ISSN 1870-4069

# Data-Driven Bayesian Network Learning: Towards a Bi-Objective Approach to Address the Bias-Variance Decomposition

Vicente Josué Aguilera Rueda, Nicandro Cruz Ramírez, Efrén Mezura Montes

Universidad Veracruzana,
Centro de Investigación en Inteligencia Artificial (CIIA),
Mexico

{vaguilera,ncruz,emezura}@uv.mx

**Abstract.** We present a novel bi-objective approach to address the data-driven learning problem of Bayesian networks. Both the log-likelihood and the complexity of each candidate Bayesian network are considered as objectives to be optimized by our proposed algorithm named Non-dominated Sorting Genetic Algorithm for learning Bayesian networks (NS2BN) which is based on the well-known NSGA-II algorithm. The core idea is to reduce the implicit selection bias-variance decomposition while identifying a set of competitive models using both objectives. Numerical results suggest that, in stark contrast to the single-objective approach, our bi-objective approach is useful to find competitive Bayesian networks with a balanced trade-off between accuracy and complexity.

**Keywords:** Bayesian networks, bias-variance, NSGA-II.

## 1 Introduction

A way to build a Bayesian Network (BN) is adopting a data-driven inductive approach; in this case, the learning task is framed as a combinatorial optimization problem with two components: a metric to assess the quality of each BN candidate, and a search procedure to move through the space of candidate networks.

In data-driven BN learning, it is common to implement metrics in the form of a penalized log-likelihood (LL) function, as minimum description length (MDL). While adding an edge to a BN never decreases the likelihood –and hence irrelevant arcs should be discarded– adding extra arcs leads to two main problems: the overfitting problem and densely connected networks. To avoid complex networks, a penalty term is used. However, complex networks may have a low LL score value but overfit the model, while a high penalty term may incur in underfitting. Thus, it is desirable to have networks with a suitable balance between the goodness of fit (accuracy) and complexity.

Some researchers point out that the trade-off between accuracy and complexity should be featured as a bi-objective problem [5, 11] however, to the best

of our knowledge, estimation of those values has not been previously used in a bi-objective approach for model selection in BN. Our work addresses this combination. The remainder of this paper is structured as follows: Section 2 describes related work. In Section 3, background about BN, MDL and bi-objective optimization problem are presented. In Section 4, our proposed algorithm is described. Section 5 presents the experiments and discusses the results. The concluding section summarizes the findings and gives an account for future work.

## 2 Related Work

There exist two main approaches to the use of crude MDL to learn BN: to find the true model (that has given rise to the data), known as the *gold-standard* network [8] and to find a model with a good trade-off between the accuracy and complexity [6]. Accuracy in this context refers to the computation of the log-likelihood of the data given a BN structure; it should not be confused with classification accuracy (see Equation 1).

Cruz-Ramírez et al. [3], performed an exhaustive experiment with four-node networks. Therefore, eventhough these results show how crude MDL produces well-balanced models in terms of complexity and log-likelihood, those experiments have a limited scope of four-node networks and they left for future work to explore the search procedure.

Previous studies have addressed the BN model selection problem using evolutionary algorithms, for instance, see [2, 13, 10]. However, none of them has tackled the problem in a multi-objective way.

Lastly, the work of Ross and Zuviria [12] uses a multi-objective genetic approach to induce dynamic BNs from data with a trade-off between likelihood and complexity. This work is focused on the modeling of biological phenomena that typically requires low-connectivity networks. However, to the best of our knowledge, this work is the only one with multi-objective criteria learning. Although, is in the context of dynamic BN.

In summary, the learning problem of BN using MDL as a metric has been dealt with mainly as a single-objective problem. However, it is possible that one objective tends to dominate the search procedure and will also add bias-variance decomposition to the kind of result obtained.

