A Transfer Learning Approach for Learning Temporal Nodes Bayesian Networks

Lindsey Fiedler Cameras, L. Enrique Sucar and **Eduardo F. Morales**
Instituto Nacional de Astrofísica, Óptica y Electrónica
Tonanzintla, Pueba, Mexico
{lfiedlerc,esucar,emorales}@inaoep.mx

Abstract

Situations where there is insufficient information to learn from often arise, and the process to recollect data can be expensive or in some cases take too long resulting in outdated models. Transfer learning strategies have proven to be a powerful technique to learn models from several sources when a single source does not provide enough information. In this work we present a methodology to learn a Temporal Nodes Bayesian Network by transferring knowledge from several different but related domains. Experiments based on a reference network show promising results, supporting our claim that transfer learning is a viable strategy to learn these models when scarce data is available.

1 Introduction

When representing domains with uncertain information probabilistic graphical models (PGMs) are a popular choice. However, when they are learned with little data they suffer the effects of over fitting, leading to an unreliable model. Transfer learning offers a solution to the problem of having scarce data by reusing knowledge learned previously for other tasks. The idea behind transfer learning is to compensate for the lack of information by applying knowledge from other domains to the learning process of a new task.

Some efforts have been made for using transfer learning strategies to learn PGMs such as Bayesian networks (Luis, Sucar, and Morales 2010), however as far as the authors know, no work exists for dynamic Bayesian models. In this work, we propose a methodology to induce a Temporal Nodes Bayesian Network (TNBN) (Arroyo-Figueroa and Suear 1999) using transfer learning. A TNBN is a type of dynamic PGM which offers a compact graphical representation and allows for the definition of multiple time intervals of variable length.

We propose a methodology for learning the structure, parameters and intervals of a TNBN by using transfer learning. The structure and the parameters are learned by extending the techniques proposed in (Luis, Sucar, and Morales 2010). In order to learn the temporal intervals, we propose a new approach where we incorporate information from other similar domains into the learning process.

To evaluate our methodology we carried out several experiments where we attempted to recover a known model from a small set of records belonging to the target domain and a larger set of records from various related auxiliary domains. Overall, our results showed promise, and suggest that transfer learning is a viable approach for learning TNBNs when there is insufficient data.

2 Related Work

The troubles brought on by having scarce data affect many domains, and several strategies have been proposed to alleviate difficulties. In (Tonda et al. 2012) the authors learned the structure of a Bayesian network when little data was available by using an evolutionary algorithm that operates directly on the graph. Their methodology follows a search-and-score strategy that uses a fitness function based on information entropy to decide between structures.

In (Zhang et al. 2010) the authors used a transfer learning approach to learn the parameters of a model. Their methodology is based on the maximum entropy model which seeks to obtain a model consistent with all the facts, while still being as general as possible. In their work, the authors transfer the learned parameters from an auxiliary domain to a target domain while adjusting the weights of the target instances to obtain the model with the highest accuracy.

In (Luis, Sucar, and Morales 2010) the authors proposed a transfer learning strategy to learn both the structure and the parameters of a Bayesian network. They developed a methodology based on the PC algorithm to induce the structure of a model from several domains, and subsequently they learned the parameters of the model by combining conditional probability tables through aggregation functions.

3 Preliminaries

In this section we provide an overview of Temporal Nodes Bayesian Networks and transfer learning.

3.1 Temporal Nodes Bayesian Network

A TNBN is a type of PGM in which each node represents an event, and each edge in the graphical structure represents a temporal probabilistic relation. TNBNs are composed by two types of nodes: instantaneous and temporal. Instantaneous nodes model events in which no time delay is seen

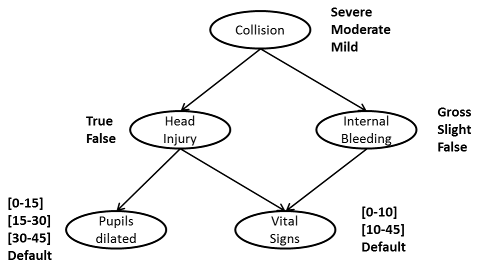


Figure 1: A Temporal Nodes Bayesian Network modeling the event of a collision. Here, "Collision", "Head injury" and "Internal bleeding" are instantaneous nodes, while "Pupils dilated" and "Vital signs" are temporal nodes each with a set of associated intervals and a default value.

before their occurrence, that is, once a parent event takes place the manifestation of the corresponding child event is immediate.

