Estimation of Distribution Algorithms in Machine Learning

Pedro Larrañaga

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

EvoStar 2022, Madrid, April 22, 2022

Outline

Outline

Optimization in Machine Learning

Huge spaces to be optimized

- **Structures. Combinatorial optimization**
  - Number of possible feature subsets, $f(n)$, for a supervised classification problem with n predictor variables (Saeys et al. 2007): $f(n) = 2^n$
  - Number of possible partitional clustering assignments, $S(N, K)$, of N objects into K groups (Sharp 1968):

$$S(N, K) = \frac{1}{K!} \sum_{i=0}^{K} (-1)^{K-i} \binom{K}{i} i^N$$

  - Number of Bayesian networks structures, $f(n)$, is super-exponential in the number of nodes, n (Robinson 1977):

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i), \text{ for } n > 2,$$

    which is initialized with $f(0) = f(1) = 1$

- **Parameters. Continuous optimization**
  - Maximum likelihood estimation is not always achieved by means of a closed form

Machine Learning

Methods

Preprocessing

Dimensionality reduction (PCA, MDS, t-SNE, ..)
Visualization
Discretization

Supervised classification

Non probabilistic classifiers
k-nearest neighbors
Classification trees
Rule induction
Artificial neural networks
Support vector machines
Probabilistic classifiers
Discriminant analysis
Logistic regression
Bayesian classifier
Metaclassifiers
Stacking. Cascading. Bagging. Boosting.
Random Forest. Hybrid classifiers
Multidimensional classification

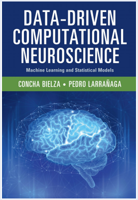
Clustering

Hierarchical clustering
Partitional clustering
Probabilistic clustering

Reinforcement learning

Probabilistic graphical models

Bayesian networks
Markov networks



Bielza and Larrañaga 2021

Optimization in Machine Learning

References

- C. Bielza, P. Larrañaga (2021). *Data-Driven Computational Neuroscience. Machine Learning and Statistical Models*. Cambridge University Press
- R. Robinson (1977). Counting unlabeled acyclic digraphs. *Lecture Notes in Mathematics* 622, 28-42, Springer
- Y. Saeys, I. Inza, P. Larrañaga (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517
- H. Sharp (1968). Cardinality of finite topologies. *Journal of Combinatorial Theory*, 5, 82–86

Outline

EDAs. A Toy Example

$$max\ O(\boldsymbol{x}) = \sum_{i=1}^{6} x_i$$

with $x_i = 0, 1$

EDAs. A Toy Example

$$max\ O(\boldsymbol{x}) = \sum_{i=1}^{6} x_i$$

with $x_i = 0, 1$

	X_1	X_2	X_3	X_4	X_5	X_6	$O(\boldsymbol{x})$
1	1	0	1	0	1	0	3
2	0	1	0	0	1	0	2
3	0	0	0	1	0	0	1
4	1	1	1	0	0	1	4
5	0	0	0	0	0	1	1
6	1	1	0	0	1	1	4
7	0	1	1	1	1	1	5
8	0	0	0	1	0	0	1
9	1	1	0	1	0	0	3
10	1	0	1	0	0	0	2
11	1	0	0	1	1	1	4
12	1	1	0	0	0	1	3
13	1	0	1	0	0	0	2
14	0	0	0	0	1	1	2
15	0	1	1	1	1	1	5
16	0	0	0	1	0	0	1
17	1	1	1	1	1	0	5
18	0	1	0	1	1	0	3
19	1	0	1	1	1	1	5
20	1	0	1	1	0	0	3

EDAs. A Toy Example

$$max\ O(\boldsymbol{x}) = \sum_{i=1}^{6} x_i$$

with $x_i = 0, 1$

	X_1	X_2	X_3	X_4	X_5	X_6	$O(\boldsymbol{x})$
1	1	0	1	0	1	0	3
2	0	1	0	0	1	0	2
3	0	0	0	1	0	0	1
4	1	1	1	0	0	1	4
5	0	0	0	0	0	1	1
6	1	1	0	0	1	1	4
7	0	1	1	1	1	1	5
8	0	0	0	1	0	0	1
9	1	1	0	1	0	0	3
10	1	0	1	0	0	0	2
11	1	0	0	1	1	1	4
12	1	1	0	0	0	1	3
13	1	0	1	0	0	0	2
14	0	0	0	0	1	1	2
15	0	1	1	1	1	1	5
16	0	0	0	1	0	0	1
17	1	1	1	1	1	0	5
18	0	1	0	1	1	0	3
19	1	0	1	1	1	1	5
20	1	0	1	1	0	0	3

EDAs. A Toy Example

Learning the probability distribution from the selected individuals

	X_1	X_2	X_3	X_4	X_5	X_6
1	1	0	1	0	1	0
4	1	1	1	0	0	1
6	1	1	0	0	1	1
7	0	1	1	1	1	1
11	1	0	0	1	1	1
12	1	1	0	0	0	1
15	0	1	1	1	1	1
17	1	1	1	1	1	0
18	0	1	0	1	1	0
19	1	0	1	1	1	1

EDAs. A Toy Example

Learning the probability distribution from the selected individuals

	X_1	X_2	X_3	X_4	X_5	X_6
1	1	0	1	0	1	0
4	1	1	1	0	0	1
6	1	1	0	0	1	1
7	0	1	1	1	1	1
11	1	0	0	1	1	1
12	1	1	0	0	0	1
15	0	1	1	1	1	1
17	1	1	1	1	1	0
18	0	1	0	1	1	0
19	1	0	1	1	1	1

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$$

EDAs. A Toy Example

Learning the probability distribution from the selected individuals

	X_1	X_2	X_3	X_4	X_5	X_6
1	1	0	1	0	1	0
4	1	1	1	0	0	1
6	1	1	0	0	1	1
7	0	1	1	1	1	1
11	1	0	0	1	1	1
12	1	1	0	0	0	1
15	0	1	1	1	1	1
17	1	1	1	1	1	0
18	0	1	0	1	1	0
19	1	0	1	1	1	1

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$$

$$p(X_1 = 1) = \frac{7}{10}$$

EDAs. A Toy Example

Learning the probability distribution from the selected individuals

	X_1	X_2	X_3	X_4	X_5	X_6
1	1	0	1	0	1	0
4	1	1	1	0	0	1
6	1	1	0	0	1	1
7	0	1	1	1	1	1
11	1	0	0	1	1	1
12	1	1	0	0	0	1
15	0	1	1	1	1	1
17	1	1	1	1	1	0
18	0	1	0	1	1	0
19	1	0	1	1	1	1

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$$

$$p(X_1 = 1) = \frac{7}{10} \quad p(X_2 = 1) = \frac{7}{10} \quad p(X_3 = 1) = \frac{6}{10}$$

$$p(X_4 = 1) = \frac{6}{10} \quad p(X_5 = 1) = \frac{8}{10} \quad p(X_6 = 1) = \frac{7}{10}$$

EDAs. A Toy Example

Learning the probability distribution from the selected individuals

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$$

$$p(X_1 = 1) = \tfrac{7}{10} \quad p(X_2 = 1) = \tfrac{7}{10} \quad p(X_3 = 1) = \tfrac{6}{10}$$

$$p(X_4 = 1) = \tfrac{6}{10} \quad p(X_5 = 1) = \tfrac{8}{10} \quad p(X_6 = 1) = \tfrac{7}{10}$$

EDAs. A Toy Example

Obtaining the new population by sampling from the probability distribution

$$p(X_1 = 1) = \frac{7}{10}; p(X_2 = 1) = \frac{7}{10}; p(X_3 = 1) = \frac{6}{10}$$

$$p(X_4 = 1) = \frac{6}{10}; p(X_5 = 1) = \frac{8}{10}; p(X_6 = 1) = \frac{7}{10}$$

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$$

$$\begin{array}{ll}
0.23 & p(X_1 = 1) = \frac{7}{10} > 0.23 \longrightarrow 1 \\
0.65 & p(X_2 = 1) = \frac{7}{10} > 0.65 \longrightarrow 1 \\
0.89 & p(X_3 = 1) = \frac{6}{10} < 0.89 \longrightarrow 0 \\
0.12 & p(X_4 = 1) = \frac{6}{10} > 0.12 \longrightarrow 1 \\
0.48 & p(X_5 = 1) = \frac{8}{10} > 0.48 \longrightarrow 1 \\
0.54 & p(X_6 = 1) = \frac{7}{10} > 0.54 \longrightarrow 1
\end{array}$$

EDAs. A Toy Example

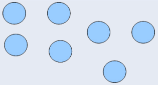
Obtaining the new population by sampling from the probability distribution