## 3 Background

### 3.1 Bayesian Networks

A BN is a graphical model that represents a joint probability distribution over a set of random variables $\{X_1, \ldots, X_n\}$. BNs are represented as a pair $(G, \Theta)$, where the directed acyclic graph (DAG) is represented by $G = (U, E_G)$; $U$ is the set of nodes or random variables, and $E_G$ is the set of arcs that represent the probabilistic relationship among these variables. The parents of $X_i$ are denoted $PA_i$; $X_i$ is independent of its non-descendant variables given its parents.

Thus, $\Theta$ is a set of parameters which quantify the network. The joint probability distribution can be recovered from local conditional probability distributions as:
$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(x_i|PA_i)$.

### 3.2 Minimum Description Length

The crude definition of MDL [6] is of the form:

$$MDL = -logP(D|\Theta) + \frac{k}{2} log\, n, \quad (1) \qquad k = \sum_{i=1}^{m} q_i(r_i - 1), \quad (2)$$

where $D$ is the dataset, $\Theta$ represents the parameters of the model, $k$ is the dimension of the model, and $n$ is the sample size. The parameter $\Theta$ is the corresponding local probability distribution for each node in the network. The dimension of the model ($k$) is given by Equation 2.

For the case of Equation 2, $m$ is the number of variables, $q_i$ is the number of possible configurations of $PA_i$; $X_i$ and $r_i$ is the number of values of the variable.

The first term of Equation 1 measures the accuracy of the model ($f_1$) and the second term measures the complexity ($f_2$). The complexity of a BN is proportional to the number of arcs, as shows the Equation 2.

### 3.3 Multi-Objective Optimization Problem

According to Deb [4], a multi-objective optimization problem (MOOP) can be seen as a search problem that aims to minimize or maximize two (or more) objectives that are usually in conflict. Without loss of generality, a MOOP can be defined as: $\vec{f}(\vec{x}) = [f_1(\vec{x}), f_2(\vec{x}), \ldots, f_l(\vec{x})]$ where $\vec{x} = [x_1, \ldots, x_n] \in N^n$ is an $n$-variable decision vector, $\vec{f}$ is the set of objective functions to be minimized or maximized, and $l$ is the number of objectives.

According to this idea, the following definitions are provided: a) a solution $x_1$ dominates a solution $x_2$ (denoted by $x_1 \preceq x_2$) if the solution $x_1$ is not worse than $x_2$ in all objectives and it is better than $x_2$ in at least one objective. In MOOPs there is not a single optimal solution, conversely, we can find a set of solutions that have no other solution which dominates them when all objectives are currently considered. Hence, the set of non-dominated solutions is called *Pareto optimal* set, and the evaluations of each non-dominated solution in each objective function are known as the *Pareto front* [4].

## 4 Non-Dominated sorting Genetic Algorithm (NSGA-II)

NSGA-II is a fast elitist multi-objective evolutionary algorithm proposed by Deb et al. [4]. In NSGA-II the individuals are ordered into non-dominated sets called fronts. A rank based on the number of the front is assigned to each individual. To know how close an individual is to its neighbors, the crowding distance is computed for each individual.

**Algorithm 1** NS2BN

1: G=0 {Generation}
2: Generate a population $P$ of random solutions $\vec{x_i}, \forall i, i = 1, \ldots, POP\_SIZE$
3: Repair cycles of each $\vec{x}_i \; \forall i, i = 1, \ldots, POP\_SIZE$
4: Evaluate the fitness functions using the first and the second term of the Eq. 1 of each $\vec{p}_i \; \forall i, i = 1, \ldots, POP\_SIZE$
5: **while** $G \leq G_{max}$ **do**
6: Crate offspring population $Q$ using: binary tournament selection, one-point crossover and bit inversion mutation.
7: Repair cycles
8: Evaluate the fitness functions using the first and the second term of the Eq. 1 of each $\vec{x}_i \; \forall i, i = 1, \ldots, POP\_SIZE$
9: Combine parents and offspring population $R = P \cup Q$
10: Sort using non-dominated criterio
11: Replacement
12: $G = G + 1$
13: **end while**

The selection of parents is performed by using a binary tournament based on the rank and the crowding distance. The selected parents generate offsprings through crossover and mutation operators.

