i specifies a multinomial distribution over values of attribute i, and we assume that these rows are independently drawn from a symmetric Dirichlet distribution with concentration parameter λ_i. A standard approach to structure learning sets $\lambda_i = 1$ for all attributes in the graph, which corresponds to a uniform prior over possible multinomial distributions for the rows in each CPT.

Suppose that we observe a data matrix D, where the rows in D represent independent samples from a Bayes net (G, θ). The posterior distribution over the components of the Bayes net is

$$p(G, \theta | D, \lambda) \propto p(D|G, \theta) p(\theta|G, \lambda) p(G) \tag{4}$$

and we assume a uniform prior $p(G)$ over graph structures G. Since we use conjugate Dirichlet priors on the rows in each CPT, we can integrate out the parameters θ and work with the posterior distribution $p(G|D, \lambda)$ over graphs (Heckerman et al., 1995). We can sample from this distribution using standard MCMC techniques for structure learning (Giudici & Castelo, 2003). If we assume that any missing entries in D are missing at random, a bag of samples from $P(G|D)$ can be used to make predictions about these missing entries.

Bayesian networks have been widely used in the psychological literature to develop formal models of learning and reasoning (Glymour, 2001; Gopnik et al., 2004) The standard approach to learning these networks, however, cannot address one-shot learning problems like the Randeria problem. This limitation depends critically on the difference between attributes (e.g. nationality) and attribute values (e.g. Brazilian). Given enough data, the standard approach will be sensitive to near-deterministic relationships between attribute *values*. After observing many Brazilian individuals, for example, the standard approach will learn parameters for the network in Figure 4a that specify a near-deterministic relationship between being Brazilian and speaking Portuguese. No amount of experience, however, will allow the standard approach to exploit near-deterministic relationships between *attributes*. The standard approach can learn that Brazilians tend to speak Portuguese, and

that Americans tend to speak English, and so on, but cannot arrive at the generalization that individuals from a given country tend to speak the same language. The next section introduces a Bayesian network approach that overcomes this limitation.

2.3 The Type-Learning Model

Our approach relies on the same basic machinery as the standard approach, except that we no longer assume λ is fixed to a single, known value for all attributes in the graph. Instead we assume that attributes come in one of two types: *non-deterministic* attributes are generated in a soft probabilistic way by their parents in the graph, but *near-deterministic* attributes are generated according to a near-deterministic function of their parent attributes. To capture the difference between these types of attributes, we assume λ_i will be smaller for near-deterministic attributes than for non-deterministic attributes. A small value of λ_i means that each row in CPT θ_i is expected have most of its probability mass concentrated on a single value of attribute i. Setting $\lambda_i = 1$, which is a standard practice when learning Bayes nets, means that each row of θ_i is drawn from a uniform prior over multinomial distributions.

A type-based approach could be implemented by assuming that each λ_i is drawn from one of two distributions: a distribution with a small mean for the near-deterministic attributes, and a distribution with mean 1 for the non-deterministic attributes. Here we take a simpler approach, and assume that $\lambda_i = 1$ for non-deterministic attributes but that $\lambda_i = 0.01$ for near-deterministic attributes. Note, however, that the type assignment for each attribute is not known in advance and must be learned.

A type-based approach can be contrasted with a type-free approach that assumes that the λ_i are independently generated from a continuous prior distribution such as an exponential distribution. These two approaches incorporate different inductive biases and should lead to slightly different predictions—for example, the type-based approach might be quicker to decide whether a given attribute is near-deterministic (low λ_i) or non-deterministic (high λ_i). Future work can consider whether a type-based or a type-free approach accounts better for human inferences. Note, however, that both approaches are consistent with our core proposal, which is that learning different values of λ_i for different attributes can allow a Bayes net approach to handle one-shot learning problems like the Randeria problem.