	X_1	X_2	X_3	X_4	X_5	X_6	$O(\boldsymbol{x})$
1	1	1	0	1	1	1	5
2	1	0	1	0	1	1	4
3	1	1	1	1	1	0	5
4	0	1	0	1	1	1	4
5	1	1	1	1	0	1	5
6	1	0	0	1	1	1	4
7	0	1	0	1	1	0	3
8	1	1	1	0	1	0	4
9	1	1	1	0	0	1	4
10	1	0	0	1	1	1	4
11	1	1	0	0	1	1	4
12	1	0	1	1	1	0	4
13	0	1	1	0	1	1	4
14	0	1	1	1	1	0	4
15	1	1	1	1	1	1	6
16	0	1	1	0	1	1	4
17	1	1	1	1	1	0	5
18	0	1	0	0	1	0	2
19	0	0	1	1	0	1	3
20	1	1	0	1	1	1	5

Probabilistic Models in EDAs

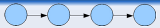
Univariate EDAs. Mühlenbein and Paaß (1996) (UMDA)

- Probabilistic model: $p_l(\boldsymbol{x}) = \prod_{i=1}^{n} p_l(x_i)$
- Structural learning: not necessary



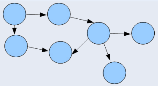
Bivariate EDAs. De Bonet et al. (1997) (MIMIC)

- Probabilistic model:
  $p_l^{\pi}(\boldsymbol{x}) = p_l(x_{i_1} \mid x_{i_2}) p_l(x_{i_2} \mid x_{i_3}) \cdots p_l(x_{i_{n-1}} \mid x_{i_n}) p_l(x_{i_n})$
- Structural learning: best permutation



Multivariate EDAs. Etxeberria and Larrañaga (1999) (EBNA); Pelikan et al. (1999) (BOA); Harik et al. (1999) (EcGA); Mühlenbein and Mahnig (1999) (LFDA)

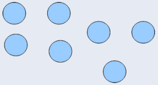
- Probabilistic model: $p_l(\boldsymbol{x}) = \prod_{i=1}^{n} p_l(x_i|\boldsymbol{pa}_i)$
- Structural learning: directed acyclic graph



Probabilistic Models in EDAs

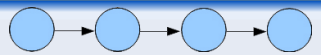
Univariate EDAs. Mühlenbein and Paaß (1996) (UMDA)

- Probabilistic model: $p_l(\boldsymbol{x}) = \prod_{i=1}^{n} p_l(x_i)$
- Structural learning: not necessary



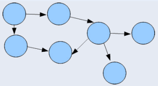
Bivariate EDAs. De Bonet et al. (1997) (MIMIC)

- Probabilistic model:
  $p_l^{\pi}(\boldsymbol{x}) = p_l(x_{i_1} \mid x_{i_2}) p_l(x_{i_2} \mid x_{i_3}) \cdots p_l(x_{i_{n-1}} \mid x_{i_n}) p_l(x_{i_n})$
- Structural learning: best permutation



Multivariate EDAs. Etxeberria and Larrañaga (1999) (EBNA); Pelikan et al. (1999) (BOA); Harik et al. (1999) (EcGA); Mühlenbein and Mahnig (1999) (LFDA)

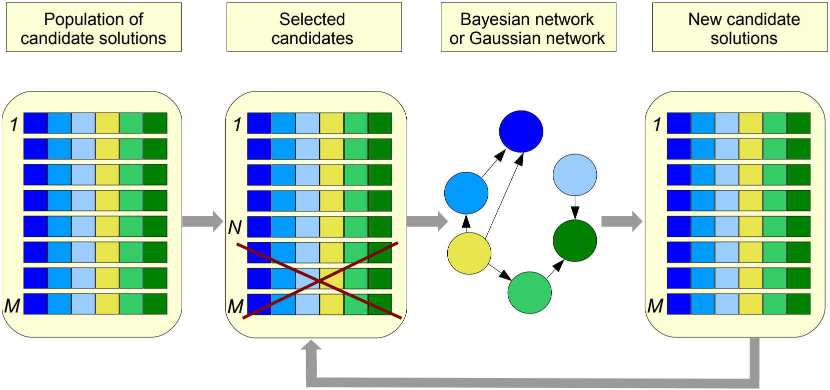
- Probabilistic model: $p_l(\boldsymbol{x}) = \prod_{i=1}^{n} p_l(x_i|\boldsymbol{pa}_i)$
- Structural learning: directed acyclic graph



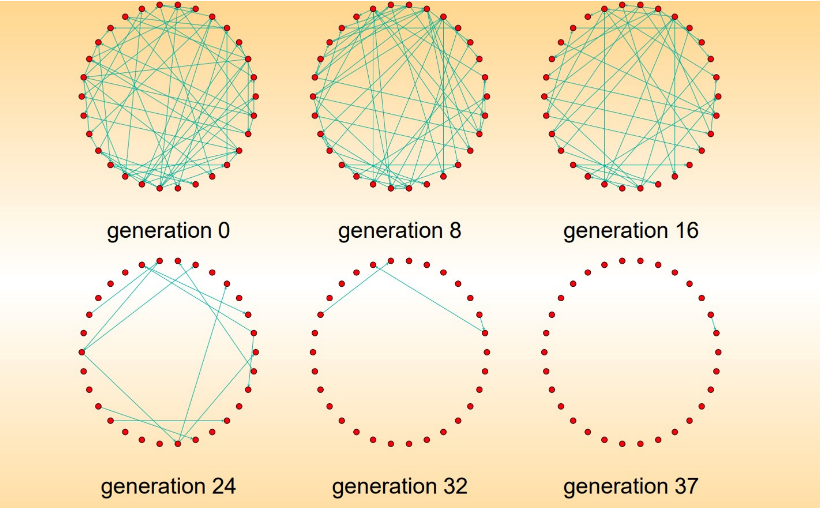
EDAs in continuous domains. Assuming Gaussianity: Larrañaga et al. (2000)

- Univariate: UMDA_c^G
- Bivariate: MIMIC_c^G
- Multivariate: EMNA_{global}^G, EMNA_{ee}^G, EGNA^G

Graphical Representation of EDAs

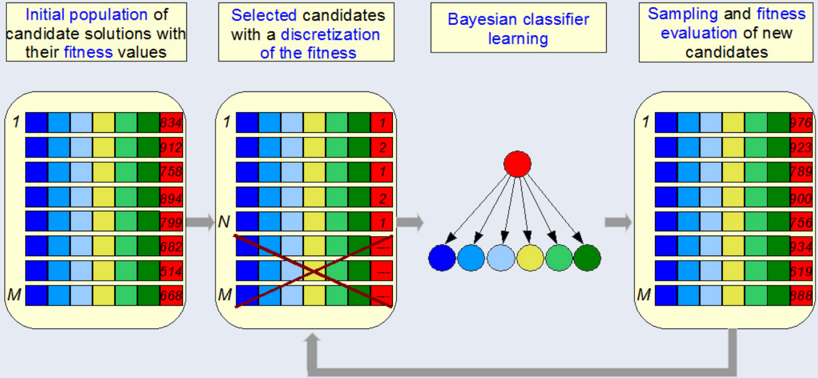


Evolution of Bayesian Network Structures in a EBNA Search (Bengoetxea 2002)



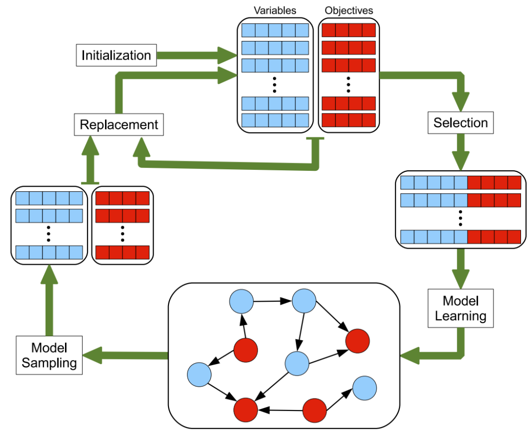
A Node for the Objective Function (Miquélez et al. 2004)

Estimation of Bayesian classifiers optimization algorithms



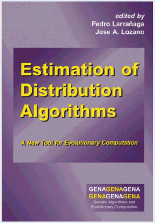
A Node for Each Objective Function (Karshenas et al. 2014)

