ISSN 1870-4069

Toward Knowledge Transfer for Learning Markov Equivalence Classes

Verónica Rodríguez López, Luis Enrique Sucar, Felipe Orihuela Espina

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Mexico

{verorl,esucar,f.orihuela-espina}@inaoep.mx

Abstract. Most algorithms for causal discovery require large sample sizes for finding Markov equivalence classes that include the structure of the true causal probabilistic graphical models. In some situation collecting data could be difficult, especially for learning models that encode the specific causal relations of a particular subject of a population. Although transfer learning techniques have shown to be useful for improving predictive associative models learned with limited datasets, their application in the field of causal discovery has not been sufficiently explored. In this paper, we explore transferring weighted instances of auxiliary datasets for improving Markov equivalence classes learned with otherwise limited datasets. A knowledge transfer algorithm extended from the Greedy equivalence search algorithm that locally selects the instances of the best auxiliary datasets is proposed. Preliminary results using synthetic datasets suggest that our knowledge transfer algorithm outperforms the base algorithm, increasing the adjacency recall from 0.58 ± 0.28 to 0.94 ± 0.13.

Keywords: causal discovery, transfer learning, causal probabilistic graphical models.

1 Introduction

Causal probabilistic graphical models (causal PGMs) are useful tools for encoding causal relations between variables of closed systems and provide information to make predictions under manipulations. From observational data, it is possible discovering Markov Equivalence Classes (MECs) that represent the structure of a set of equivalent causal PGMs with the same joint probability distribution [2].

Learning MECs that include the true causal structure from a limited sample size could be challenging using many existing algorithms, since they find these MECs in the large sample limit [18, 5]. In some situations, it can be difficult collecting data, especially for learning casual PGMs that encode the specific causal relations for a particular member of a population. Transfer learning has shown to be useful for improving models learned with limited datasets, allowing the use of auxiliary data that come from different models with different probability distributions [15].

Many works have explored knowledge transfer for learning PGMs. However, most of these studies have relied on the learning of associative PGMs [10, 12–14]. Limited work [8] has been done on learning causal PGMs from observational data. Although other algorithms have been proposed for learning MECs from multiple datasets, their aim is different of that for knowledge transfer algorithms. These algorithms aim to discover MECs that include the common causal relations in all datasets, assuming that all datasets include a representative number of samples [3, 16, 17].

The knowledge transfer algorithm proposed in [8] is a modification of the PC algorithm that assumes all auxiliary datasets have the same relevance for learning a target MEC, ignoring their differences in probability distributions. Moreover, like other PC-based algorithms, require large sample sizes for the independence conditional tests [5]. Score-based algorithms have shown to be more accurate for learning MECs with small samples than constraint-based algorithms as PC [11]. In this paper, we present a preliminary knowledge transfer algorithm, based on the score-based algorithm, Greedy Equivalence Search [2], for improving MECs learned with limited datasets. We propose locally transferring the instances of the best auxiliary datasets, considering their differences in probability distributions with that of the target dataset.

The paper is organized as follows. In Section 2 concepts related to graphs and the Greedy Equivalence Search algorithm are described. Our knowledge transfer algorithm is presented in Section 3. In Section 4, the experimental results are shown. Finally, the conclusions of this paper are presented in Section 5.

2 Preliminaries

2.1 Graph Concepts

Definition 1. *A **graph** is a pair* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *formed by a set of nodes* $\mathbf{V} = \{V_1, ..., V_N\}$*, and a set of edges* $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$*.*

Two nodes are **adjacent** in a graph $\mathcal{G}$, if there is an edge associating them. When a graph only contains directed edges in the form $(V_1 \rightarrow V_2)$, it is called a **directed graph**. In a directed edge in the form $V_1 \rightarrow V_2$, V_1 is said to be the **parent** of V_2, and V_2, the **child** of V_1. The set of parents of a node V is denoted as $\mathbf{Pa}(V)$.

Definition 2. *Within a graph* $\mathcal{G}$*, a **directed path** between two nodes* V_1 *and* V_k *is a sequence of nodes,* $(V_1, V_2, ..., V_k)$*, starting at* V_1 *and ending at* V_k*, where* $k \geq 2$*, and* $V_i \rightarrow V_{i+1} \in \mathbf{E}$ *for* $i = 1, ..., k-1$*.*

A directed path where the last node coincides with the first one is a **directed cycle**. A directed graph in which there are no directed cycles is called a **directed acyclic graph** (DAG). If an acyclic graph contains directed and undirected edges, it is called a **partially directed graph** (PDAG).