The pseudocode of the proposed approach named Non-dominated Sorting Genetic Algorithm for learning BN (NS2BN) is presented in Algorithm 1.

For the carried out of NS2BN: i) the representation of the individual is adjacency matrix, and ii) a repair operator that replaces values randomly when a cycle is identified.

## 5 Experiments and Results

### 5.1 Experimental Setup

This section presents the experimental setup. Firstly, we proposed four golden-standard networks with 6-nodes and the following characteristics: i) two of them with 8 arcs each one, and ii) two of them with 7 and 9 arcs, respectively; we call them A RDP, B RDP, C LED, and D LED, accordingly. In the four networks, all the random variables are binary, since this does not produce any qualitative impact on results in comparison to non-binary variables [1]. Then, we generate the datasets through these networks in instances of 1000, 5000, and 10000 cases. The first two of these databases were generated using a random probability distribution and the next two were generated with distribution $p = 0.1$ that according to [1] changing the parameters to be high or low tends to produce low-entropy distributions which have more potential for data compression.

Additionally, we include the following datasets: Asia that has 8-nodes and 8 arcs and Car Diagnosis that has 18-nodes and 20 arcs. Both networks were tested using the dataset with 1000, 5000 and 1000 instances.

Ten independent runs were made by each algorithm per database, with $20,000$ evaluations. A single objective Genetic Algorithm [9] (GA) was carried out for comparison propose. The individual representation consists of an adjacency matrix; the fitness function is the crude MDL, as described in the previous subsection (3.2).

In this algorithm, binary tournament parent selection, one-point crossover and bit inversion mutation are employed. The GA finds a single network for each execution, the network with the best MDL is chosen as the "genetic solution", meanwhile, in NS2BN the result of a run is a set of solutions with a variety of accuracy and structural complexity measurements. To choose an overall best solution from the Pareto front is scientifically invalid due to all the solutions are equally desirables, and normally the decision corresponds to a high-level expert knowledge in the modeling field.

In this work, to have a comparison between the multi-objective approach and the single-objective approach, from the accumulated Pareto front of ten executions, the solution nearest to a reference point which is $(0, 0)$ is chosen. To find this solution, all of them were normalized and the Euclidean distances were computed between the reference point and each Pareto solution. The solution with the shortest Euclidean distance is referred to as the "chosen solution" in this work.

The experimentation is presented in two parts 1) the comparison against the *gold-standard* network, the genetic solution and the chosen solution in terms of the Kullback-Leibler divergence (KLD) computed as the $log_2$ of the ratio of gold-standard network/chosen solution or genetic solution, according to the case, and 2) the analysis of the plots of the accumulated Pareto fronts.

The parameters setting employed by NS2BN and the GA were tuning empirically. The parameters are the follows: $POP\_ZIZE = 100$, $G_{max} = 200$, $C = 0.9$ and $M = 0.3$.