EDAs for multiobjective optimization

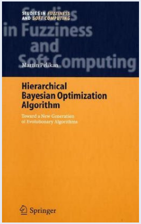


EDAs

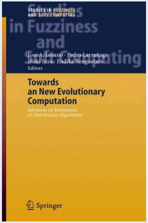
Books



Larrañaga and Lozano 2002



Pelikan 2005



Lozano et al. 2006

Special issues

2002

2005

2009

EDAs. Methodological Papers

References I

- L. Bao, X. Sun, Y. Chen, G. Man, H. Shao (2018). Restricted Boltzmann machine-assisted estimation of distribution algorithm for complex problems. *Complexity*, Article ID 2609014
- Bengoetxea E (2002). *Inexact Graph Matching Using Estimation of Distribution Algorithms*. PhD Thesis. Ecole Nationale Supérieure des Télécommunications
- S. Bhattacharjee (2019). *Variational Autoencoder Based Estimation of Distribution Algorithms and Applications to Individual Based Ecosystem Modeling Using EcoSim*. PhD Thesis. University of Windsor
- J.S. De Bonet, C.L. Isbell, P. Viola (1997). MIMIC: Finding optima by estimating probability densities. *Advances in Neural Information Processing Systems*, Vol. 9, 424-430
- P.A.N. Bosman, D. Thierens (2006). Numerical optimization with real-valued estimation-of-distribution algorithms. *Scalable Optimization via Probabilistic Modelling*. Springer, 91-120
- B. Doerr, M.S. Krejca (2020). Significance-based estimation-of-distribution algorithms. *IEEE Transactions on Evolutionary Computation* 24(6), 1025-1034
- W. Dong, X. Yao (2008). Unified eigen analysis on multivariate Gaussian based estimation of distribution algorithms. *Information Sciences*, 178(15), 3000-3023
- W. Dong, T. Chen, P. Tiňo, X. Yao (2013). Scaling up estimation of distribution algorithms for continuous optimization. *IEEE Transactions on Evolutionary Computation*, 17 (6), 797-822
- R. Etxeberria, P. Larrañaga (1999). Global optimization using Bayesian networks. *Second International Symposium on Artificial Intelligence*, 332-339
- G. Harik, F.G. Lobo, D.E. Goldberg (1998). The compact genetic algorithm. *Proceedings of the IEEE Conference on Evolutionary Computation*, 523-528
- M. Henrion (1988). Propagating uncertainty in Bayesian networks by probabilistic logic sampling. *Uncertainty in Artificial Intelligence*, Vol. 2, 149-163
- H. Karshenas, R. Santana, C. Bielza, P. Larrañaga (2013). Regularized continuous estimation of distribution algorithms. *Applied Soft Computing*, 13(5), 2412–2432

EDAs. Methodological Papers

References II

- H. Karshenas, R. Santana, C. Bielza, P. Larrañaga, (2014). Multi-objective estimation of distribution algorithms based on joint modeling of objectives and variables. *IEEE Transactions on Evolutionary Computation*, 18(4), 519-542
- P. Larrañaga, R. Etxeberria, J. A. Lozano, and J. M. Peña (1999). *Optimization by Learning and Simulation of Bayesian and Gaussian Networks*. Technical Report EHU-KZAA-IK-4-99. Department of Computer Science and Artificial Intelligence. University of the Basque Country
- P. Larrañaga, H. Karshenas, C. Bielza, R. Santana (2012). A review on probabilistic graphical models in evolutionary computation. *Journal of Heuristics*, 18(5), 795–819
- Y. Liang, Z. Ren, X. Yao, Z. Feng, A. Chen, W. Guo (2020). Enhancing Gaussian estimation of distribution algorithm by exploiting evolution direction with archive. *IEEE Transactions on Cybernetics*, 50 (1), 140-152
- Q. Lu, X. Yao (2005). Clustering and learning Gaussian distribution for continuous optimization. *IEEE Transactions on Systems, Man and Cybernetics*, 35(2), 195-204
- T. Miquélez, E. Bengoetxea, P. Larrañaga (2004). Evolutionary computation based on Bayesian classifiers. *International Journal of Applied Mathematics and Computer Science*, 14, 101-115
- T. Miquelez, E. Bengoetxea, A. Mendiburu, P. Larrañaga (2007). Combining Bayesian classifiers and estimation of distribution algorithms for optimization in continuous domains. *Connection Science*, 19(4), 297-319
- H. Mühlenbein, G. Paaß (1996). From recombination of genes to the estimation of distributions I. Binary parameters. *Lecture Notes in Computer Science 1411*, 178-187
- H. Mühlenbein, T. Mahnig, A. Ochoa (1999). Schemata, distributions and graphical models in evolutionary optimization. *Journal of Heuristics*, 5 (2), 213-247
- T.K. Paul, H. Iba (2003). Reinforcement learning estimation of distribution algorithm. *Genetic and Evolutionary Computation Conference*, 1259-1270
- M. Pelikan, D.E. Goldberg, E. Cantú-Paz (1999). BOA: The Bayesian optimization algorithm. *Proceedings of the Genetic and Evolutionary Computation Conference*, Vol. 1, 525-532

EDAs. Methodological Papers

References III

- M. Pelikan, D. E. Goldberg, F. Lobo (2002). A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21 (1), 5-20
- J. M. Peña, J. A. Lozano, P. Larrañaga (2005). Globally multimodal problem optimization via an estimation of distribution algorithm based on unsupervised learning of Bayesian networks. *Evolutionary Computation*, 13, 1, 43-66
- L. PourMohammadBagher, M.M. Ebadzadeh, R. Safabakhsh (2017). Graphical model based continuous estimation of distribution algorithm. *Applied Soft Computing*, 58,388-400
- M. Probst, F. Rothlauf (2020). Harmless overfitting: Using denoising autoencoders in estimation of distribution algorithms. *Journal of Machine Learning Research*, 21(78), 1-31
- R. Santana (2005). Estimation of distribution algorithms with Kikuchi approximations. *Evolutionary Computation*, 13(1), 67-97
- R. Santana, P. Larrañaga, J. A. Lozano (2008). Combining variable neighborhood search and estimation of distribution algorithms. *Journal of Heuristics*, 14, 519–547
- R. Santana, J. A. Lozano, P. Larrañaga (2008). Research topics in discrete estimation of distribution algorithms. *Memetic Computing*, 1, 35-54
- S Shakya, J McCall (2007). Optimization by estimation of distribution with DEUM framework based on Markov random fields. *International Journal of Automation and Computing*, 4(3), 262-272
- A. Utamina (2021). A comparative study of hybrid estimation distribution algorithms in solving the facility layout problem. *Egyptian Informatics Journal*, 22(4), 505-513
- H. Xu, J. Yang, P. Jia, Y. Ding (2013). Effective structure learning for estimation of distribution algorithms via L1-regularized Bayesian networks. *International Journal of Advanced Robotic Systems*, 10(1)
- Q. Zhang, A. Zhou, Y. Jin (2008). RM-MEDA: A regularity model based multiobjective estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation*, 12 (1), 41-63

EDAs. Theoretical Papers

References

- T. Chen, K. Tang, G. Chen, X. Yao (2010). Analysis of computational time of simple estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation*, 14(1), 1-22
- C. González, J. A. Lozano, P. Larrañaga (2001). Analyzing the PBIL algorithm by means of discrete dynamical systems. *Complex Systems*, 12 (4), 465-479
- C. González, J. A. Lozano, P. Larrañaga (2002). Mathematical modeling of UMDAc algorithm with tournament selection. Behaviour on linear and quadratic functions. *International Journal of Approximate Reasoning*, 31 (4), 313-340
- M.S. Krejca, C. Witt (2020). Theory of estimation-of-distribution algorithms. *Theory of Evolutionary Computation*, Springer, 405-442
- B. Doerr, M.S. Krejca (2021). A simplified run time analysis of the univariate marginal distribution algorithm on LeadingOnes. *Theoretical Computer Science* 851, 121-128
- A. H. Wriht, S. Pulavarty (2005). On the convergence of an estimation of distribution algorithm based on linkage discovery and factorization. *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, 695-702
- Q. Zhang, H. Mühlenbein (2004). On the convergence of a class of estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation*, 8 (2), 127-136

Outline

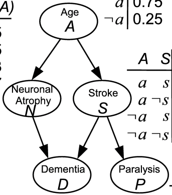
Bayesian Networks

DAG + CPTs

- Conditional independence: **W** and **T** are conditionally independent given **Z** $\Leftrightarrow p(\mathbf{W}|\mathbf{T},\mathbf{Z}) = p(\mathbf{W}|\mathbf{Z})$
- Directed acyclic graph (DAG)
- Conditional probability tables (CPTs)
- $p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{Pa}(X_i))$

A	$p(A)$
a	0.75
$\neg a$	0.25

A	N	$p(N\|A)$
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97



A	S	$p(S\|A)$
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	$p(D\|N,S)$
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	$p(P\|S)$
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

