ISSN 1870-4069

Discriminative Parameter Learning of Bayesian Networks Using Differential Evolution: A Preliminary Analysis

Alejandro Platas López, Nicandro Cruz Ramírez, Efrén Mezura Montes,
Alejandro Guerra Hernández

Universidad Veracruzana, Centro de Investigación en Inteligencia Artificial,
México

alejandroplatasl@gmail.com,{ncruz,emezura,aguerra}@uv.mx

Abstract. This work proposes Differential Evolution (DE) to train parameters of Bayesian Networks (BN) for optimizing the Conditional Log-Likelihood (Discriminative Learning) instead of the log-likelihood (Generative Learning). Although Discriminative Parameter Learning algorithms have been proposed, to the best of the authors' knowledge, a metaheuristic approach has not been devised yet. Thus, the objective of this research is to come up with this kind of solution and evaluate its behavior so that its feasibility in this domain can be determined. According to the theory such a solution tends to generate low-bias classifiers that minimize classification error but this is not reflected in results, regarding proposed method, bias in search for best solutions improves DEs performance.

Keywords: Bayesian networks, differential evolution, discriminative parameter learning.

1 Introduction

Two paradigms are distinguished for parameter learning of Bayesian networks. One of them, called Generative Learning (GL), optimizes Log-Likelihood in order to obtain the parameters that characterize the joint distribution in the form of local conditional distributions, and subsequently estimates class conditional probabilities using the Bayes rule. Even though this paradigm is computationally efficient, it is likely to generate biased classifiers [12].

The other paradigm optimizes Conditional Log-Likelihood (CLL) to directly estimate the parameters associated with conditional class distribution. Such paradigm is known as Discriminative Learning (DL) and generates low-bias classifiers that typically tend to minimize the classification error. In addition, the effect caused by the assumption of conditional independence among attributes in the network structure, but which may be violated in the data, is reduced.

pp. 75–82; rec. 2019-09-17; acc. 2019-09-30

However, the huge search space defined by the parameters that optimize the CLL function motivates this work to find efficient and effective search algorithms for discriminative parameter learning in BN classifiers [1,9].

Based on the above, different algorithms have been developed with the purpose of generating unbiased classifiers that mitigate the assumption of conditional independence among attributes. To the best of the authors' knowledge, there are nor proposals that apply evolutionary algorithms for DL of parameters. In this paper, we propose the use of the Differential Evolution (DE) algorithm for learning parameters in BN optimizing CLL. The aim is to understand the behavior of this evolutionary algorithm in this particular optimization task in both optimized structures for classification purposes, learned with a Bi-Objective PSO [2] and structures that are not optimized, learned by Tree-Augmented Network [3]. A comparison is made against some parameter learning algorithms for Log-Likelihood optimization.

The rest of the paper is organized as follows. Section 2 describes the optimization problem and introduces notations and terminologies. Section 3 gives details about the implementation of algorithms and experimental settings. The obtained results are presented in section 4. Finally, some conclusions and possible paths of future work are given in section 5.

2 Parameter Learning

GL is based on two steps, the first involves the maximization of $P(y, \mathbf{x})$, where y is the class and $\mathbf{x}$ is the set of attributes; and the second step is the application of the Bayes rule to obtain $P(y|\mathbf{x})$. In DL, it is possible to directly optimize $P(y|\mathbf{x})$, maximizing CLL.

Although there are approaches for parameter learning (not structures) with a discriminative approach [4]-[12], no meta-heuristic algorithms for DL of parameters in BNs have been adopted. A related work was proposed by [13], where they optimize LL (Generative approach) with a Genetic Algorithm combined with Expectation Maximization (GAEM). This proposal, according to the authors, combines the global search and local search properties of the respective algorithms. Part of notation and definitions used throughout this paper are taken from that work.

The proposed methods in this paper is based on Differential Evolution, which has been used for optimization problems in real-world applications[14]. DE was introduced in 1996 [15], and improved with some mechanisms to decrease the dependence to its parameter values such as the mutation factor F and the crossover rate CR [14], so as to increase its search performance[16].

To determine which search strategies provide a better performance, four DE variants will be used in this study: DE/rand/1/bin [15], JADE without archive, JADE with archive [14] and L-SHADE [16]. Such variant selection was made to include the most popular DE variant (DE/rand/1/bin), a variant with a novel differential mutation operator (JADE) and a recent one with a memory-based parameter adaptation mechanism (L-SHADE).

Let $\mathbf{X} = \{X_1, X_2, \ldots, X_R\}$ denote the set of random variables in a BN. Each random variable X_k is associated with a Conditional Probability Table (CPT). An individual ρ^t consists of a random variables vector of CPTs in a BN. The estimated CPT of an individual i at generation t is denoted by Θ_i^t. An individual is defined as a vector consisting of CPTs: $\rho_i^t = (\Theta_1^t, \Theta_2^t, \ldots, \Theta_R^t)$.