Unlike an instantaneous node, a temporal node models the possible time delays between the occurrence of the cause and the observing of the effect. Each temporal node consists of a set intervals in which an event may happen. A "Default" value indicates that the event does not occur. It is important to point out that all root nodes of a TNBN must be instantaneous. An example of a TNBN that models the event of a collision is provided in Figure 1.

3.2 Transfer Learning

A domain is characterized by two components: a feature space X and a marginal probability distribution $P(X)$ over X. Two domains are different when they have different feature spaces or their probability distributions are not the same.

Given a specific domain, $D = \{X, P(X)\}$, a task is defined by a label space Y and an objective predictive function $f(\cdot)$ to be learned from the data. This function allows us to predict the corresponding label y of a new instance x.

Transfer learning improves the learning of a target predictive function by using knowledge from one or several different but related auxiliary domains or tasks.

4 Methodology

To learn a TNBN three elements must be obtained: 1) the structure, 2) the probability distributions that parametrize the model, and 3) the intervals in which temporal events occur. We propose a transfer learning strategy to learn each component of a TNBN. For the structure and the parameter learning we adopted the method introduced in (Luis, Sucar, and Morales 2010). A new approach was taken to learn the temporal intervals, which implements the ideas behind transfer learning in order to compensate for the small amount of data.

4.1 Structure Learning

In order to learn the graph for the model, we implemented the PC-TL algorithm which is an extension of the PC algorithm (Spirtes, Glymour, and Scheines 2001) that incorporates knowledge transfer from auxiliary domains. PC-TL begins with a fully connected graph and removes edges based on the results of some procedure to determine independence. After obtaining the skeleton, it orients the edges by measuring conditional independence between variable triplets and directing the remaining edges with care to avoid cycles. Its main difference with PC lies in how the independence tests are evaluated, as these are now a linear combination of the results for the tests performed on the target domain and the closest auxiliary domain. A similarity measure based on common dependencies and independencies is defined to determine the closest auxiliary domain from which to transfer. The full algorithm is described in (Luis, Sucar, and Morales 2010).

4.2 Parameter Learning

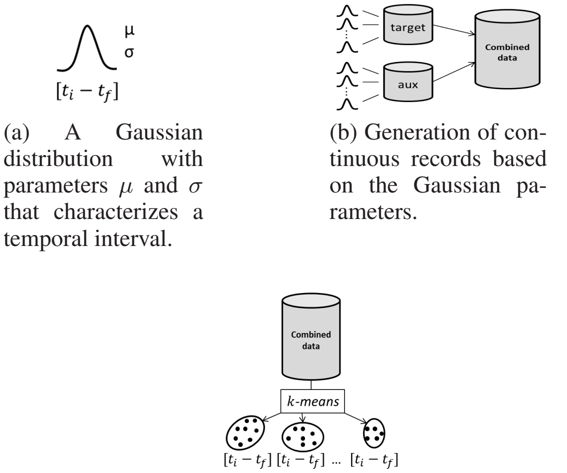
Once a structure is obtained, the probability values that parametrize the model must be learned. The approach we used consists of combining the conditional probability tables (CPTs) from the target task and the auxiliary tasks using linear aggregation.

In order to combine CPTs, the auxiliary tasks must have the same parents as the target task. If this is not the case, transformations to the auxiliary structures must be made. Three situations are considered: 1) the auxiliary structure has more parents, 2) the auxiliary structure has less parents, and 3) a combination of 1 and 2. In the first scenario the additional parents are removed by marginalizing over them. For the second scenario, values for the additional parents are obtained by repeating the values seen in the auxiliary CPT for the common parents. For example, we repeat the values of $P(X|Y)$ for all values of Z on $P(X|Y, Z)$.

4.3 Interval Learning

Since the data for the temporal fields is continuous, before a structure can be learned we must first transform all the continuous information into discrete values. This is analogous to learning the time intervals for each temporal node. We based our method on the strategy used in (Hernandez-Leal, Sucar, and Gonzalez 2011) to learn the initial time intervals. The authors of that work used k-means to obtain a set of clusters, where each cluster corresponds to a temporal interval. The number of intervals a temporal node has is therefore defined by the parameter k.