$p(A,N,S,D,P) = p(A)p(N|A)p(S|A)p(D|N,S)p(P|S)$

Bayesian Networks

DAG + CPTs

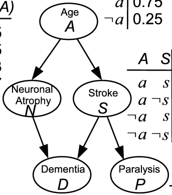
- Conditional independence: **W** and **T** are conditionally independent given **Z** $\Leftrightarrow p(\mathbf{W}|\mathbf{T},\mathbf{Z}) = p(\mathbf{W}|\mathbf{Z})$
- Directed acyclic graph (DAG)
- Conditional probability tables (CPTs)
- $p(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{Pa}(X_i))$

Inference

- Exact: variable elimination, message passing
- Approximate: sequential simulation and MCMC

A	$p(A)$
a	0.75
$\neg a$	0.25

A	N	$p(N\|A)$
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97

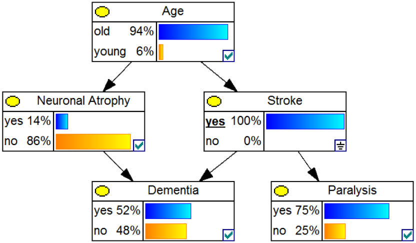


A	S	$p(S\|A)$
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	$p(D\|N,S)$
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	$p(P\|S)$
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

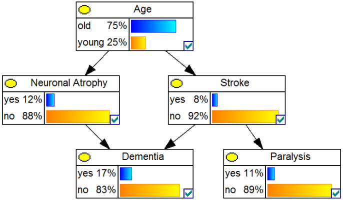
$p(A,N,S,D,P) = p(A)p(N|A)p(S|A)p(D|N,S)p(P|S)$



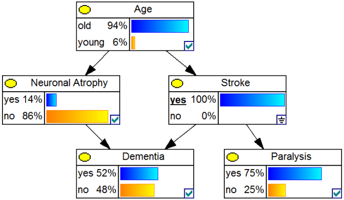
$p(X_i|\texttt{Stroke=yes})$

Bielza and Larrañaga 2021

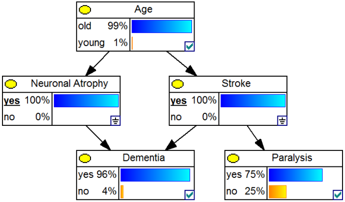
Conditional Independence



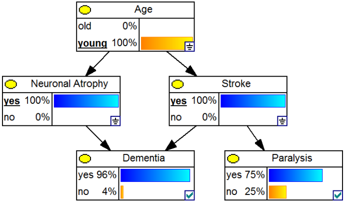
$p(X_i)$



$p(X_i|\texttt{Stroke=yes})$



$p(X_i|\texttt{Stroke=yes, Neural Atropy=yes})$

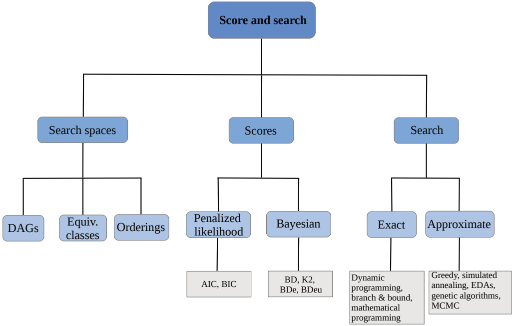


$p(X_i|\texttt{Stroke=yes, Neural Atropy=yes, Age=young})$

Learning Bayesian Networks from Data

Two tasks

- Parameters $p(X_i = x_i \mid \mathbf{Pa}(X_i) = \mathbf{pa}_i^j)$: MLE or Bayesian
- Structure: conditional independence tests or by optimizing a score



Scores

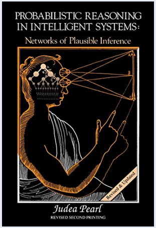
- Penalized likelihood: avoid structural overfitting
- Bayesian: $\arg\max_{\mathcal{G}} p(\mathcal{G}|\mathcal{D})$, with

$$p(\mathcal{G}|\mathcal{D}) \propto \overbrace{p(\mathcal{D}|\mathcal{G})}^{\text{marginal likeli.}} \overbrace{p(\mathcal{G})}^{\text{prior}}, \text{ with}$$

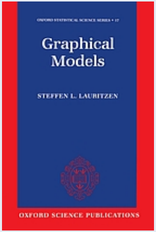
$$p(\mathcal{D}|\mathcal{G}) = \int \underbrace{p(\mathcal{D}|\mathcal{G}, \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{f(\boldsymbol{\theta}|\mathcal{G})}_{\text{prior param.}} d\boldsymbol{\theta}$$

Bayesian Networks

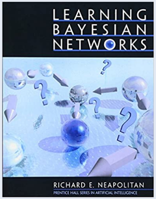
Books



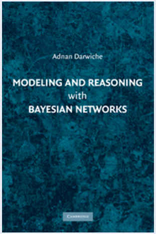
Pearl 1988



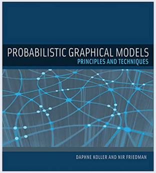
Lauritzen 1996



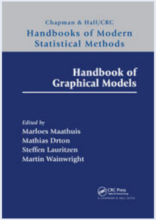
Neapolitan 2003



Darwiche 2009



Koller and Friedman 2009



Maathius et al. 2019

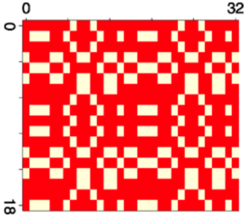
Bayesian Networks

References

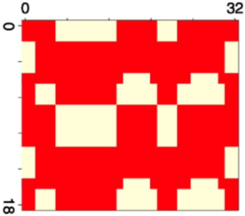
- Castillo E, Gutierrez JM, Hadi A (1997). *Expert Systems and Probabilistic Network Models*. Springer
- Darwiche A (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press
- Jensen F, Nielsen TD (2007). *Bayesian Networks and Decision Graphs*. Springer
- Koller D, Friedman N (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press
- Lauritzen S (1996). *Graphical Models*. Oxford University Press
- Maathuis M, Drton M, Lauritzen S, Wainwright M (2019). *Handbook of Graphical Models*. CRC Press
- Neapolitan (2003). *Learning Bayesian Networks*. Prentice Hall
- Pearl J (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann
- Sucar E (2015). *Probabilistic Graphical Models: Principles and Applications*. Springer

Outline

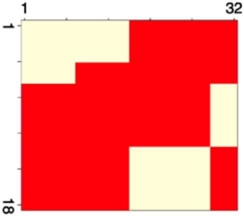
Improving Table Interpretation (Bengoetxea et al. 2011)



(a) Stress=1936



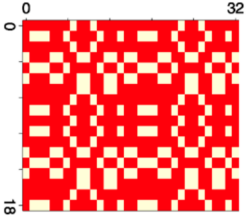
(b) Stress=848



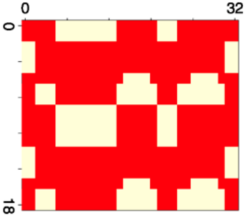
(c) Stress=330

- The problem. The way rows and columns are ordered in a table is a very sensitive issue that affects its readability. The optimal ordering of tables is equivalent to solving two travelling salesman problems, one for the R rows and the other for the C columns $\Rightarrow$ cardinality of the search space: $R! \cdot C!$
- Individual representation.
  - (a) Discrete: $\mathbf{x} = (x_1, \ldots, x_R, x_{R+1}, \ldots, x_{R+C})$, where $x_i = k$ means that the order of the original ith row is k, and $x_{R+j} = l$ means that the order for the jth column is l
  - (b) Continuous: the real vectors of values were transformed into permutations as the respective order in the continuous individual
- EDAs. Univariate, bivariate and multivariate in both discrete and continuous domains

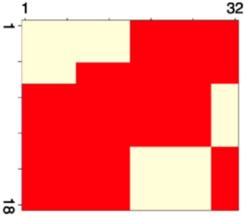
Improving Table Interpretation (Bengoetxea et al. 2011)