The undirected graph resulting from ignoring the direction of edges in a DAG is the **skeleton** of the DAG.

A **v-structure** in a DAG is an ordered triple of nodes (X, Y, Z), such that, the edges $X \rightarrow Y$ and $Y \leftarrow Z$ are in the DAG, and there is no edge between the nodes X, Z [2].

Definition 3. *A **Markov equivalence class** is a set of directed acyclic graphs that have the same skeletons and the same v-structures [6].*

2.2 Greedy Equivalence Search Algorithm

Greedy Equivalence Search (GES) [2] is a score-based algorithm for heuristically searching the best Markov equivalence class that represents a set of equivalent DAGs including a true causal probabilistic graphical model. Given a dataset $\mathbf{D} = \{d_1, ..., d_m\}$ containing m instances, where each d_i represent an assignment of value to each variable of a set $\mathbf{X} = \{X_1, X_2, X_3, ..., X_n\}$, the best MEC $\mathcal{G}^* = (\mathbf{X}, \mathbf{E})$ is found by maximizing a scoring function such that:

$$\mathcal{G}^* = \arg\max_{\mathcal{G} \in \mathcal{G}_C} score(\mathcal{G}, \mathbf{D}), \tag{1}$$

where $score(\mathcal{G}, \mathbf{D})$ is a scoring function that measures the adjustment of $\mathbf{D}$ with a candidate MEC $\mathcal{G}$, and $\mathcal{G}_C$ is the set of all MECs defined over $\mathbf{X}$.

In the GES algorithm, Bayesian Dirichlet equivalent and Uniform (BDeU) score function is used for learning MECs defined over discrete variables with complete datasets $\mathbf{D}$ (without missing values). BDeU score is a descomposable function that can be expressed as the product of local functions $BDeU(X_i, \mathbf{Pa}(X_i), \mathbf{D})$ that only depends of a node $X_i \in \mathbf{X}$ and their parents $\mathbf{Pa}(X_i)$ as follows [7]:

$$BDeU(\mathcal{G}, \mathbf{D}) = \prod_{i=1}^{n} \{BDeU(X_i, \mathbf{Pa}(X_i), \mathbf{D})\}, \tag{2}$$

$$BDeU(X_i, \mathbf{Pa}(X_i), \mathbf{D}) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{3}$$

where n is the number of nodes in $\mathcal{G}$, q_i is the number of values of $\mathbf{Pa}(X_i)$, r_i is the number of values of X_i, N_{ijk} is the number of cases in which $X_i = k$ and its parents $\mathbf{pa}(X_i = k) = j$, $N_{ij} = \sum_k N_{ijk}$, and $\alpha_{ijk} = \frac{1}{r_i q_i}$ is a Dirichlet prior parameter with $\alpha_{ij} = \sum_k \alpha ijk$.

BDeU score assigns the same value to all equivalent DAGs in the same MEC. It is used in each iteration of the GES algorithm for evaluating the improvement of the score when an edge is added or deleted. In the first stage of GES, starting with a empty graph, the scoring function is used for heuristically searching the edges that could be added of a MEC. And in the second stage, for searching the edges that could be removed of a MEC.

3 Instance-Based Transfer Learning GES

Our proposed preliminary algorithm, denominated as Knowledge Transfer Learning with Weighted instances GES (KTL-WeGES), is an extension of the Greedy Equivalence Search (GES) algorithm that using the instances of two auxiliary observational datasets tries to improve the skeleton identification of Markov equivalence class (MEC) learned with limited dataset.

Under the assumptions of causal sufficiency and faithfulness conditions, the best target MEC $\mathcal{G}_T^*$ is found by maximizing a scoring function that combines the instances of target $\mathbf{D}_T$ and auxiliary $\mathbf{D}_S$ datasets:

$$\mathcal{G}_T^* = \arg \max_{\mathcal{G}_T \in \mathcal{G}_C} score(\mathcal{G}_T, \mathbf{D}_T, \mathbf{D}_S). \tag{4}$$