To incorporate knowledge from auxiliary domains we transferred temporal information to the data on which k-means will be applied. However, because the data for the temporal fields in the target domain is continuous and the data in the auxiliary domains is already discrete in the form of intervals, information cannot be directly transferred. Instead, continuous records are generated from the intervals of the temporal nodes from the auxiliary domains by assuming that each one is characterized by a Gaussian distribution where μ is the middle point of the interval and σ is the distance from that point to either of the extremes. Based on these parameters, continuous values that follow this Gaussian distribution can be generated. These values are constricted to the range $\mu \pm \sigma$. With the discrete records in continuous form, we can now transfer auxiliary knowledge.



(a) A Gaussian distribution with parameters μ and σ that characterizes a temporal interval.

(b) Generation of continuous records based on the Gaussian parameters.

(c) New intervals learned with the transferred knowledge.

Figure 2: The basic procedure followed for interval learning with knowledge transfer.

Aside from how to transfer, we must also decide how much to transfer. To avoid the auxiliary information from overwhelming the target data, we control the proportion of transferred data in the total amount of records. The number of records each auxiliary domain transfers is decided by how similar it is to the target domain. We used the global similarity metric defined by (Luis, Sucar, and Morales 2010) to determine the strength of the relation between two domains.

In addition to controlling the proportion of auxiliary records, we can also increase the amount of target records by applying the same process. In order to obtain initial time intervals from which continuous values can be generated, we apply k-means to the target data for the temporal field of interest. We then repeat the process applied on the auxiliary domains to increase the proportion of target information in the total amount of records. Finally, we apply k-means one last time to this combined set of target and auxiliary records to obtain k intervals learned with knowledge transfer. The basic steps for this procedure are illustrated in Figure 2.

5 Experiments

In the following experiments we learn a simple Temporal Nodes Bayesian Network (TNBN) from scarce data and a set of different but similar auxiliary domains each of which has sufficient data to learn a reliable model. The TNBN to be learnt presents the possible consequences of a collision as modeled in (Hanks, Madigan, and Gavrin 1995). Figure 1 displays this TNBN which is used as a gold standard with which we compare our resulting models.

Three auxiliary models[1] were created by using the original TNBN as a base model, and subsequently altering the structure by adding or removing links. The conditional probability tables were also changed by adding Gaussian noise. The data for the experiments was generated with the Elvira System (Consortium and others 2002).

We performed two sets of experiments. Our first experiments evaluate how the size of the target data set affects the produced models, while our second experiments aim to measure the impact the amount of auxiliary records have on the learned models. To assess the quality of the resulting models we used different metrics to evaluate the structure, the parameters and the time intervals learned with our algorithm. For the structure we measured the edit distance of the learned network with our original target model. The quality of the parameters was evaluated by calculating the mean square error (MSE) between the resulting parameters and the original values. We used the relative temporal error (RTE) to assess the quality of the time intervals. This metric evaluates how far the real events are from the learned intervals. It is defined as the difference between the time the event occurred and the middle point of the predicted time interval divided by the range of the temporal node. Finally, the relative Brier score (RBS) was used to evaluate the predictive accuracy of the resulting models. The relative Brier Score expressed as a percentage is defined as:

$$RBS = \left(1 - \frac{1}{n}\sum_{i=1}^{n}(1 - P_i)^2\right) \times 100$$

where n is the number of selected nodes to infer, and P_i is the marginal posterior probability of the correct value for each node given the evidence. To calculate the RBS we randomly selected a set of nodes to be "observed" and then inferred the remaining hidden nodes. We used a 5-fold cross validation process with disjoint sets to gain a better estimate of how the resulting models behave.

For our experiments we used a confidence value of 0.01 as a threshold for the conditional independence tests and established that the target data would account for 40% of the final values of the learned parameters. We also assume we know the total number of intervals each temporal node has, eliminating the need to discover this parameter. Since these are only initial experiments, we designated the temporal nodes of the auxiliary models to have the same temporal intervals as the target model.

5.1 Varying the Amount of Target Records

In these experiments we vary the number of target records used to learn the model. We generated a total of 2200 target records and we randomly selected a subset of these. To assess how the amount of target records affects the learned model, we tested with subset sizes of 44, 200 and 440 records while leaving the amount of auxiliary records fixed.

To ensure that each auxiliary domain has sufficient records to learn a reliable TNBN we defined that a minimum of 10 records were required to learn each independent parameter of the network. For example, if a model requires 35 probabilistic values to fully specify all the probability tables that parametrize the model, then a minimum of 350 records were generated to learn a reliable model.

