Building Bayesian Influence Ontologies
Annotated Bibliography

Tamlin Love
1438243

March 7, 2019

References

[Ajoodha and Rosman 2017] Ritesh Ajoodha and Benjamin Rosman. Tracking influence between naïve bayes models using score-based structure learning. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*. IEEE, November 2017.

Aim: To present a method that learns the high-level influence structure present between a set of independently learned naïve Bayes models.

Style: Conference paper, theoretical

Cross References: This paper builds on many of the principles outlined in Koller and Friedman [2009b], making heavy use of Bayesian networks. Although the approach outlined in this paper is novel, it is composed of many parts described in detail in Koller and Friedman [2009a], such as the score-based structure learning approach and the greedy hill-climbing search. The Bayesian information criterion used in this paper was first developed in Schwarz [1978]. Although using different approaches, the idea of tracking influence presented here is somewhat related to the idea of mapping concepts via probabilistic semantic linkage presented in Pan *et al.* [2005]. The work done in this paper is directly extended to stochastic processes in Ajoodha and Rosman [2018].

Summary: This paper presents an algorithm for learning the influence structure between naïve Bayes models (NBMs). The algorithm achieves this by first learning a set of independent NBMs. It then computes a score used to evaluate the fitness of the network. This approach makes use of the Bayesian information criterion (BIC) for scoring, which provides an acceptable trade-off between model complexity and data fitting. The algorithm then refines the model given the new influence structure using expectation maximisation. After this, the candidate network is subjected to a graph operation (edge addition, reversal or deletion) chosen to optimally improve the network's score. This is achieved using a greedy hill-climbing heuristic, which guarantees monotonically improving score between iterations. Finally, these steps are repeated until an optimum is found.
The result is a method which, in the authors' tests achieved 60-82% accuracy when

compared to the ground truth structure. Additionally, the method outperformed the random structure and the structure with no conditional independence assertions, and tended towards the true structure as the number of samples increased.

[Ajoodha and Rosman 2018] Ritesh Ajoodha and Benjamin Rosman. Learning the influence structure between partially observed stochastic processes using iot sensor data. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Publications, 2018.

Aim: To present an algorithm for learning the influence structure between a set of stochastic processes represented as hidden Markov models (HMMs).

Style: Conference paper, theoretical

Cross References: This paper directly extends the work in Ajoodha and Rosman [2017] to stochastic processes, using hidden Markov models rather than naïve Bayes models. The basic concepts used in this paper, those relating to Bayesian networks, are discussed in Koller and Friedman [2009b], while concepts relating to structure learning, such as the greedy hill-climbing heuristic, are discussed in Koller and Friedman [2009a]. The Bayesian information criterion, one of two scoring criteria used in the paper, was first derived in Schwarz [1978].

Summary: This paper presents a method, referred to as the Greedy structure search (GESS), for recovering the delayed influence structure between a set of HMMs. It does so by first learning each HMM independently using partially observed Internet-of-Things (IoT) data. It then sets the independence assumptions between the models and uses expectation maximisation to learn the associated influence network. The algorithm then evaluates the candidate network's score. The authors empirically test the algorithm using both the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for delayed dynamic influence networks (DDINs). Then the algorithm applies the graph operator (edge addition, deletion or reversal) to result in the best improvement of the network's score with respect to the data. This is done using greedy hill-climbing, which works by applying a graph operation which increases the score, until no such changes can be made. The above steps are repeated until no improvement can be made to the score or the algorithm exceeds the maximum number of iterations.
In the authors' tests, the DDINs produced by the GESS algorithm with the aforementioned scoring criteria more closely recovered the ground truth structure than the tree structures and no structure for a large number of observations. However, for fewer (less than 200) observations, the tree structures and no structure performed better than the GESS-produced structures in this regard.

[Koller and Friedman 2009a] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques - Chapter 18: Structure Learning in Bayesian Networks.* MIT Press, 2009.