For combining the instances of target and auxiliary datasets, local knowledge transfer of the auxiliary datasets is explored. In this local knowledge transfer, weighted instances of the auxiliary datasets are used for finding the best local structure for a target MEC composed by a node $X_i \in \mathbf{X}$ with their parents $\mathbf{Pa}_T(X_i)$. The local BDeU score defined in the Equation 3 is used for evaluating the adjustment of the combination of weighted instances of the auxiliary $\mathbf{D}_S$ and target $\mathbf{D}_T$ datasets, with a candidate local structure for a target MEC. In this equation, N_{ijk} counting the combination of auxiliary and target instances as follows:

$$N_{ijk} = (N_{ijk})_T + W_i(N_{ijk})_S, \tag{5}$$

where $(N_{ijk})_T$ represents the number of cases in $\mathbf{D}_T$ in which $X_i = k$ and its parents $\mathbf{pa}_T(X_i = k) = j$, and $(N_{ijk})_S$, the number of cases in $\mathbf{D}_S$ in which $X_i = k$ and its parents $\mathbf{pa}_T(X_i = k) = j$. W_i encode the relatedness of the auxiliar dataset with the candidate local structure for a target MEC.

In the estimation of this relatedness, differences in the conditional probability distribution of X_i and its parents $\mathbf{Pa}_T(X_i)$, between the target dataset $P_T(X_i|\mathbf{Pa}_T(Xi))$ and the auxiliary dataset $P_S(X_i|\mathbf{Pa}_T(Xi))$, are considered. The difference between these distributions is evaluated with the Kullback-Leibler divergence D_{KLD} [1] as follows:

$$D_{KLD}(P_T(X_i|\mathbf{Pa}_T(X_i)), P_S(X_i|\mathbf{Pa}_T(X_i))) \approx \sum_{x_i, \mathbf{pa}_T(x_i)} log\left(\frac{P_T(x_i|\mathbf{pa}_T(x_i))}{P_S(x_i|\mathbf{pa}_T(x_i))}\right). \tag{6}$$

Using this difference, W_i is estimated by:

$$W_i = 2^{-|D_{KLD}(P_T(X_i|\mathbf{Pa}_T(X_i)), P_S(X_i|\mathbf{Pa}_T(X_i)))|}. \tag{7}$$

With this function, when the difference between target and auxiliary datasets increases, it is penalized with weights nearly to zero; and it assigns weights nearly to one, to small differences lower to one.

4 Experiment and Results

4.1 Generation of Synthetic Datasets

Synthetic datasets are generated from ground truth Bayesian networks which are BN with known structure and parameters. Target and auxiliary datasets are generated in the following form [10]. Target dataset is sampled from the ground truth BN, and auxiliary datasets, from related BNs. Related BNs are generated modifying in certain percent ($pMod$) the edges of the ground truth models, adding $pMod$ edges, followed by deleting edges in the same $pMod$ percent. Increasing the $pMod$, we generate BN less related to the ground truth model. From each related BN are estimated its parameters using a dataset sampled from the ground truth BN. Each auxiliary dataset is sampled from its corresponding related BN using forward sampling, in which the values of each variable X_i are sampled in ancestral order (parents before their children), in such form that its values x_i are drawn from $P(x_i|\mathbf{pa}(x_i))$.

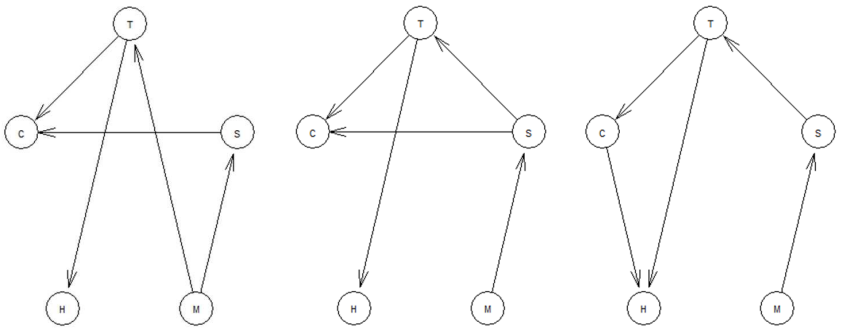
4.2 Experimental Design

In this experiment, we hypothesized that the KTL-WeGES algorithm outperform the GES algorithm. The performance of the KTL-WeGES algorithm was evaluated in its ability for finding the skeleton of the ground truth models. In the evaluation, the Coma [4] and Asia [9] binary BNs with five and eight nodes, respectively, were used as ground truth models. The edges of the original BNs were modified in 10% and 40%, for generating the two related BNs. Considering extreme cases of relatedness (most and least related) were selected these parameters. Coma and Asia BNs and their corresponding related BNs are presented in Figures 1 and 2, respectively. Datasets with 1600 and 12800 samples for Coma and Asia were used for estimating the parameters of related BNs.