(a) Stress=1936



(b) Stress=848



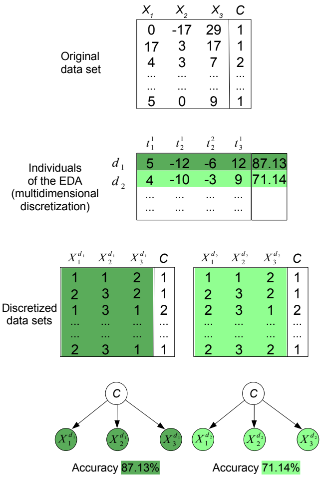
(c) Stress=330

- The problem. The way rows and columns are ordered in a table is a very sensitive issue that affects its readability. The optimal ordering of tables is equivalent to solving two travelling salesman problems, one for the R rows and the other for the C columns $\Rightarrow$ cardinality of the search space: $R! \cdot C!$
- Individual representation.
  - (a) Discrete: $\mathbf{x} = (x_1, \ldots, x_R, x_{R+1}, \ldots, x_{R+C})$, where $x_i = k$ means that the order of the original ith row is k, and $x_{R+j} = l$ means that the order for the jth column is l
  - (b) Continuous: the real vectors of values were transformed into permutations as the respective order in the continuous individual
- EDAs. Univariate, bivariate and multivariate in both discrete and continuous domains

References

- E. Bengoetxea, P. Larrañaga, C. Bielza, J.A. Fernández del Pozo (2011). Optimal row and column ordering to improve table interpretation using estimation of distribution algorithms. *Journal of Heuristics*, 17(5), 567–588

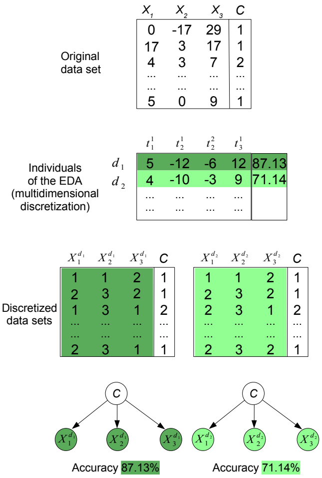
Wrapper Multidimensional Discretization (Flores et al. 2007)



Multidimensional discretization

- The problem. Optimal (maximizing the estimated accuracy) multidimensional discretization
- Individual representation depends on the discretization process (number of bins and cutpoints for each predictor variable)
  - An individual in the EDA represents a discretization policy that transforms the original dataset into a discretized one
- EDAs. UMDA_c^G

Wrapper Multidimensional Discretization (Flores et al. 2007)



Multidimensional discretization

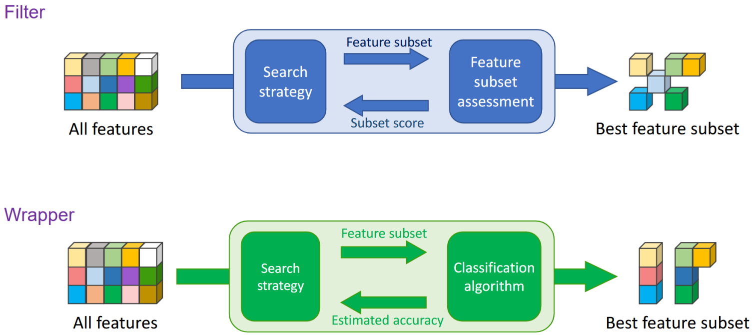
- The problem. Optimal (maximizing the estimated accuracy) multidimensional discretization
- Individual representation depends on the discretization process (number of bins and cutpoints for each predictor variable)
  - An individual in the EDA represents a discretization policy that transforms the original dataset into a discretized one
- EDAs. UMDA_c^G

References

- J. L. Flores, I. Inza, P. Larrañaga (2007). Wrapper discretization by means of estimation of distribution algorithms. *Intelligent Data Analysis Journal*, 11(5), 525–546

Outline

Feature Subset Selection (Inza et al. 2000)



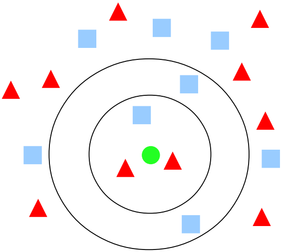
- The problem. Optimal (maximizing the accuracy) subset of features for a given supervised classification paradigm
- Individual representation. $\mathbf{x} = (x_1, \ldots, x_n)$ where $x_i = 1$ if variable X_i is selected, and 0 otherwise
- EDAs. EBNA for ID3 and naive Bayes

Feature Subset Selection

References

- M. Ayodele (2019). Application of estimation of distribution algorithm for feature selection. *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2019*, 43-44
- E. Cantú-Paz (2002). Feature subset selection by estimation of distribution algorithms. *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, 303-310
- I. Inza, P. Larrañaga, B. Sierra (2001). Feature subset selection by Bayesian networks: A comparison with genetic and sequential algorithms. *International Journal of Approximate Reasoning*, 27, 143–164
- I. Inza, P. Larrañaga, R. Etxeberria, B. Sierra (2000). Feature subset selection by Bayesian network–based optimization. *Artificial Intelligence*, 123, 157–184
- S. Maza, M. Touahria (2019). Feature selection for intrusion detection using new multi-objective estimation of distribution algorithms. *Applied Intelligence*, 49, 4237-4257
- G. Neuman, D. Cairns (2013). Applying a hybrid targeted estimation of distribution algorithm to feature selection problems. *Proceedings of the 5th International Joint Conference on Computational Intelligence, ECTA-2013*, 136-143
- Y. Saeys, S. Degroeve, D. Aeyels, Y. Van de Peer, P. Rouzé (2003). Fast feature selection using a simple estimation of distribution algorithm: A case study on splice site prediction. *Bioinformatics*, 19, 2, ii179-ii188

k-Nearest Neighbors (Inza et al. 2002)

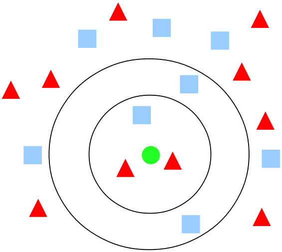


$$d(\mathbf{x}, \mathbf{x}^i) = \sum_{j=1}^{n} w_j \delta(x_j, x_j^i)$$

Feature weighting

- The problem. Search for the optimal (in terms of accuracy) feature weighting
- Individual representation: discrete (three possible values), or continuous
- EDAs. EBNA and EGNA

k-Nearest Neighbors (Inza et al. 2002)



$$d(\mathbf{x}, \mathbf{x}^i) = \sum_{j=1}^{n} w_j \delta(x_j, x_j^i)$$

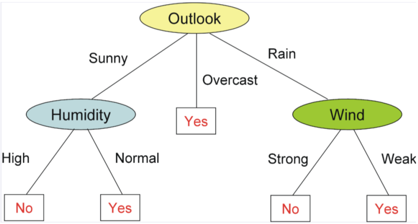
Feature weighting

- The problem. Search for the optimal (in terms of accuracy) feature weighting
- Individual representation: discrete (three possible values), or continuous
- EDAs. EBNA and EGNA

References

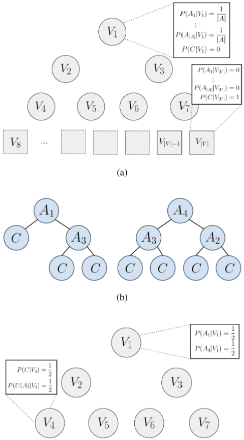
- I. Inza, P. Larrañaga, B. Sierra (2002). Feature weighting for nearest neighbor by EDAs. in *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*, Kluwer Academic Publishers, 295–311

Classification Trees (Cagnini et al. 2017)

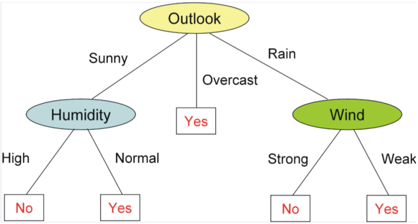


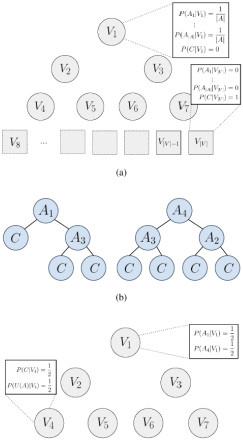
Optimal classification tree

- The problem. Optimal (maximizing the estimated accuracy) classification tree for continuous predictors
- Individual representation. Binary tree with a given maximal depth. Root node and internal nodes selected from probability distributions
- EDAs. UMDA



Classification Trees (Cagnini et al. 2017)





Optimal classification tree

- The problem. Optimal (maximizing the estimated accuracy) classification tree for continuous predictors
- Individual representation. Binary tree with a given maximal depth. Root node and internal nodes selected from probability distributions
- EDAs. UMDA

References

- H.E.L. Cagnini, R.C. Barros, M.P. Basgalupp (2017). Estimation of distribution algorithms for decision-tree induction. *2017 IEEE Congress on Evolutionary Computation*