Aim: To present and discuss approaches to solving the problem of learning the structure of a Bayesian network from data; namely the constraint-based and score-based techniques, and to introduce several scores used to evaluate a network; namely the maximum likelihood estimate (MLE) and the Bayesian score.

Style: Textbook chapter, theoretical

Cross References: This chapter discusses and builds upon concepts outlined in Koller and Friedman [2009b]. The score-based structure learning using the greedy hill-climbing heuristic presented here is used directly in both Ajoodha and Rosman [2017] and Ajoodha and Rosman [2018]. Among the concepts compiled in this chapter, that of the Bayesian information criterion (which sees use in Ajoodha and Rosman [2017] and Ajoodha and Rosman [2018]) was first developed in Schwarz [1978].

Summary: The chapter introduces the concept of constraint-based structure learning as a method in which dependencies between variables are first tested for and then, based on these dependencies, a network is constructed. However, as the authors discuss, failure in individual independence queries can lead to a network which poorly matches the data.

The authors then discuss score-based structure learning, in which entire networks are chosen and then evaluated based on some score. The chapter discusses such scores at length. The authors present the maximum likelihood estimate, which maximises the probability of the data given the graph and its parameters. The limitations of the MLE are discussed, namely that it always prefers more connected networks and is thus prone to overfitting. The chapter then presents the Bayesian score, which is derived from Bayes rule. The Bayesian score balances model complexity with model fit, preferring complex structures only when more data is available. An approximation for the Bayesian score, the Bayesian information criterion, is presented and is shown to be consistent and to satisfy score equivalence.

Finally, structure search is discussed, which can be divided into the search space and the search procedure. The authors present the search space as the set of candidate graphs connected by possible operations between them. These operations include edge addition, deletion and reversal. Possible search procedures are then discussed. The greedy hill-climbing algorithm is presented, but is shown to be susceptible to local maxima and plateaus between I-equivalent graphs. Methods such as basin flooding, tabu search and random restarts are discussed as possible improvements to the heuristic.

[Koller and Friedman 2009b] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques - Chapter 3: The Bayesian Network Representation.* MIT Press, 2009.

Aim: To present the notion of a Bayesian network (BN), to prove some fundamental properties of BNs, to present the notion of I-equivalence and show that a partially directed acyclic graph (PDAG) can be used to represent all members of an I equivalence class, and to present an algorithm for constructing such a PDAG.

Style: Textbook chapter, theoretical

Cross References: The concepts compiled in this chapter, namely that of a Bayesian network as well as related concepts such as perfect maps and I-equivalence, are built upon in Ajoodha and Rosman [2017], Ajoodha and Rosman [2018] and

Pan *et al.* [2005], and are discussed further and built upon in Koller and Friedman [2009a].

Summary: This chapter presents the concept of a Bayesian network and shows that it can be used to reduce the number of parameters needed to represent a joint distribution. The book provides two definitions of a BN, one as a data structure for compact representation of a joint distribution, and the other as a representation of the set of conditional independence assumptions that hold for such a distribution, and then shows that these definitions are in fact equivalent.
The authors then present the definition of I-equivalence: that two graphs belong to the same I-equivalence class if and only if they represent the same set of independence assumptions. The authors show that a PDAG can be used to represent an I-equivalence class, in which all the undirected edges of the graph can be oriented in any way to produce a graph that belongs to the class. The authors also provide a definition for an immorality between three variables.
Finally, the chapter provides a set of algorithms used to construct a PDAG for a given set of random variables and a distribution over said variables. It consists of an algorithm which constructs an undirected skeleton of the final graph, an algorithm which identifies the immoralities in the class and applies them to the skeleton, and finally an algorithm which applies the previous two algorithms and further directs any edges which could result in the creation of new immoralities or of cycles.

[Pan *et al.* 2005] Rong Pan, Zhongli Ding, Yang Yu, and Yun Peng. A Bayesian network approach to ontology mapping. In *The Semantic Web – ISWC 2005*, pages 563–577. Springer Berlin Heidelberg, 2005.