Since the type assignments that determine λ are not known in advance, we work with a posterior distribution created by summing over all possible values of λ:

$$p(G, \theta|D) \propto p(D|G, \theta)p(\theta|G)p(G) \tag{5}$$

$$= \sum_{\lambda} p(D|G, \theta)p(\theta|G, \lambda)p(G)p(\lambda) \tag{6}$$

We use a uniform prior over type assignments, which amounts to a uniform prior over the two possible values of λ_i for any attribute i. Standard MCMC techniques for structure learning can be extended to sample from $P(G, \lambda|D)$, but for the small data sets considered here we compute Equation 6 by enumerating all possible values of λ. As for the standard approach in Equation 4, the parameters θ can be integrated out for any given value of λ, and we make inferences about missing values in D using a bag of samples from the learned distribution $P(G, \lambda|D)$.

2.3.1 Related Work

A special case of our general approach has previously been discussed in the psychological literature. Kemp et al. (2007) describe a Bayesian model that can discover, for example, that objects in the same category tend to have the same same shape—in other words, that the relationship between category label and shape is near-deterministic. Their model, however, works with a restricted class of Bayes nets where there is an arrow from the category label attribute to each other attribute, and where no other edges are allowed. The model developed here can handle Bayes nets with arbitrary structure, including networks that specify relationships between attributes (e.g. hair color and eye color) that do not correspond to category labels.

Our emphasis on near-deterministic relationships is consistent with previous suggestions that humans assume by default that causal relationships will be deterministic (Schulz & Sommerville, 2006). Previous researchers have developed probabilistic approaches that can exploit deterministic

Table 2: Passenger Data Attributes

Attribute	Example	# Values
Nationality	Spain	24
Race	Spanish	16
Language	Spanish	12
Birth Country	Spain	24
Complexion	Dark	2
Hair	Black	4
Eyes	Brown	7

relationships when they are present. Closest to our own approach is the work of Lucas and Griffiths (2007), who describe a hierarchical Bayesian model that can learn whether causal observations are better explained by a deterministic relationship or a noisy-OR relationship between variables. Note, however, that this model does not handle settings where a single network includes both near-deterministic and non-deterministic relationships, and cannot address one-shot learning problems like the Randeria problem considered here.

Our approach to one-shot learning relies critically on the concentration parameters λ_i used to define the Dirichlet priors on the Bayes net parameters θ. We know of no previous work that explores one-shot learning with Bayesian networks, but several previous researchers have emphasized the role of the Dirichlet priors. One line of work explores structure learning in the standard setting where there is a single value of λ for all nodes in the network, and has demonstrated that the value of this parameter plays an important role in determining the graph structure G that maximizes $P(G|D)$ (Steck, 2008; Silander, Kontkanen, & Myllymäki, 2007). When λ is very small, the best graph structure will often have very few edges, and as λ increases the number of edges in the inferred graph will also tend to increase. This result suggests that the value of λ matters, and supports the idea that predictive accuracy may be improved by choosing different λ_i values for near-deterministic and non-deterministic nodes.

Previous authors have explored the possibility of learning a single λ parameter for the entire network (Giudici & Green, 1999), but there are few attempts to learn different values of λ_i for different attributes. One possible reason is that this approach is inconsistent with the assumption of likelihood equivalence, or the assumption that networks in the same Markov equivalence class should receive the same prior probability (Heckerman et al., 1995). Although likelihood equivalence is often appealing, it will not always apply in settings where prior knowledge is available about network parameters. Our setting is one example, and the knowledge in this case specifies that some relationships are near-deterministic but that others are probabilistic.

2.4 Experiments

We evaluate our approach in two ways using a real-world data set. First, we directly model the Randeria problem to show the practical consequences of modeling near-deterministic relationships. Second, we use a larger test set to demonstrate the quantitative differences between inferences made by our model and a standard Bayes net approach.

2.4.1 Passenger Data

Our experiments used a real-world version of the data set shown schematically in Figure 3(a). The data specify physical and cultural properties of immigrants who arrived at Ellis Island during the 1920s and 1930s, and were extracted from passenger manifests available at `ellisisland.org`. We took manifests for 4 ships and created a data set with 85 people and 7 categorical attributes[1]. Table 2 shows each attribute, its number of possible values, and example values for one person. The relationships between the attributes include both near-deterministic relationships (country determines language) and soft probabilistic relationships (hair color predicts eye color). Note, however, that the near-deterministic relationships are not perfectly clean (e.g. not everyone from Spain speaks Spanish).