A CPT is generated based on the constraint that the sum of probabilities for different states of the random variable should be equal to 1 for a parent instantiation. A CPT is given by: $\Theta_i^t = (\theta_{1,1}^t, \ldots, \theta_{1,b}^t, \ldots, \theta_{a,1}^t, \ldots, \theta_{a,b}^t)$, where $\theta_{ab}^t \in [0, 1]$ denotes a probability value for a particular state given a parent instantiation.

3 Implementation

The obtained performance by the DE variants was compared based on both CLL optimization and predictive accuracy. Such results were further contrasted against those obtained by three GL algorithms: (1) Bayesian estimation, (2) maximum-likelihood and (3) Attribute-Weighted Naive Bayes. The parameter learning was applied to (1) BN structures optimized for classification with a bi-objective PSO algorithm that seeks trade-offs between predictive power and compression of data with the MDL metric [2]; the solution found in the "knee" of the Pareto front was selected as the best BN structure, and (2) BN structures learned with TAN-CL[1]. The datasets shown in Table 1 were used for comparison purposes and predictive accuracy was tested with 15 rounds of 2-fold stratified cross validation. 2-fold cross validation is used in order to maximize the variation in the training data from trial to trial [12].

Table 1. Details of datasets. Abrev: Abbreviation. Class = Number of classes. Att: Number of attributes. Case: Number of cases. θs: Number of parameters to be optimzed.

Data	**Abbrev**	**Class**	**Att**	**Case**	θs	**Data**	**Abbrev**	**Class**	**Att**	**Case**	θs
australian	aust	2	15	690	130	hepatitis	hepa	2	20	80	162
chess	ches	2	37	3296	290	lymphography	lymp	4	19	148	1220
cleve	clev	2	12	296	1005	Mofn-3-10	mofn	2	11	1324	78
corral	corr	2	7	128	46	pima	pima	2	9	768	102
crx	crx	2	16	653	848	segment	segm	7	20	2310	2548
diabetes	diab	2	9	768	102	Soybean-large	soyb	19	36	316	5265
flare	flar	8	11	1389	276	Tic-tac-toe	tic-	2	10	958	152
german	germ	2	21	1000	866	vehicle	vehi	4	19	958	1152
glass2	glas	2	10	163	1038	vote	vote	2	18	436	278
heart	hear	2	14	270	118	Waveform-21	wave	3	22	301	3186

Two repairs were applied to satisfy the constraints for $\theta_{ab}^t \in [0, 1]$:

$$\theta_{ab}^{'} = \begin{cases} |\theta_{ab}| \mod 1 & \text{if} \quad \theta_{ab} < 0 \\ 1 - (\theta_{ab} \mod 1) & \text{if} \quad \theta_{ab} > 1, \end{cases}$$

and to keep the sum of row vectors equal to 1:

$$\theta^{'}_{ab} = \theta_{ab} \Big/ \sum_{b=1}^{n} \theta_{ab}.$$

For the DE variants, 31 independent runs were performed on each dataset. The parameter values used in each DE variant are detailed in Table 2. Such values were adopted from the specialized literature [14] and by further experimentation.

Table 2. Parameter values of DE variants

DE algorithm	NP	G	F	CR	c	p	|A|
rand/1/bin	200	25 × Att	0.5	0.7			
JADE without A	200	10 × Att			0.05	0.05	∅
JADE with A	200	10 × Att			0.05	0.05	NP
L-SHADE	200	10 × Att			0.05	0.05	NP_g

4 Results

Based on the results summarized in Fig. 1, in datasets with a few number of parameters θ, the DE variants provided better results than those of the GL algorithms. Such behaviour was less marked in complex BN. Graphically there is no difference among structure types. As expected, those algorithms that had CLL as objective function, gave better results (Fig. 2). Regarding predictive accuracy, there is no clear evidence in favor of any approach, although DE variants are not the best, as shown in Figs. 3 and 4.

5 Conclusion and Future Work

A comparison of representative DE variants in an open problem about discriminative learning of parameters in BNs was presented. This would lead to the generation of classifiers with low bias that minimize the classification error. Based on the results obtained, difficulties were noted for DE variants when the number of parameters θ to be optimized increased. On the other hand, it was also observed that bias in search for high-quality solutions as well as the reduction in population size improved the DE variants performance. Future work contemplates the application of strategies that are capable of contending with big networks. Although it was not the purpose of this research, it is important to evaluate the performance of the proposed DE variants against state-of-art discriminative learning algorithms.