Rule Induction (Sierra et al. 2002)

$$\left\{ \begin{array}{llll} \mathcal{R}: & \texttt{IF} \; (X_{25} = 2 \;\; \texttt{AND} \;\; X_{56} = 3) & \texttt{OR} & \\ & (X_2 = 1 \;\; \texttt{AND} \;\; X_5 \neq 1) & \texttt{THEN} & C = \texttt{I} \end{array} \right\}$$

Pittsburgh-like approach

- The problem. Optimal (maximizing the estimated accuracy) rule. Classifier system
- Individual representation. The antecedent of the rule consists on disjunction of simple antecedents, where a single rule dimension is given by n, the number of predictor variables, allowing for each variable to take values that are equal to, different from, and any possible value
- EDAs. UMDA, EBNA

Rule Induction (Sierra et al. 2002)

$$\left\{ \mathcal{R}: \quad \texttt{IF} \; \begin{array}{l} (X_{25} = 2 \;\; \texttt{AND} \;\; X_{56} = 3) \\ (X_2 = 1 \;\; \texttt{AND} \;\; X_5 \neq 1) \end{array} \quad \begin{array}{l} \texttt{OR} \\ \texttt{THEN} \quad C = \texttt{I} \end{array} \right\}$$

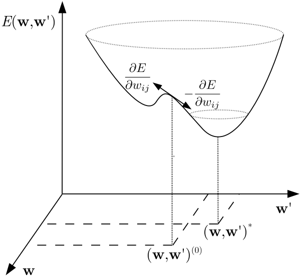
Pittsburgh-like approach

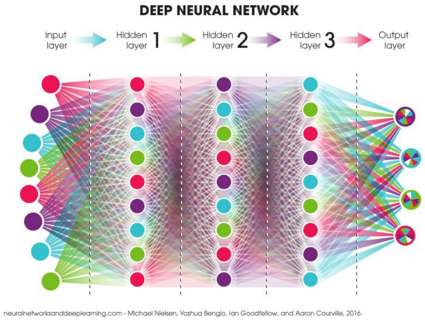
- The problem. Optimal (maximizing the estimated accuracy) rule. Classifier system
- Individual representation. The antecedent of the rule consists on disjunction of simple antecedents, where a single rule dimension is given by n, the number of predictor variables, allowing for each variable to take values that are equal to, different from, and any possible value
- EDAs. UMDA, EBNA

References

- B. Sierra, E.A. Jiménez, I. Inza, P. Larrañaga (2002). Rule induction by estimation of distribution algorithms. In *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*, Kluwer Academic Publishers, 313-322

Artificial Neural Networks (Baluja 1995)





Optimal weights

- The problem. Minimization of $E(\mathbf{w}, \mathbf{w}') = \frac{1}{N}\sum_{k=1}^{N}(c^k - \hat{c}^k)^2$. Alternative to the backpropagation algorithm, a gradient descent method
- Individual representation. k-dimensional vectors of real numbers, where k denotes the number of weights in the artificial neural network (a feed-forward multilayer perceptron)
- EDAs. PBIL: $p_{l+1}(\mathbf{x}) = (1-\alpha)p_l(\mathbf{x}) + \alpha\frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_{k:M}^l$

Artificial Neural Networks

References

- S. Baluja (1995). *An empirical comparison of seven iterative and evolutionary function optimization heuristics*. Technical Report CMU-CS-95-193, Carnegie Mellon University
- E. Cantú-Paz (2003). Pruning neural networks with distribution estimation algorithms. *Genetic and Evolutionary Computation Conference*
- C. Cotta, E. Alba, R. Sagarna, P. Larrañaga (2002). Adjusting weights in artificial neural networks using evolutionary algorithms. in *Estimation of Distribution Algorihtms. A New Tool for Evolutionary Computation*, Kluwer Academic Publishers, 361–377
- Y. Chen, A. Abraham (2006) Estimation of distribution algorithms for optimization of neural networks for intrusion detection. *8th International Conference on Artificial Intelligence and Soft Computing*
- E. Galić, M. Höhfeld (1996). Improving the generalization performance of multi-layer-perceptrons with population-based incremental learning. *Parallel Problem Solving from Nature IV*, 740-750
- M.R. Gallacher (2000). *Multi-layer Perceptron Error Surfaces: Visualization, Structure and Modelling*. PhD Thesis, University of Queensland
- G. Holker, M.V. dos Santos (2010). Toward an estimation of distribution algorithm for the evolution of artificial neural networks. *Proceedings of the Third C* Conference on Computer Science and Software Engineering*
- J.-Y. Li, Z.-H. Zhan, J. Xu, S. Kwong, J. Zhang (2021). Surrogate-assisted hybrid-model estimation of distribution algorithm for mixed-variable hyperparameters optimization in convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*
- N.F.A. Rasli, M.S.M. Kasihmuddin, M.A. Mansor, M.F.M. Basir, S. Sathasivam (2020). *k* satisfiability programming by using estimation of distribution algorithm in Hopfield neural network. *Proceedings of the 27th National Symposium on Mathematical Sciences*
- Q. Xu, A. Liu, X. Yuan, Y. Song, C. Zhang, Y. Li (2021). Random mask-based estimation of the distribution algorithm for stacked auto-encoder one-step pre-training. *Computers and Industrial Engineering*, 158, 107400

Logistic Regression (Robles et al. 2008)

- The logistic regression model: $p(C=1|\mathbf{x},\boldsymbol{\beta}) = \theta_{\mathbf{x}} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}$
- The likelihood function is: $\mathcal{L}(\boldsymbol{\beta}|\mathbf{x}^1,\ldots,\mathbf{x}^N) = p(c^1,\ldots,c^N|\mathbf{x},\theta_{\mathbf{x}}) = \prod_{i=1}^N \theta_i^{c^i}(1-\theta_i)^{1-c^i}$
- The likelihood equations are:

$$\left\{\begin{array}{rcl}
\frac{\partial \ln \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_0} & = & \sum_{i=1}^N c^i - \sum_{i=1}^N \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}} = 0 \\
\frac{\partial \ln \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_1} & = & \sum_{i=1}^N c^i x_{i1} - \sum_{i=1}^N x_{i1} \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}} = 0 \\
 & \vdots & \\
\frac{\partial \ln \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_n} & = & \sum_{i=1}^N c^i x_{in} - \sum_{i=1}^N x_{in} \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}} = 0
\end{array}\right\}$$

- The (iterative) Newton-Raphson method: $\widehat{\boldsymbol{\beta}}^{\text{new}} = \widehat{\boldsymbol{\beta}}^{\text{old}} - \left(\frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ln \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$

Pareto front for calibration and discrimination

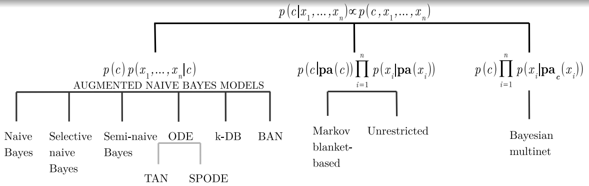
- Each individual in the EDA is represented as a vector of real numbers with cardinality $n+1$
- Two UMDA_c^Gs were developed, one for calibration (log-likelihood) and the other for discrimination (area under the ROC curve)
- The best individuals obtained with each of these UMDA_c^Gs were evaluated in the other objective, thus obtaining an approximation to the Pareto front for the bi-objective problem

Logistic Regression

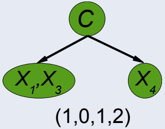
References

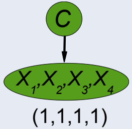
- C. Bielza, V. Robles, P. Larrañaga (2009). Estimation of distribution algorithms as logistic regression regularizers of microarray classifiers. *Methods of Information in Medicine*, 48(3), 236-241
- C. Bielza, V. Robles, P. Larrañaga (2011). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*, 38(5), 5110-5118
- V. Robles, C. Bielza, P. Larrañaga, S. González, L. Ohno-Machado (2008). Optimizing logistic regression coefficients for discrimination and calibration using estimation of distribution algorithms. *TOP*, 16(2), 345-366

Bayesian Classifiers (Robles et al. 2004)



Semi-naive Bayes

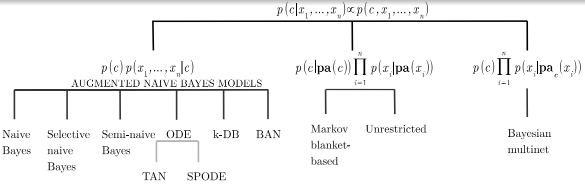




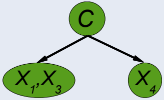
$p(c|x_1, x_2, x_3, x_4) \propto p(c)p(x_1, x_3|c)p(x_4|c)$ $\quad$ $p(c|x_1, x_2, x_3, x_4) \propto p(c)p(x_1, x_2, x_3, x_4|c)$