[1]The data set is available online at `www.andrew-maas.net`

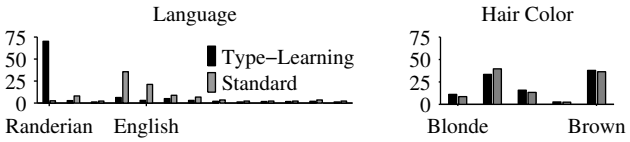

Figure 6: Conditional distributions on the language and hair color of a new person given only the information that she is Randerian. These marginals are analogous to those in Figure 3(b), but are computed by models trained on real-world passenger data.

Our first experiment addresses the Randeria problem schematically described in Figure 3. Our second experiment explores prediction of missing attributes when these hidden attributes were specifically chosen to create one-shot learning problems similar to the Randeria example. Both experiments rely on learning the structure of a Bayesian network, and we first present structure-learning results for the passenger data.

2.4.2 Learning Model Structure

Structure learning for the standard model can be achieved by drawing a MCMC sample from $P(G|D,\lambda)$, where each λ_i is set to 1. For the type-learning model we drew an MCMC sample from $P(G|D,\lambda)$ for each possible setting of λ. Given these samples, we constructed an approximate posterior $P(G,\lambda|D)$ by computing the relative posterior probabilities of each pair (G,λ) then normalizing.

Both models learned distributions on graph structures which capture some of the intuitive relationships between the seven attributes. For example, both models predict with high confidence that there is an edge between the birth country and nationality attributes. The structures assigned high probability by the type-learning model tend to have more edges than the structures preferred by the standard model. Adding more edges allows the model to explain certain attributes as near-deterministic functions of their parents.

For any training set D, we use the above training technique to obtain structure distributions $P(G|D,\lambda)$ for the standard model and $P(G,\lambda|D)$ for the type-learning model. These distributions serve as the basis for predictions about unobserved attributes.

2.4.3 Meeting a Randerian

Our first test directly corresponds to the Randeria problem mentioned in the introduction. We took the passenger data already described and added a record for a single Randerian—an individual with blonde hair, a fair complexion, and blue eyes, but a new nationality, race, language and birth country. Using the training technique described in the previous section, the models infer structure distributions and network parameters. Both models were then asked to predict the language and hair color of a second individual that was known to be Randerian, but had no other attributes observed. Figure 6 shows the marginal distributions over language and hair-color for both models.

Only the type-learning model was able to confidently predict that a second Randerian would also speak Randerian based on the single training instance provided. When predicting hair color, both models produce similar distributions over the possible values. Despite allowing for near-deterministic relationships, the type-learning model correctly realizes that hair color is not a near-deterministic function of nationality.

2.4.4 One-Shot Learning Tests

Figure 6 suggests that the type-learning model matches our intuitive notion about correct performance on the Randeria problem, and our next analysis explores a setting where model success can be assessed more objectively. We took the passenger data and created a series of one-shot learning problems for each attribute value. For example, we create a one-shot learning problem for the case where Language=French by removing all French-speaking passengers except one from the training set. The test set contains all of the French speakers that were removed, and the task is to predict the

Table 3: One-shot learning tests. Each model was shown a single instance with a given attribute value (e.g. a single French-speaking passenger) and asked to make inferences about all other instances with this attribute value.

Missing	**KL Divergence**		**Accuracy (%)**	
Attribute	**TL**	**Standard**	**TL**	**Standard**
Nationality	1.46	2.72	73	58
Race	1.74	2.16	63	36
Language	1.38	2.16	60	60
Country	1.23	2.32	82	45
Complexion	1.99	1.96	13	18
Hair	3.22	3.28	0	0
Eyes	3.26	3.33	0	0

language of each individual given all of their other attributes. In other words, we explore whether the models can confidently identify French speakers after observing a single example of this category. We repeated this process for each value of each attribute in the passenger data.

To evaluate the models we measure both model accuracy and model confidence. We expect that near-deterministic relations will allow confident predictions based on a single training instance, and use Kullback-Leibler(KL) divergence as a metric of model confidence. We considered the models' inferred marginals as approximating distributions to the true marginal, KL(true||inferred). The true marginal is a point-mass distribution which assigns all of its probability to the correct attribute value. In this case, the KL-divergence simplifies to $-\log[p(v_t)]$ where $p(v_t)$ is the probability a model assigns to the true attribute value.