Taking into account that after modifying the ground truth BNs would increase the number of parents for some nodes. The sample size was estimated using $samplesize = 100(2^k)$, considering that a node in a related BN may have at most $k = n - 1$ parents (where n is the number of nodes in the BN). For each auxiliar dataset, 1600 samples from related BNs of Coma and 12800 samples from related BNs of Asia (using the same formula for the parameters estimation), were obtained. Ten datasets varying the sample size were obtained for the target domain. For Coma, the set of target datasets includes datasets with size $\{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$, and for Asia, with $\{80, 160, 240, 320, 400, 480, 560, 640, 720, 800\}$. Ten runs of this scenario were used to evaluate the algorithms.

The models obtained by the algorithms were evaluated using normalized structural Hamming distance (NSHD), adjacency precision (TPR), and adjacency recall (TDR). Normalized structural Hamming distance is the minimum number of edge insertions, deletions, and changes needed to transform a model into another. Adjacency precision is the ratio $TP/(TP + FP)$, and the ratio $TP/(TP + FN)$ is the adjacency recall.

Where TP is the number of adjacencies that are in common in the estimated model and ground truth model without considering the edge orientation; FP is the number of adjacencies that are present in the estimated model but not in the ground truth model; and FN is the number of adjacencies that are present in the ground truth model but not in the estimated model [17].



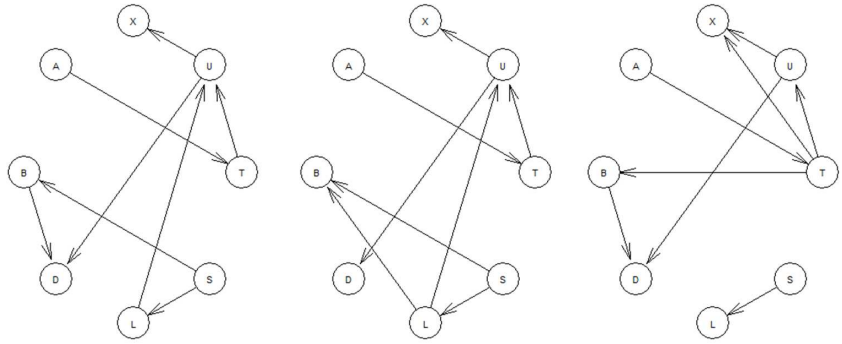
(a) (b) (c)

Fig. 1. (a) Coma and its related BNs created by modifying the edges in (b) 10% and (c) 40%.

4.3 Results

The experimental results are summarized in Tables 1 and 2 for Coma and Asia, respectively. In these tables, the averages for each metric, over the ten test target datasets and all experimental runs, obtained by transferring instances from the most related, the least related, and both auxiliary datasets, are presented.

The results show that KTL-WeGES seems to improve the skeleton identification of the ground truth models with respect to GES. In the case of Coma, considering the results for NSHD (the best NSHD is obtained when it is zero), KTL-WeGES seems to decrease the differences between the skeleton of the true and that one of the estimated model. The results for this model also show that, although the performance of the TPR decrease, KTL-WeGES are discovering more number of edges, increasing the TDR. The results for Asia show an improvement in the TPR and TDR rates.



(a) (b) (c)

Fig. 2. (a) Asia and its related BNs created by modifying the edges in (b) 10% and (c) 40%.

Table 1. Averages ± standard deviations of TPR, TDR, and NSHD for Coma.

Method	**TPR**	**TDR**	**NSHD**
GES	0.90 ± 0.13	0.59 ± 0.25	0.54 ± 0.25
KTL-WeGES (most related)	0.86 ± 0.11	0.94 ± 0.10	0.40 ± 0.31
KTL-WeGES (least related)	0.85 ± 0.09	0.98 ± 0.06	0.36 ± 0.28
KTL-WeGES (both auxiliar datasets)	0.84 ± 0.09	0.96 ± 0.08	0.38 ± 0.27

They also show that the differences between the skeleton of the true and that one of the estimated model increase, which indicate that the estimated model has more edges than the true model (spurious edges).

From the results, it also can be observed that KTL-WeGES with all strategies (transferring from all datasets, the best and least related auxiliary dataset) improves the TDR, being better transferring from the least related auxiliary dataset. Although, the NSHD and TPR results show that KTL-WeGES discovers more spurious edges when the number of nodes increases. It indicates that the scoring function prefers dense graphs, and hence KTL-WeGES has problems for deleting edges.