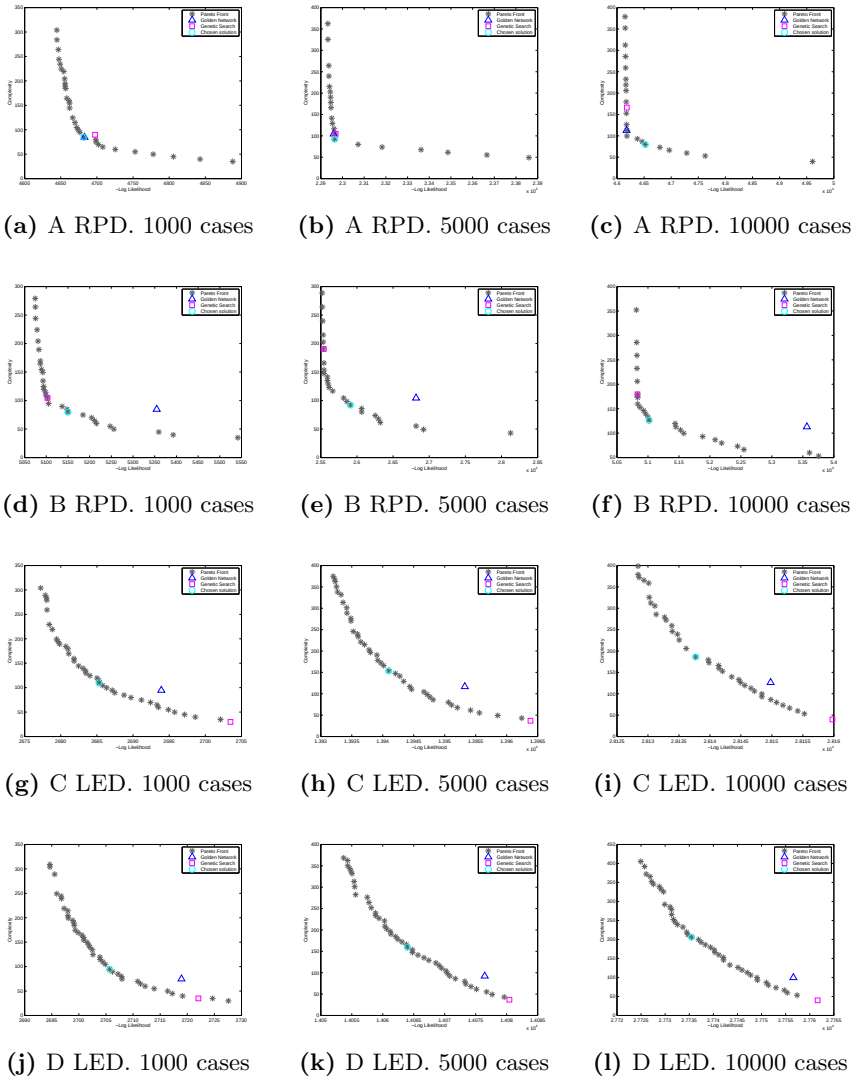
### 5.2 Results

Table 1 shows the results of the computation of the KLD. According to such a test, there were in ten databases significant differences in favor of the chosen solution that means that the chosen solution is closest to the gold-standard network concerning the subjacent distribution.

Since the genetic algorithm is searching for the minimum value of MDL, the genetic solutions show a minor MDL in sixteen databases. However, one of the objectives is punished in those results. The Figures 1d to 1f show how the genetic solution tends to choose solutions with a smaller log-likelihood but more complex, a similar situation occurs in the Figures 1g to 1i where the genetic algorithm chooses solutions less complex but with a worst log-likelihood value.

Regarding the sample size, Grünwald [6] points that crude MDL does not work well when the sample size is small or moderate and Hastie et al. [7] point out that a metric like crude MDL, in a finite sample, tends to select models less complex. However, these results agree with Grünwald and in contrast to Hastie's et al. our work, show a bias when the sample size is greater in the Genetic Solution, which is used a weighted sum, since this solution tends to select a more complex model (see Figures 1b, 1c, 1e, 1f, 2b, 2c, 2e and 2f).