Table 3 shows the results of the one-shot learning tests for both models. As expected, the type-learning model made more confident inferences for attributes with near-deterministic relations given only a single training example. Given a single instance of a passenger from a new country, for example, the model achieves high accuracy and confidence (as measured by a low KL divergence) when predicting the country attribute for subsequent passengers from that country. In contrast, the standard model was often unable to make confident one-shot inferences. Although this model made inferences from the single target instance at a rate better than chance, it had substantially lower confidence and accuracy for attributes with near-deterministic relations. Both models performed comparably for the three non-deterministic attributes. We do not expect one-shot learning to be possible for these attributes, and accuracy was low in all cases.

2.4.5 Randomized Test Set

Our results so far suggest that the type-learning model outperforms the standard model when applied to one-shot learning problems. It is important, however, to verify that this success is not achieved at a cost to performance on more traditional inference tasks. We therefore considered a second task where twenty passengers are randomly chosen from the full set of 85 to serve as a test set. Since the test set is randomly chosen, one-shot learning is very unlikely to be required.

In a first experiment, the models inferred attribute values for data instances that were otherwise fully observed. This problem corresponds to meeting a new person and making an inference about a single attribute (e.g. language) after observing all of his or her other attributes. The results of this experiment are shown in Table 4. The accuracies achieved by the two models are comparable. Reasoning about relation type should not significantly impact model accuracy in this scenario because there was sufficient training data overcome the uninformative prior placed by the standard model on the CPTs of the Bayesian network. Note, however, that the KL divergence metric shows a substantial difference in the confidence of the two models. For the variables with near-deterministic relations, the type learning model has half the KL divergence of the standard model for some variables. For variables with non-deterministic relations, the type learning and standard models have similar KL divergences.

We then ran a second experiment where the number of attributes observed for the target individual varied from 6 down to 1. For each number of observed attributes k, we averaged across all possible ways in which an instance could have only k attributes visible during inference. For example, to

Table 4: Standard learning tests. The models infer attribute values for passengers that are otherwise fully observed.

Missing Attribute	KL Divergence		Accuracy (%)	
	TL	Standard	TL	Standard
Nationality	0.57	1.46	85	85
Race	0.81	0.96	85	85
Language	0.32	0.74	90	85
Country	1.24	1.85	70	65
Complexion	0.44	0.46	85	85
Hair	1.21	1.35	50	40
Eyes	1.24	1.43	45	45

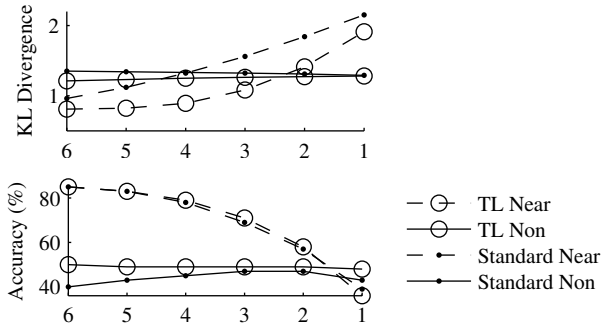

Figure 7: Performance on the standard learning test as the number of observed attributes for each test instance decreases. The four curves in each plot show average results for two models and two types of variables: near-deterministic variables and non-deterministic variables (See Table 2). Over the entire range of observed attributes, the type-learning model achieves higher confidence than the standard model as measured by the KL divergence metric.

compute the inference performance for language with $k = 1$ observed attribute, we considered predictions of language given nationality, language given race, language given country, and so on.

Figure 7 shows how model performance degrades as the number of variables observed during inference decreases. The models achieve similar accuracies for each number of observed variables, suggesting that the more complex graph structures learned by the type learning model do not have detrimental effects on inference as the number of observed variables decreases. The KL metric shows that even as the number of observed attributes decreases, the type learning model is able to make inferences with substantially higher confidence.