Regarding execution time for learning a single MEC on average, KTL-WeGES takes 0.46 and 10.78 seconds for learning models of Coma and Asia, respectively, with a 1.8 GHz Intel Core i7 processor with 8 GB RAM, using Matlab 2019a.

Table 2. Averages ± standard deviations of TPR, TDR, and NSHD for Asia.

Method	TPR	TDR	NSHD
GES	0.71 ± 0.27	0.58 ± 0.31	0.98 ± 0.44
KTL-WeGES (most related)	0.95 ± 0.07	0.90 ± 0.19	1.94 ± 0.33
KTL-WeGES (least related)	0.97 ± 0.05	0.90 ± 0.19	1.98 ± 0.35
KTL-WeGES (both auxiliar datasets)	0.97 ± 0.05	0.90 ± 0.19	1.99 ± 0.37

5 Conclusions

A preliminary instance-based transfer algorithm for improving Markov equivalence classes learned with limited datasets was presented. Our algorithm locally selects the instances from the two auxiliary datasets for searching the best set of parents of each node in a target MEC.

Experimental results show that our algorithm outperforms the GES algorithm in the skeleton identification for MECs, transferring weighted instances from the most related, the least related and both auxiliary datasets. Preliminary results suggest that our algorithm seems to be promising for discovering MECs.

As future work, we consider extending the local knowledge transfer of the weighted-instances for more than two auxiliary datasets and also analyzing other scoring functions and score-based algorithms that have shown better performance deleting false edges. Also, it is contemplated improving the algorithm for discovering the v-structures of MECs.

Acknowledgements. We acknowledge the support from the Secretaría de Educación, Ciencia, Tecnología e Innovación de la Ciudad de México SECITI/042/2018 (INGER-DI-CRECITES-001-2018) "Red Colaborativa de Investigación Traslacional para el Envejecimiento Saludable de la Ciudad de México" (RECITES). The first author acknowledges scholarship support from CONACYT as a PhD student.

References

1. Campos Ibáñez, L.M.: A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. Journal of Machine Learning Research 7(Oct), 2149–2187 (2006)
2. Chickering, D.M.: Optimal structure identification with greedy search. Journal of Machine Learning Research 3(Nov), 507–554 (2002)
3. Claassen, T., Heskes, T.: Causal discovery in multiple models from different experiments. In: Advances in Neural Information Processing Systems. pp. 415–423 (2010)

4. Cooper, G.F.: Nestor: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. Tech. rep., Stanford University CA, Dept of Computer Science (1984)
5. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in Genetics 10, 1–15 (2019)
6. He, Y., Jia, J., Yu, B.: Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. The Journal of Machine Learning Research 16(1), 2589–2609 (2015)
7. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning 20(3), 197–243 (1995)
8. Jia, H., Wu, Z., Chen, J., Chen, B., Yao, S.: Causal discovery with Bayesian networks inductive transfer. In: International Conference on Knowledge Science, Engineering and Management. pp. 351–361. Springer (2018)
9. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society: Series B (Methodological) 50(2), 157–194 (1988)
10. Luis, R., Sucar, L.E., Morales, E.F.: Inductive transfer for learning Bayesian networks. Machine learning 79(1-2), 227–255 (2010)
11. Malinsky, D., Danks, D.: Causal discovery algorithms: A practical guide. Philosophy Compass 13(1), e12470 (2018)
12. Niculescu-Mizil, A., Caruana, R.: Inductive transfer for Bayesian network structure learning. In: Artificial Intelligence and Statistics. pp. 339–346 (2007)
13. Oates, C.J., Smith, J.Q., Mukherjee, S., Cussens, J.: Exact estimation of multiple directed acyclic graphs. Statistics and Computing 26(4), 797–811 (2016)
14. Oyen, D., Lane, T.: Bayesian discovery of multiple Bayesian networks via transfer learning. In: 2013 IEEE 13th International Conference on Data Mining. pp. 577–586. IEEE (2013)
15. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10), 1345–1359 (2010)
16. Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C.: Six problems for causal inference from fMRI. Neuroimage 49(2), 1545–1558 (2010)
17. Tillman, R., Spirtes, P.: Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. pp. 3–15 (2011)
18. Zhang, K., Schölkopf, B., Spirtes, P., Glymour, C.: Learning causality and causality-related learning: some recent progress. National Science Review 5(1), 26–29 (2018)