The experiments generated by a low-entropy distribution show, as was pointed by Cruz-Ramírez et al. [3] that the presence of noise rate affects the behavior



**(a)** A RPD. 1000 cases **(b)** A RPD. 5000 cases **(c)** A RPD. 10000 cases

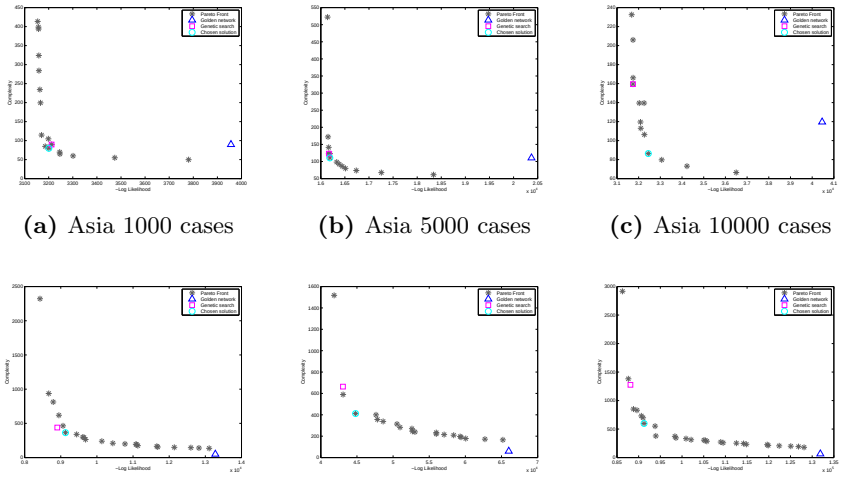
**(d)** B RPD. 1000 cases **(e)** B RPD. 5000 cases **(f)** B RPD. 10000 cases

**(g)** C LED. 1000 cases **(h)** C LED. 5000 cases **(i)** C LED. 10000 cases

**(j)** D LED. 1000 cases **(k)** D LED. 5000 cases **(l)** D LED. 10000 cases

**Fig. 1.** Accumulated Pareto front of the twelve first databases with 6-nodes with random probability distribution (RPD) and low-entropy probability distribution (LED). Gray stars - the accumulated front obtained by five runs of NSGA-II. Blue triangle - the golden-standard network. Pink square - the genetic solution and then green circle - the chosen solution from the Pareto front.

of MLD which tends to prefer the less complex models, even a network with no arcs.



**(a)** Asia 1000 cases **(b)** Asia 5000 cases **(c)** Asia 10000 cases

**(d)** Car diagnosis 1000 cases **(e)** Car diagnosis 5000 cases **(f)** Car diagnosis 10000 cases

**Fig. 2.** Accumulated Pareto front of the well-known benchmark databases with the different number of cases. Gray stars - the accumulated front obtained by five runs of NSGA-II. Blue triangle - the golden-standard network. Pink square - the genetic solution and the green circle - the chosen solution from the Pareto front.

However, these results show, independent of the sample size, solutions with better values in both terms (see Figures 1g to 1l).

## 6 Conclusion and Future Work

In this paper, a novel evolutionary bi-objective optimization approach for model selection of BN was presented.

The accuracy and the complexity, which are related to bias and variance respectively, were adopted as the objectives to be optimized to obtain models with an acceptable generalization performance. A set of trade-off solutions was obtained per database. A solution nearest to the origin was chosen as a competitive solution with a suitable trade-off between the objectives. This chosen solution was compared with a single-objective solution. The chosen solution achieved competitive results, especially in the complexity. It is important to note, that one of the main advantages of this approach is the set of trade-off solutions and that the selection of a model can be a high-level decision and must be performed by a domain expert of the modeling phenomenon. Additional advantages are that the proposed method can be applied to a database from different domains and can be extended to other models. As future work, different

**Table 1.** Kullback-Leibler divergence computed between the gold-standard network with the genetic search solution and the gold-standard network with the chosen solution of the Pareto front. Values in **boldface** mean the best value found.