2.5 Modeling Human One-Shot Learning

The type learning model is inspired in part by the idea that human learners are able to detect and exploit near-deterministic relationships between variables. Our second set of analyses evaluates our approach as an account of human learning. We consider a study of one-shot learning conducted by Billman and Dávila (Billman & Davila, 1 October 2001). Participants in this study observed instances of three categories, and then observed a single instance of a new category. One training example was enough to support inferences about this new category, and we demonstrate that the type learning model can account for this result.

The data provided during training are summarized in the left section of Table 5. Each column in the table represents an animal, and the animals vary along six dimensions. Category labels for each animal were provided during training, and there was a deterministic relationship between the category of an animal and its values along two of the dimensions. These two *relevant* dimensions are shown as the top two rows in Table 5: note that any two instances with the same category label share the same values along these dimensions. The two relevant dimensions were randomized across participants. In some cases, for example, an animal's mode of locomotion and its sound were the

Table 5: Training and Test Data for the Human Learning Experiment. Each column is an animal instance.

	Training													Test			
Relevant	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	6	5
	1	1	1	1	2	2	2	2	3	3	3	3	4	4	5	4	5
Not Relevant	1	1	3	3	2	3	2	3	1	3	2	3	4	5	5	8	5
	1	3	1	3	2	1	2	2	3	3	2	2	4	5	5	5	7
	1	1	2	2	1	2	2	3	1	2	2	3	4	5	5	5	4
	1	2	1	2	1	1	3	3	2	2	3	3	4	5	5	4	4
Label	1	1	1	1	2	2	2	2	3	3	3	3	4	?	?	?	?

two relevant dimensions, and any animals that flew and made a whistling sound were members of the same category.

Following the training phase, participants were shown a *novel item*: an instance of a novel category that had new values along each dimension. In Table 5, the novel item is represented by the single column with category label 4. Participants were then shown a sequence of test animals, and asked to respond 'yes' or 'no' depending on whether each animal belonged to the same category as the novel item (i.e. category 4 in Table 5). These test items varied in the number of attributes that matched the novel item, and in the extent to which the matching attributes overlapped with the relevant attributes. Some representative test animals are shown in the right part of Table 5. Half of the test items were 'yes' items: they matched the novel item along the two relevant dimensions, and should therefore belong to the same category as the novel item. Given only a single example of the novel category, participants were able to distinguish the 'yes' items from the 'no' items, and chose the correct response for 76% of the test items.

To model this task, we consider Bayes Nets defined over the seven variables shown in Table 5. One of these variables—the category label—is qualitatively different from the others, and we assume that the remaining attributes are generated from this variable. In other words, we restrict the space of possible graph structures so that all candidate edges emerge from the category label variable. Our approach is therefore closely related to a Naive Bayes classifier, and to similar approaches from the psychological literature (Rehder & Burnett, 2005).

Given the 12 training examples, the standard model and the type learning model both learn a distribution over graph structures. Since the space of possible structures is relatively small, we create an exact posterior distribution by enumerating all structures and computing the relative probability of each one. Both models discover that the edges most likely to exist are the edges joining the category label to the two relevant dimensions, but other edges are also assigned non-negligible probability. These additional edges appear to capture some of the statistical properties that are weakly present in Table 5: for example, items in category 1 never have value 3 along dimension 6, but this value appears half of the time for items in categories 2 and 3. The type learning model discovered that the two relevant attribute variables were near-deterministic functions of the category label with 99% probability. All other variables were treated as having soft probabilistic relations. After observing the 12 training examples, each model observed the novel item and then made inferences about the category label of each item in the test set. To model the yes/no choice used in the behavioral experiment, we assume that a model says 'yes' to a test item if the posterior probability that this item belongs to the same category as the novel item exceeds 50%.

Figure 8 shows the percentage of correct responses for the 'yes' items (i.e. the items that matched the novel item along the two relevant dimensions). These responses are organized into three groups: responses for 'yes' items that match the novel item only along the two relevant dimensions, and responses for test items that match along two or three dimensions. (Billman & Davila, 1 October 2001) suggest that 76% of the 'yes' items are correctly identified, but do not report the percentage of 'yes' responses for each group. They state, however, that there is no correlation between the probability that an item is successfully identified and the number of dimensions along which it matches the novel item, and we therefore assume that the percentage of 'yes' responses is roughly 76% for all groups.