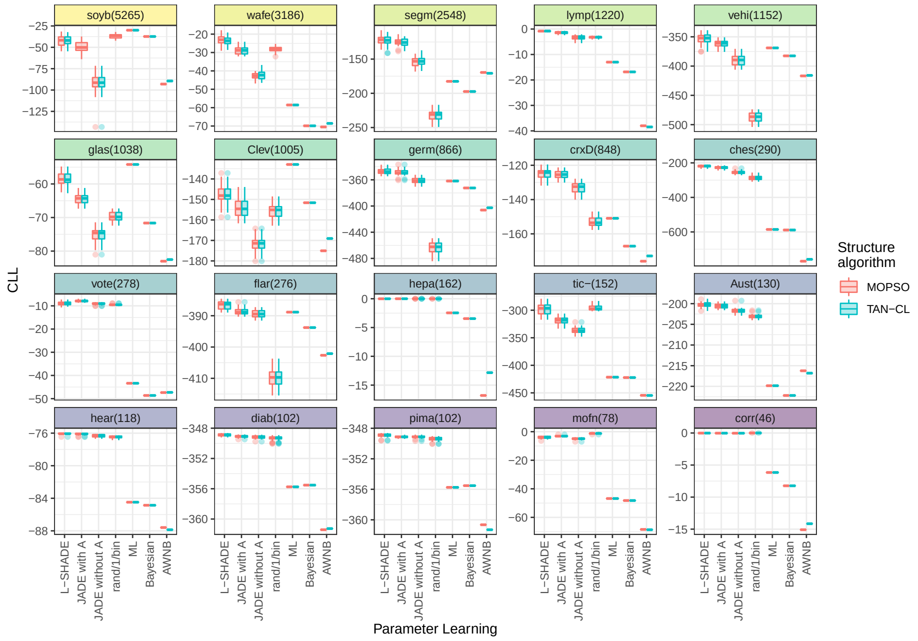

Fig. 1. Best CLL obtained in 31 independent runs by the DE variants and CLL obtained by the GL algorithms. Number of parameters θ are shown in parentheses.

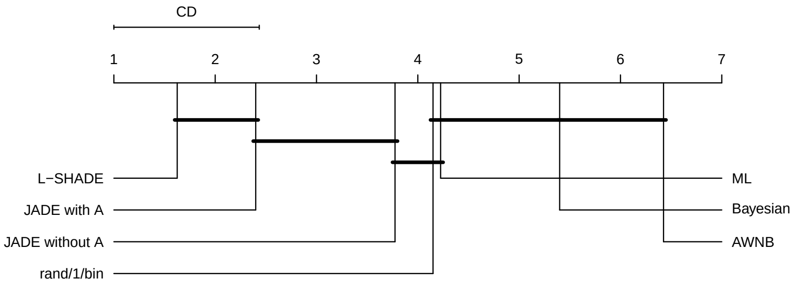

Fig. 2. Critical Differences diagram for the median CLL value of 31 independent runs (DE variants) and CLL value (GL algorithms). Horizontal line segments group together algorithms with CLL that are not significantly different (at $\alpha = 0.05$). Top line axis ranks methods from best (left) to worst (right).

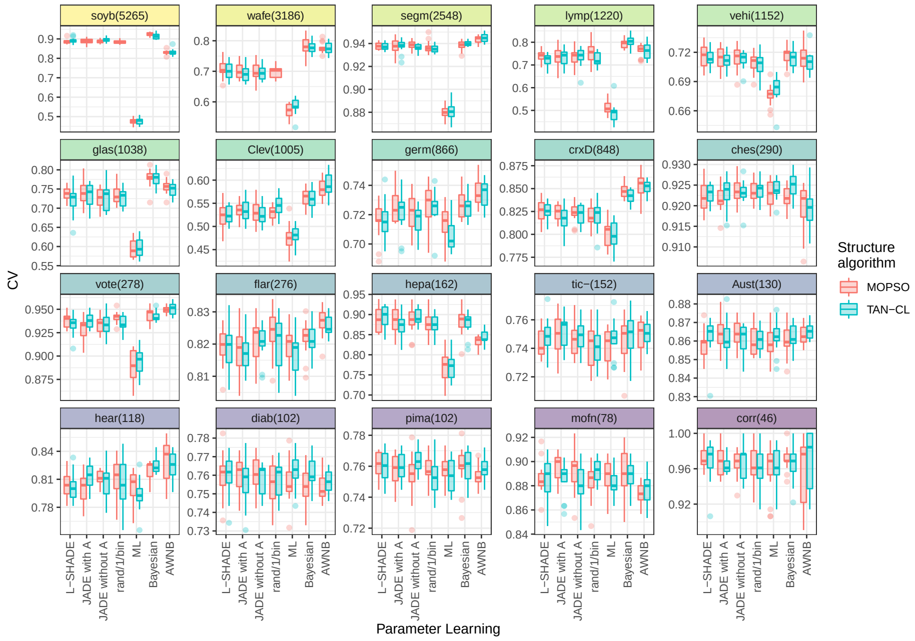

Fig. 3. Predictive accuracy of 15 rounds of 2-fold CV with the parameters learned by median of best solutions among 31 independent runs by DE variants and solution of GL algorithms. Datasets are sorted by number of parameters θ (shown in parentheses).