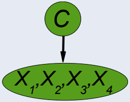
- UMDA based approach. Individuals will have n variables each one with an integer value in $\{0, 1, 2, \ldots, n\}$ representing the (super)node each variable belongs to

Bayesian Classifiers (Robles et al. 2004)



Semi-naive Bayes





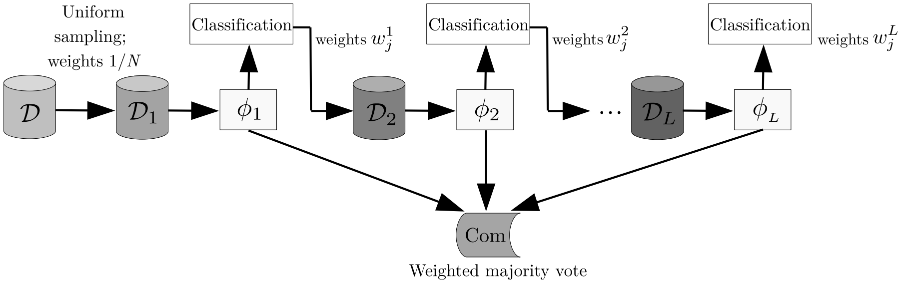
$p(c|x_1, x_2, x_3, x_4) \propto p(c)p(x_1, x_3|c)p(x_4|c)$ $\quad p(c|x_1, x_2, x_3, x_4) \propto p(c)p(x_1, x_2, x_3, x_4|c)$

- UMDA based approach. Individuals will have n variables each one with an integer value in $\{0, 1, 2, \ldots, n\}$ representing the (super)node each variable belongs to

References

- V. Robles, P. Larrañaga, J. M. Peña, E. Menasalvas, M. S. Pérez, V. Herves, A. Wasilewska (2004). Bayesian networks as consensed voting system in the construction of a multi–classifier for protein secondary structure prediction. *Artificial Intelligence in Medicine*, 31, 117-136

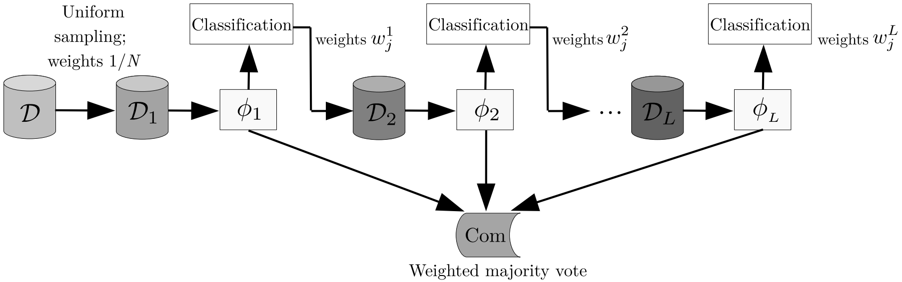
Boosting (Cagnini et al. 2018)



Improving AdaBoost with UMDA$_c^G$

- The base classifiers are classification trees. Each classifier ϕ_i is trained on data set $\mathcal{D}_i$ of size N sampled from $\mathcal{D}$ which focuses more on the mistakes of the previous classifier ϕ_{i-1}
- Voting weights given by AdaBoost as starting point for the UMDA$_c^G$

Boosting (Cagnini et al. 2018)



Improving AdaBoost with UMDA_c^G

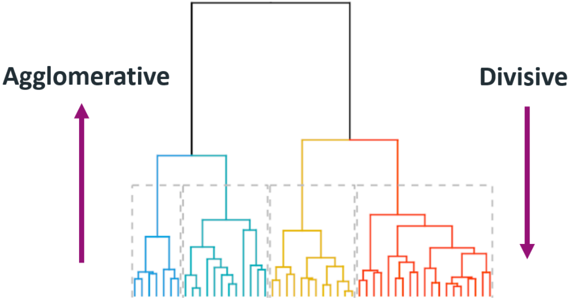
- The base classifiers are classification trees. Each classifier ϕ_i is trained on data set $\mathcal{D}_i$ of size N sampled from $\mathcal{D}$ which focuses more on the mistakes of the previous classifier ϕ_{i-1}
- Voting weights given by AdaBoost as starting point for the UMDA_c^G

References

- H.E.L. Cagnini, M.P. Basgalupp, R.C. Barros (2018). Increasing boosting effectiveness with estimation of distribution algorithms. *IEEE Congress on Evolutionary Computation*

Outline

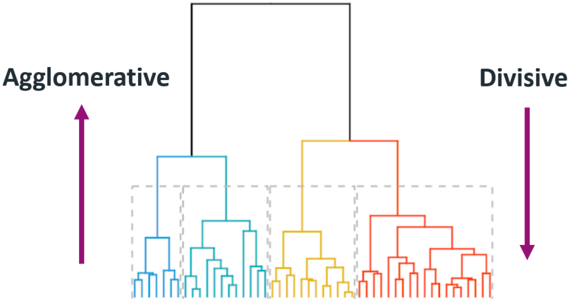
Hierarchical Clustering (Fan 2019)



Stochasticity in the merging operations in agglomerative hierarchical clustering

- UMDA promoting that the subsets with more instances have a greater probability of being joined as long as the value of the distance with the centroid linkage does not exceed a certain threshold
- The constructed dendrogram does not necessarily have to be complete

Hierarchical Clustering (Fan 2019)



Stochasticity in the merging operations in agglomerative hierarchical clustering

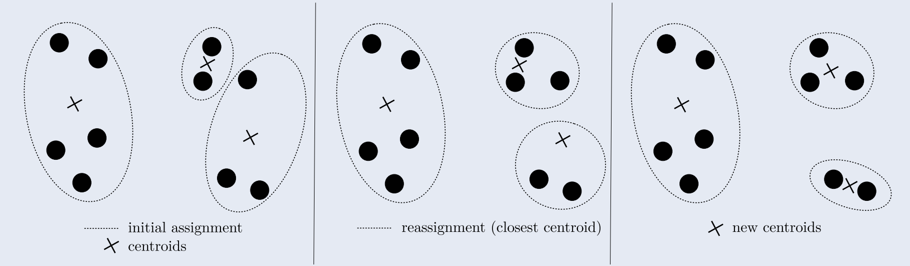
- UMDA promoting that the subsets with more instances have a greater probability of being joined as long as the value of the distance with the centroid linkage does not exceed a certain threshold
- The constructed dendrogram does not necessarily have to be complete

References

- J. Fan (2019). OPE-HCA: An optimal probabilistic estimation approach for hierarchical clustering algorithm. *Neural Computation and Applications*, 31, 2095-2105

Partitional Clustering (Roure et al. 2002)

K-means. Forgy (1965)



K-means (hill-climbing strategy) with computations of the new centroids once all objects are assigned to their respective clusters

EDAs with an object membership representation

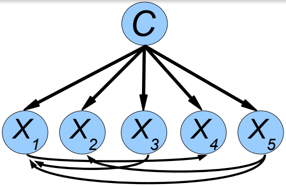
- Each individual is a string of lenght N (number of objects to clusters) where each position can take one value in $\{1, \ldots, K\}$, with K denoting the number of clusters
- The ith position of the string represents the cluster number to which object $\mathbf{x}^i$ belongs
- EDAS: MIMIC, EBNA_{BIC}

Partitional Clustering

References

- H.E.L. Cagnini, R.C. Barros, C.V. Quevedo, M. P. Basgalupp (2016). Medoid-based data clustering with estimation of distribution algorithms. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, 112-115
- H.E.L. Cagnini, R.C. Barros (2016). PASCAL: An EDA for parameterless shape-independent clustering. *IEEE Congress on Evolutionary Computation*, 3434-3440
- A.S.G. Meiguins, R.C. Limăo, B.S. Meiguins, S.F.S. Junior, A.A. Freitas (2012). AutoClustering. An estimation of distribution algorithm for the automatic generation of clustering algorithms. *IEEE World Congress on Computational Intelligence*
- J. Roure, P. Larrañaga, R. Sangüesa (2002). An empirical comparison between k-means, GAs and EDAs in partitional clustering. In *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*, Kluwer Academic Publishers, 343-360
- R. Santana, C. Bielza, P. Larrañaga (2011). Affinity propagation enhanced by estimation of distribution algorithms. *Proceedings of the 2011 Genetic and Evolutionary Conference*, 331-338

Probabilistic Clustering (Peña et al. 2004)