| Golden-network | Genetic search | Chosen solution |
|---|---|---|
| A RPD. 1000 cases | 0.006256036 | **0.000412874** |
| A RPD. 5000 cases | 0.000735484 | **0.000166667** |
| A RPD. 10000 cases | **0.000622825** | 0.010558429 |
| B RPD. 1000 cases | **0.5008542** | 0.512832286 |
| B RPD. 5000 cases | **0.50817743** | 0.527715617 |
| B RPD. 10000 cases | **0.501635069** | 0.506660672 |
| C LED. 1000 cases | 0.006859061 | **0.000558415** |
| C LED. 5000 cases | 0.001254388 | **8.84927E-06** |
| C LED. 10000 cases | 0.000630321 | **0.000231126** |
| D LED. 1000 cases | 0.005505678 | **0.001674059** |
| D LED. 5000 cases | 0.001196043 | **0.0007695** |
| D LED. 10000 cases | 0.000561088 | **0.000529102** |
| Asia 1000 cases | 0.184669176 | **0.183903387** |
| Asia 5000 cases | 0.279944777 | **0.277977466** |
| Asia 10000 cases | 0.272191288 | **0.262362486** |
| Car diagnosis 1000 cases | **0.161505741** | 0.278079726 |
| Car diagnosis 5000 cases | **0.160725004** | 0.192815203 |
| Car diagnosis 10000 cases | **0.200548739** | 0.223971025 |

methods can be used to evaluate accuracy and complexity. Also, alternatives to reduce the computational cost of the algorithm can be included.

## References

1. Allen, T.V., Greiner, R.: Model Selection Criteria for Learning Belief Nets: An Empirical Comparison. In: Int Conf Mach Learn. pp. 1047–1054 (2000)
2. Blanco, R., Inza, I., Larrañaga, P.: Learning Bayesian networks in the space of structures by estimation of distribution algorithms. International Journal of Intelligent Systems 18(2), 205–220 (2003)
3. Cruz-Ramírez, N., Acosta-Mesa, H.G.G., Mezura-Montes, E., Guerra-Hernández, A., Hoyos-Rivera, G.d.J., Barrientos-Martínez, R.E.E., Gutiérrez-Fragoso, K., Nava-Fernández, L.A.A., González-Gaspar, P., Novoa-del Toro, E.M.M., Aguilera-Rueda, V.J., Ameca-Alducin, M.Y.: How good is crude MDL for solving the bias-variance dilemma? An empirical investigation based on Bayesian networks. PloS one 9(3) (2014)
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast Elitist Multi-Objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6, 182–197 (2000)
5. Gräning, L., Jin, Y., Sendhoff, B.: Generalization improvement in multi-objective learning. In: International Joint Conference on Neural Networks. p. 9893–9900. IEEE Press (2006)
6. Grünwald, P.D.: The Minimum Description Length Principle (Adaptive Computation and Machine Learning). The MIT Press (2007)

7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA (2001)
8. Heckerman, D.: A Tutorial on Learning with Bayesian Networks. In: Jordan, M.I. (ed.) Learning in Graphical Models, pp. 301–354. MIT Press, Cambridge, MA, USA (1999)
9. Holland, J.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI, USA (1975)
10. Li, G., Xing, L., Chen, Y.: A New BN Structure Learning Mechanism Based on Decomposability of Scoring Functions, pp. 212–224. Springer Berlin Heidelberg, Berlin, Heidelberg (2015)
11. Rosales, A., Escalante, H.J., C. A. Reyes, J.A.G., Coello, C.A.: Bias and Variance Optimization for SVMs Model Selection. In: Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference. pp. 136–141. St. Pete Beach, Florida, USA (2013)
12. Ross, B.J., Zuviria, E.: Evolving dynamic Bayesian networks with Multi-objective genetic algorithms. Applied Intelligence 26(1), 13–23 (Feb 2007)
13. Xiao-Lin, L., H., X.D., Chuan-Ming, C.: A method for learning Bayesian networks by using immune binary particle swarm optimization. In: Slezak, D., Kim, T.h., Zhang, Y., Ma, J., Chung, K.i. (eds.) Database Theory and Application, Communications in Computer and Information Science, vol. 64, pp. 115–121. Springer Berlin Heidelberg (2009)