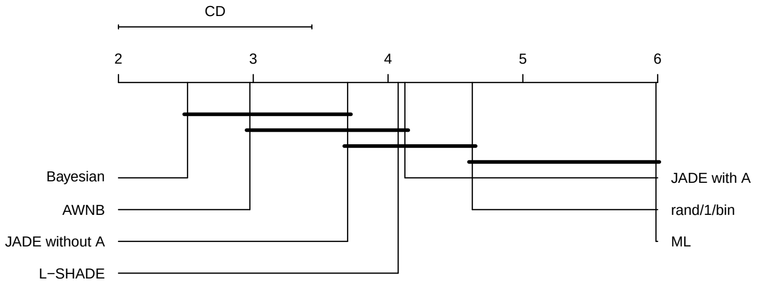

Fig. 4. Critical Differences diagram for median predictive accuracy of 15 rounds of 2-fold CV among algorithms. Horizontal line segments group together algorithms with predictive accuracy that are not significantly different (at $\alpha = 0.05$). Top line axis ranks methods from best (left) to worst (right).

References

1. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning, 29:2,131–163. (1997)
2. Aguilera-Rueda, V.J., Cruz-Ramírez, N., Mezura-Montes, E., Vilalta, R.: Learning bi-objective Bayesian Networks Structure from data using Particle Swarm Optimization. Elsevier (forthcoming)
3. Chow, C.K., Liu, C.N.: Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 14(3), pp. 462–467. (1968)
4. Shen, B., Su, X., Greiner, R., Musilek, P., Cheng, C.: Discriminative parameter learning of general Bayesian network classifiers. In: 15th IEEE International Conference on Tools with Artificial Intelligence, pp. 296–305. (2003)
5. Grossman, D., Domingos, P.: Learning Bayesian Network classifiers by maximizing conditional likelihood. In R. Greiner, and D. Schuurmans (Eds.), 21st International Conference on Machine Learning, ICML. pp. 361–368. (2004)
6. Guo, Y., Greiner, R.: Discriminative Model Selection for Belief Net Structures. In: Proceedings of the National Conference on Artificial Intelligence. 2. pp. 770–776. (2005)
7. Jing, Y., Pavlović, V., Rehg, J.M.: Efficient discriminative learning of Bayesian network classifier via boosted augmented naive Bayes. In: Proceedings of the 22nd international conference on Machine learning. ACM, New York, NY, USA, pp. 369–376. (2005)
8. Greiner, R., Su, X., Shen, B., Zhou, W.: Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers. Maching Learning 59, 297–322 (2005).
9. Su, J., Zhang, H., Ling, C.X., Matwin, S.: Discriminative parameter learning for Bayesian networks. In: Proceedings of the 25th international conference on Machine learning. ACM, New York, NY, USA, pp. 1016–1023. (2008)
10. Pernkopf, F., Wohlmayr, M.: On Discriminative Parameter Learning of Bayesian Network Classifiers. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II (ECML PKDD '09), Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 221–237. (2009)
11. Carvalho, A.M., Roos, T., Oliveira, A.L., Myllymäki, P.: Discriminative Learning of Bayesian Networks via Factorized Conditional Log-Likelihood. J. Mach. Learn. Res. 12, 2181–2210. (2011)
12. Zaidi, N.A., Webb, G.I., Carman, M.J., Petitjean, F., Buntine, W., Hynes, M., Sterck, H.: Efficient parameter learning of Bayesian network classifiers. Maching Learning. 106, 1289–1329. (2017)
13. Sundararajan, P.K., Mengshoel, O.J.: A Genetic Algorithm for Learning Parameters in Bayesian Networks using Expectation Maximization. In: Proceedings of the 8th International Conference on Probabilistic Graphical Models, PMLR 52, pp. 511–522. (2016)
14. Zhang, J., Sanderson, A.: JADE: adaptive differential evolution with optional external archive. IEEE Transactions on evolutionary computation, 13(5), pp. 945–958. (2009).
15. Price, K., Storn, R.: Minimizing the Real Functions of the ICEC'96 contest by Differential Evolution, IEEE International Conference on Evolutionary Computation (ICEC'96), pp. 842–844. (1996)

16. Tanabe, R., Fukunaga, A.S.: Improving the Search Performance of SHADE Using Linear Population Size Reduction. IEEE Congress on Evolutionary Computation (CEC), Beijing, pp. 1658–1665. (2014)