Aim: To present an approach to automatically mapping concepts between two ontologies via two Bayesian networks using the BayesOWL framework for the semantic web.

Style: Conference paper, theoretical

Cross References: This paper uses the Bayesian network representation discussed in Koller and Friedman [2009b] as a means of mapping concepts between ontologies. Although it makes use of different methods (Jeffrey's rule and IPFP), the approach here is quite similar in concept to the influence-tracking developed in Ajoodha and Rosman [2017] and Ajoodha and Rosman [2018].

Summary: This paper presents a framework for mapping concepts between ontologies using an ”m to n” probabilistic mapping rather than a simple ”1 to 1” mapping. This framework consists of three parts: a learner module, a BayesOWL module and concept mapping module. The learner module is responsible for learning the prior, conditional and joint distributions over concepts in two ontologies. It does this by using text classification to associate concepts with sample documents. In order to correctly label the sample documents, the concept, along with all of its ancestors in the ontology, is searched using a search engine and is associated to any documents returned by the search engine.
The BayesOWL module is responsible for translating each ontology to a Bayesian

network. It does so by translating each class into a node and each predicate relation between two classes into an arc, from superclass to subclass, between the corresponding nodes. The module also creates a set of control nodes to represent logical relations between concepts in the original ontology. The authors also present an algorithm named D-IPFP, which extends the iterative proportional fitting procedure (IPFP) and is used to construct the conditional probability table of each regular node given the set of all control nodes.

Finally, the mapping module uses evidential reasoning across the two networks and the learned similarities calculated earlier to discover mappings. The authors present the notion of pair-wise probabilistic semantic linkage and show that it can be thought of as two subsequent applications of Jeffrey's rule. The authors go on to present a method to map one concept in one ontology to many concepts in another, using a combination of Jeffrey's rule and IPFP. The paper then briefly explores the notion of reducing the number of linkages between variables while still preserving the probability constraints of the system so as to improve the performance of the already computationally expensive IPFP algorithm.

The authors show in their experiments that the framework can successfully map semantically identical concepts and can detect overlap between related but not identical concepts.

[Schwarz 1978] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 03 1978.

Aim: To present a criterion by which models can be selected from a finite set of possible models.

Style: Journal article, theoretical

Cross References: The criterion presented by this paper, now known as the Bayesian information criterion, is used heavily in both Ajoodha and Rosman [2017] and Ajoodha and Rosman [2018], and is discussed in detail in Koller and Friedman [2009a]. In turn, this paper is heavily reliant on the concept of maximum likelihood discussed in Koller and Friedman [2009a].

Summary: This paper proposes a criterion for model selection which balances fit to data with model complexity. The procedure is given as the selection of the model which maximises

$$logM_j(X_1,...,X_n) - \frac{logn}{2}k_j$$

where M_j is the maximum likelihood estimate for the jth model, n is the number of observations X_i and k_j is the dimension of the jth model. The author then derives the proposed criterion as an asymptotic approximation of the Bayesian score in the case of independent observations of identical distribution coming from a distribution with density of the form

$$f(x,\theta) = exp(\theta \cdot y(x) - b(\theta))$$

where y is the sufficient statistic, and where it is also assumed that the penalty for guessing an incorrect model is fixed.

This derivation is achieved by first showing that the Bayes solution to the problem of model selection is equivalent to selecting the j that maximizes

$$S(Y,n,j) = log \int \alpha_j exp((Y \cdot \theta - b(\theta))n) d\mu_j(\theta)$$

where α_j is the probability of the jth model being true, Y is the mean of y, and μ_j is the conditional probability of θ given the jth model. The author then shows that the proposed criterion is derived from $S(Y,n,j)$ as $n \to \infty$.

Finally, it is noted that the resulting criterion is merely a modification of the Akaike Information Criterion (AIC) in which the dimensionality term is multiplied by a factor of $\frac{logn}{2}$. The author also notes that the criterion has greater bias towards lower-dimensional models than the AIC and concludes that the AIC cannot be asymptotically optimal given the earlier assumptions.