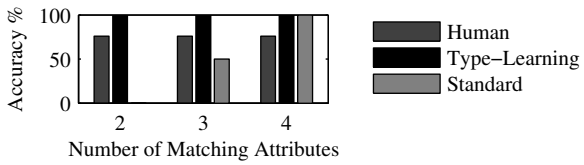

Figure 8: Behavioral data and model predictions on a one-shot learning task. The left groups of bars shows the percentage of correct 'yes' responses for 'yes' items that matched the novel item only along the two relevant dimensions. The two remaining groups show results for 'yes' items that matched the novel item along three or four dimensions respectively.

As shown in Figure 8, the type-learning model performs the task perfectly, saying 'yes' with greater than 90% confidence for all cases. The responses of the standard model depend on the number of attributes shared by the novel item and a given test item. When a test item shares only the two relevant attributes, the model says 'no' in all cases, but when a test item shares four attributes, the model says 'yes' in all cases. This profile of responses departs significantly from the behavioral data in Figure 8, and suggests that the type-learning model accounts better than the standard model for one-shot learning in humans.

2.6 Conclusion

Humans often make accurate inferences given a single example of a novel situation, and we presented a model that attempts to match this ability. Our model uses a Bayes net to capture relationships between attributes, and learns which of these relationships are soft and probabilistic and which are near-deterministic. The ability to exploit near-deterministic relationships gives our approach a different inductive bias than a standard Bayes net approach, and we showed that this inductive bias supports one-shot learning about novel situations.

Here we focused on a specific one-shot learning problem—the Randeria problem—that is motivated by real-world inferences made by human learners. Future studies can design behavioral experiments to test our approach, and can explore, for example, how people make inferences about unobserved entries in the passenger data that we analyzed. Future experimental studies can also explore one-shot learning in other settings. Kemp et al. (2007) describe a special case of our approach that helps to explain word-learning data collected by Smith et al. (2002), and our current approach should account for all of the findings captured by this previous model. This previous model, however, can only learn Bayesian networks that belong to a very restricted class. Future studies of one-shot learning can test our prediction that people can learn and reason about a much broader class of relationships.

References

Anderson, W., Kossoff, E., Bergey, G., & Jallo, G. (2008). Implantation of a responsive neurostimulator device in patients with refractory epilepsy. *Neurosurgical Focus*, *25*.

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Billman, D., & Davila, D. (1 October 2001). Consistent contrast aids concept learning. *Memory and Cognition*, *29*, 1022–1035.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and reports on child language development*, *15*, 17–29.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*(2).

Davies, T. R., & Russell, S. J. (1987). A logical approach to reasoning by analogy. In *IJCAI 10* (pp. 264–270).

Ellis, T., & Stevens, A. (2008). Deep brain stimulaiton for medically refractory epilepsy. *Neurosurgical Focus*, *25*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874.

Fei-Fei, L., Fergus, R., & Perona, P. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *ICCV 9*.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Giudici, P., & Castelo, R. (2003). Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, *50*, 127–158.

Giudici, P., & Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, *86*, 785-801.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

Lucas, C., & Griffiths, T. (2007). Learning the functional form of causal relationships. In *Proceedings of the 29th annual conference of the cognitive science society* (p. 1810). Austin, TX: Cognitive Science Society.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *IJCAI 19* (pp. 1352–1359).

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*, 107–136.

Russell, S. J. (1989). *The use of knowledge in analogy and induction*. London: Pitman.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: causal determinism and children's inferences about unobserved causes. *Child Development*, *77*(2), 427–442.

Silander, T., Kontkanen, P., & Myllymäki, P. (2007). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *UAI 23*.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Steck, H. (2008). Learning the Bayesian network structure: Dirichlet prior vs data. In *UAI 24* (p. 511-518).

Vincent, R., Pineau, J., Guzman, P. de, & Avoli, M. (2007). Recurrent boosting for classification of natural and synthetic time-series data. In *Canadian conference on artificial intelligence (canai)* (pp. 192–293). (Publication in refereed conference proceedings. Co-authors: Philip de Guzman and Massimo Avoli (Montreal Neurological Institute))