Table 1 shows the most significant results for this experiment. As is expected, better results are achieved as the

[1] All models and experimental results can be seen at http://ccc.inaoep.mx/˜lfiedlerc/tnbn-tl/

	44 records	200 records	440 records
Avg. RBS	78.67%	80.97%	81.96%
Avg. RTE	8.71	8.41	8.01
Avg. MSE	0.0127	0.0091	0.0067
Avg. Edit distance	6.0	4.2	3.0

Table 1: Results of learning a TNBN with transfer learning using three auxiliary domains with 350, 470 and 600 records each and varying the number of target records.

amount of target records grows. The models built using 44 target records showed lower values for the edit distance and MSE, providing an explanation for the lower RBS as they impact the results of inference. The temporal intervals for the model were also affected by the smaller amount of target data, and in general we observed a higher RTE when less target data was available.

We carried out this same experiment two more times, each time incrementing the sizes of the auxiliary data sets. While we do not provide these results explicitly, we note that in cases where only 44 target records were used all metrics see an improvement, with the RBS surpassing values of 80%. This proves that by incrementing the amount of auxiliary data we can indeed compensate for the lack of target data.

5.2 Varying the Amount of Auxiliary Records

To observe the effect the number of auxiliary records has in the resulting models, we conducted another experiment where we fixed the amount of target records used for learning to 200 and varied the amount of auxiliary records. We began by setting the three auxiliary data sets to 350, 470 and 600 records each and then increased the amount of records to twice as many each, and finally five times as many. The results of this experiment are shown in Table 2.

	Aux × 1	Aux × 2	Aux × 5
Avg. RBS	80.17%	82.66%	82.47%
Avg. RTE	8.17	8.04	8.18
Avg. MSE	0.0097	0.0065	0.0058
Avg. Edit distance	5.0	1.6	3.0

Table 2: Results of learning a TNBN with transfer learning using 200 target records and varying the number of auxiliary records.

Our results show that by doubling the amount of auxiliary data, we can improve the predictive accuracy of the model. However, when the data sets were increased by a factor of 5, a decline in almost all metrics was observed. This phenomenon can be explained as the result of the decrease in contribution of the target data as the amount of auxiliary data increases, such that a point is reached where the target data becomes completely overwhelmed.

6 Conclusions

In this paper we presented a transfer learning strategy to induce a Temporal Nodes Bayesian Network when little data is available. For the learning of the structure and the parameters we extended the strategy defined in (Luis, Sucar, and Morales 2010) for Bayesian networks. For the temporal intervals, we proposed a new methodology where the intervals for each temporal node are learned by incorporating in the learning process, information about the distributions followed by the auxiliary temporal intervals.

Several experiments were carried out to test our methodology, and overall we obtained promising results, achieving a predictive accuracy of over 80% for almost all cases. We note that the relative temporal error is maintained low, proving that the learned intervals are performing satisfactorily.

As future work, we propose to test the effects that providing a node ordering has on the learned structure. We expect this additional information to improve the structure and as a result the predictive accuracy. We would also like to explore a strategy to learn the number of intervals each temporal node has as this information is not necessarily available.

Acknowledgements

The first author is thankful to CONACyT for the financial support given to her through the scholarship 261257.

References

Arroyo-Figueroa, G., and Suear, L. E. 1999. A temporal bayesian network for diagnosis and prediction. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 13–20. Morgan Kaufmann Publishers Inc.

Consortium, E., et al. 2002. Elvira: An environment for creating and using probabilistic graphical models. In *Proceedings of the first European workshop on probabilistic graphical models*, 222–230.

Hanks, S.; Madigan, D.; and Gavrin, J. 1995. Probabilistic temporal reasoning with endogenous change. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 245–254. Morgan Kaufmann Publishers Inc.

Hernandez-Leal, P.; Sucar, L. E.; and Gonzalez, J. A. 2011. Learning temporal nodes bayesian networks. In *The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24). Palm Beach, Florida, USA*.

Luis, R.; Sucar, L. E.; and Morales, E. F. 2010. Inductive transfer for learning bayesian networks. *Machine learning* 79(1):227–255.

Spirtes, P.; Glymour, C.; and Scheines, R. 2001. *Causation, prediction, and search*, volume 81. MIT press.

Tonda, A.; Lutton, E.; Reuillon, R.; Squillero, G.; and Wuillemin, P.-H. 2012. Bayesian network structure learning from limited datasets through graph evolution. *Genetic Programming* 254–265.

Zhang, Y.; Hu, X.; Mei, C.; and Li, P. 2010. A weighted algorithm of inductive transfer learning based on maximum entropy model. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 6, 2740–2745. IEEE.