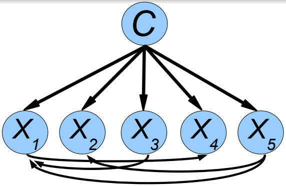


$$p(c, \boldsymbol{x} \mid \boldsymbol{\theta}^S) = p(c \mid \boldsymbol{\theta}_c) \prod_{i=1}^{n} p(x_i \mid c, \boldsymbol{pa}_i^S, \boldsymbol{\theta}_c, \boldsymbol{\theta}_i)$$

- The value of variable C is unknown, it is estimated with the EM algorithm
- UMDA to search for the optimal structure of dependency between the variables. Each individual in the UMDA represents an upper triangular connectivity matrix $\boldsymbol{H}$ with $\frac{n^2-n}{2}$ elements h_{ij}, such that:

$$h_{ij} = \begin{cases} 1 & \text{if } X_j \in \boldsymbol{Pa}_i \\ 0 & \text{otherwise} \end{cases}$$

Probabilistic Clustering (Peña et al. 2004)



$$p(c, \boldsymbol{x} \mid \boldsymbol{\theta}^S) = p(c \mid \boldsymbol{\theta}_c) \prod_{i=1}^{n} p(x_i \mid c, \boldsymbol{pa}_i^S, \boldsymbol{\theta}_c, \boldsymbol{\theta}_i)$$

- The value of variable C is unknown, it is estimated with the EM algorithm
- UMDA to search for the optimal structure of dependency between the variables. Each individual in the UMDA represents an upper triangular connectivity matrix $\boldsymbol{H}$ with $\frac{n^2-n}{2}$ elements h_{ij}, such that:

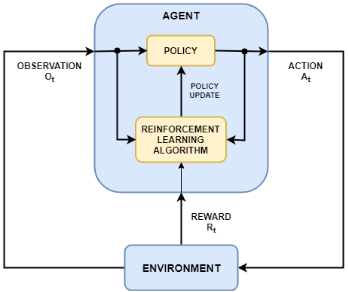
$$h_{ij} = \begin{cases} 1 & \text{if } X_j \in \boldsymbol{Pa}_i \\ 0 & \text{otherwise} \end{cases}$$

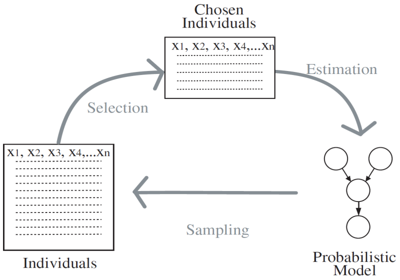
References

- D. Brookes, A. Busia, C. Fannjiang, K. Murphy, J. Lostgarten (2020). A view of estimation of distribution algorithms through the lens of expectation-maximization. *arXiv:1905.10474v10*
- J. M. Peña, J. A. Lozano, P. Larrañaga (2004). Unsupersived learning of Bayesian networks via estimation of distribution algorithms: An application to gene expression data clustering. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12, 63-82
- B. Maxwell, S. Anderson (1999). Training hidden Markov models using population-based learning. *Proceedings of the 1999 Genetic and Evolutionary Computation Conference*, 944-944

Outline

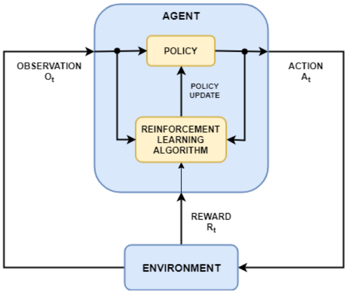
Reinforcement Learning (Handa and Nishimura 2008)

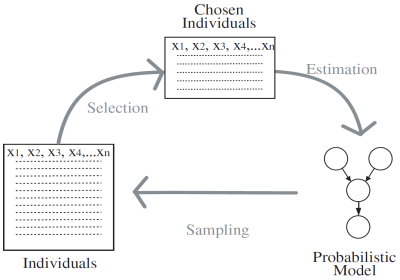




Handa and Nishimura 2008

Reinforcement Learning (Handa and Nishimura 2008)





Handa and Nishimura 2008

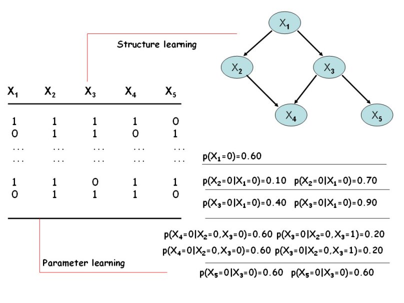
References

- H. Handa, T. Nishimura (2008). Solving reinforcement learning problems by using estimation of distribution algorithms. *2nd International Conference on Soft Computing and 9th Intelligent Systems and International Symposium on Advanced Intelligent Systems*, 676-681

Outline

Learning from Data (Blanco et al. 2003)

Learning structure and parameters



Space of DAGs

- Univariate EDAs: UMDA
- Individual representation: connectivity matrix (Bayesian network structure)
- If no total ordering between the variables is assumed: simple repair operator (randomly delete cycles)

Bayesian Networks

References

- R. Blanco, I. Inza, P. Larrañaga (2003). Learning Bayesian networks in the space of structures by estimation of distribution algorithms. *International Journal of Intelligent Systems*, 18, 205-220
- L.M. de Campos, J.A. Gámez, P. Larrañaga, S. Moral, T. Romero (2002). Partial abductive inference in Bayesian networks: An empirical comparison between GAs and EDAs. In *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*, Kluwer Academic Publishers, 323-341
- S. Fukuda, Y. Yamanaka, T. Yoshihiro (2014). A probability-based evolutionary algorithm with mutations to learn Bayesian networks. *International Journal of Artificial Intelligence and Interactive Multimedia*, 3(1), 7-13
- D.W. Kim, S. Ko, B.Y. Kang (2013). Structure learning of Bayesian networks by estimation of distribution algorithms with transpose mutation. *Journal of Applied Research and Technology*, 11(4), 586-596
- P. Larrañaga, H. Karshenas, C. Bielza, R. Santana (2013). A review on evolutionary algorithms in Bayesian network learning and inference tasks. *Information Sciences*, 233, 109–125
- T. Romero, P. Larrañaga, B. Sierra (2004). Learning Bayesian networks in the space of orderings with estimation of distribution algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 18 (4), 607-625
- T. Romero, P. Larrañaga (2009). Triangulation of Bayesian networks with recursive estimation of distribution algorithms. *International Journal of Approximate Reasoning*, 50(3), 472–484
- G. Thibault, S. Bonnevay, A. Aussem (2007). Learning Bayesian network structures by estimation of distribution algorithms: An experimental analysis. *IEEE International Conference on Digital Information Management*, 127-132

Outline

Estimation of Distribution Algorithms in Machine Learning

Methods

Preprocessing

Dimensionality reduction (PCA, MDS, t-SNE, ..)
Visualization
Discretization

Supervised classification

Non probabilistic classifiers
- *k*-nearest neighbors
- Classification trees
- Rule induction
- Artificial neural networks
- Support vector machines

Probabilistic classifiers
- Discriminant analysis
- Logistic regression
- Bayesian classifier

Metaclassifiers
- Stacking
- Cascading
- Bagging
- Boosting
- Random Forest
- Hybrid classifiers

Multidimensional classification

Clustering

Hierarchical clustering
Partitional clustering
Probabilistic clustering

Reinforcement learning

Probabilistic graphical models

Bayesian networks
Markov networks

Conclusions and Further Topics

Conclusions

- Estimation of distribution algorithms competitive with the state of the art heuristics in machine learning
- Bayesian networks as a framework providing interpretability for machine learning and optimization

Further topics

- EDAs
  - Development of EDAs that incorporates advances in methodology for learning Bayesian networks from data
  - Continuous optimization: Semiparametric Bayesian networks
- EDAs in machine learning
  - Preprocessing: Parallel coordinates
  - Feature subset selection: Multivariate filtering approach
  - k-nearest neighbors: Prototype selection, distance election
  - Multilabel classification: Chain classifiers
  - Clustering: Divisive hierarchical clustering, K-medians, K-modes, fuzzy C-means, SOM, biclustering, clustering multi-view
  - Bayesian networks
    - Mallows distribution for the best order in (a) structure learning in the space of ordering, (b) triangulation of the moral graph
    - Evidence explanation: Most relevant explanation, MAP-independence explanation, counterfactual reasoning

ESTIMATION OF DISTRIBUTION ALGORITHMS IN MACHINE LEARNING

Pedro Larrañaga

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

EvoStar 2022, Madrid, April 22, 2022

Computational Intelligence Group

Fundación BBVA

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN