

Modelos gráficos probabilísticos para analítica escalable de datos: TIN2013-46638-C3-2-P

Tareas del equipo de Granada

Andrés Cano



Dpto. de Ciencias de la Computación e I.A.



ETSIT Universidad de Granada

Granada, 8 de Marzo de 2017

10 Members and 3.5 EDPs

- Staff at UGR:
 - Andrés Cano (1 EDP)
 - Manuel Gómez-Olmedo (1 EDP)
 - Serafín Moral (1 EDP)
 - Francisco Javier García-Castellano (0.5 EDP)
- R+D contracts:
 - Rafael Cabañas de Paz (Hired by TIN2013)
- Otros:
 - Carlos Bernardo Morales Ramos (Ministerio Defensa)
 - Andrés R. Masegosa (Universidad de Almería)
 - Cora B. Pérez-Ariza (Junta de Andalucía, Consejería de Educación, Cultura y Deporte)
- Foreign collaborators:
 - Peter Lucas (Radboud University Nijmegen)
 - Anders Madsen (Aalborg, Hugin E/S)

Objective 3: Approximate inference and learning with recursive probability trees

Objective

To define new algorithms for **learning PGMs from data** (*massive data sets*)

- We propose to define approximate algorithms for learning **canonical models** from data.
- Conditional probability distributions (CPTs) of a variable given a big set of parents will be represented as a mixture (sum) of smaller CPTs: we will use **Recursive Probability Trees** (RPTs).

$$P(X_i | \Pi(X_i)) = \sum_k \alpha_k \cdot P_k(X_i | \Pi(X_i))$$

- Algorithms for making **exact and approximate inference** in PGMs with potentials represented with RPTs must be defined.

Objective 3: Approximate inference and learning with recursive probability trees

Planning

Participants: Manuel Gómez Olmedo, Andrés Cano Utrera, Anders Madsen, Serafín Moral Callejón, Cora B. Pérez Ariza; Jose M. Puerta; Antonio Salmerón Cerdán

Execution period: T1 - T6

Milestone M2: Inference developments successfully completed, T6

Deliverables:

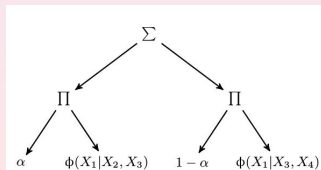
- D5 - State of the art of canonical models literature, T1
- D6 - Report describing the solutions designed for learning, T3
- D7 - Report describing the solutions designed for inference, T6

Objective 3: Approximate inference and learning with recursive probability trees

[Cano et al., 2014]

A. Cano, M. Gómez, S. Moral, C.B. Pérez-Ariza (2014). **Extended Probability Trees for Probabilistic Graphical Models**. *PGM'2014*.

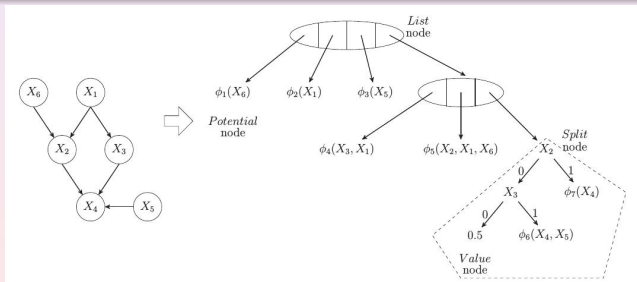
- **Sum nodes** have been incorporated to recursive probability trees: **Extended Probability Trees (ePTs)**
- **Inference algorithms** are provided using ePTs, providing a description of the operations on potentials with ePTs.
- A way of **approximating** a CPT with an ePT is given.



Objective 3: Approximate inference and learning with recursive probability trees

[Cano et al., 2015]

A. Cano, M. Gómez-Olmedo and C.B. Pérez-Ariza (2015). **An extended approach to learning recursive probability trees from data.** *International Journal of Intelligent Systems*, 30:355–383.



- It describes an algorithm for **learning RPTs from a database.**

Objective 3: Approximate inference and learning with recursive probability trees

- The RPTs contains **only multiplicative factorizations**.
- Experiments with **artificial databases**:
 - Sampled from random RPTs with different levels of factorization and cs-independences: we tried to detect if the learning algorithm was able to detect those patterns.
 - We measured the Kullback-Leibler divergence of the learned RPTs with respect to the original ones, and the sizes (number of probability values).
- Experiments with **databases from the UCI repository**.
 - We calculated the BIC score and sizes.

Objective 3: Approximate inference and learning with recursive probability trees

Work in progress

Estimating Conditional Probabilities by Mixtures of Low Order Conditional Distributions

It is given a procedure to estimate a large conditional probability distributions $P(Y|X_1, \dots, X_n)$ by means of an average of low order conditional probability distributions $P(Y|\mathbf{x}) = \sum_{i=1}^k \alpha_i P_i(Y|\mathbf{x}_{\mathbf{Z}_i})$:

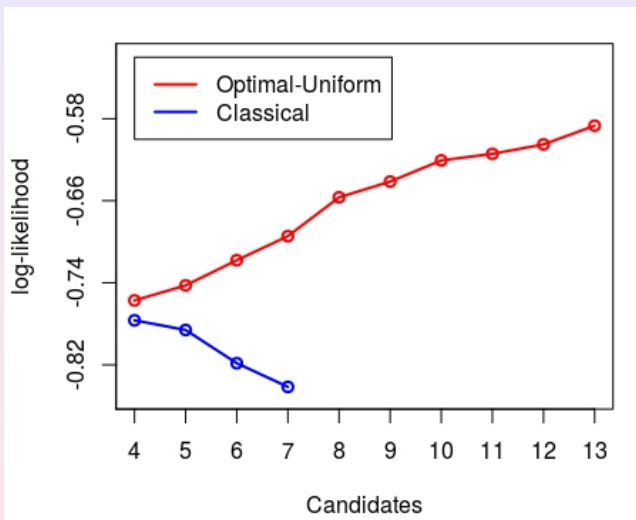
- 1 Determine a family $\mathbf{Z}_1, \dots, \mathbf{Z}_k$ of subsets of \mathbf{X} , where each \mathbf{Z}_i is of a moderate size.
- 2 Compute a conditional probability distribution $P_i(Y|\mathbf{Z}_i)$ for each $i = 1, \dots, k$, by using maximum likelihood or Laplace correction.
- 3 Estimate $P(Y|\mathbf{X})$ as a convex combination $P(Y|\mathbf{x}) = \sum_{i=1}^k \alpha_i P_i(Y|\mathbf{x}_{\mathbf{Z}_i})$, where $\alpha_i \in [0, 1]$ and $\sum_{i=1}^k \alpha_i = 1$.

Objective 3: Approximate inference and learning with recursive probability trees

Experiments: 31 datasets from UCI Machine Learning Repository have been used

- **Objective:** To estimate a conditional distribution of the class (variable Y) given the attributes (\mathbf{X}): $P(Y|\mathbf{X})$. We assume that the class is dependent of all the attributes.
- Different runs of the algorithm are performed selecting as \mathbf{X} the first l attributes in the dataset.
- The average of the log-likelihood of the conditional probabilities $P(y|\mathbf{x})$ (for all the test cases \mathbf{x} in the dataset) is evaluated to compare a **classical method** of estimation of probabilities by maximum likelihood (using Laplace correction) with our proposals.

Objective 3: Approximate inference and learning with recursive probability trees



Objective 6: Learning Bayesian networks from data streams

Objective

To design methods to learn and update BNs from data streams (parameter and structure updating)

- Our proposal will be based on the approach by Friedman and Goldszmidt (1997): a set of neighbour structures is kept.
 - An online computation gets the required statistics for evaluating them.
 - Changes induce new neighbours and statistics computation.
- Based on this methodology, we will consider the following points:
 - Local structures for representing CPTs: probability trees or/and RPTs
 - A method for dimming past data could be integrated within a Bayesian score
 - Integration of expert knowledge

Objective 6: Learning Bayesian networks from data streams

Planning

Participants: A. Cano Utrera, F.J. García Castellano, Manuel Gómez Olmedo, Andrés R. Masegosa Arredondo, Serafín Moral Callejón; Antonio Salmerón Cerdán; Jose A. Gámez

Execution period: T1 - T8

Milestone M3: Inference developments successfully completed, T8

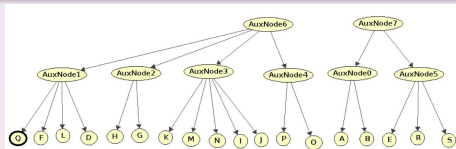
Deliverables:

- D14 - State of the art on learning from data streams, T3
- D15 - Report describing the solutions designed for learning, T8

Objective 6: Learning Bayesian networks from data streams

[Oviedo et al., 2016a]

B. Oviedo, S. Moral and A. Puris (2016). **A Hierarchical Clustering Method: Applications to Educational Data.** *Intelligent Data Analysis Journal*.



- It proposes procedures for learning PGMs in situations with many variables which are **dependent of some hidden common variables**: classical methods produce too complex graphs.
- This is the case in **socio-economic data of students**.
- It proposes a model to learn a set of **hidden variables with a tree structure**.

Objective 6: Learning Bayesian networks from data streams

[Oviedo et al., 2015], [Oviedo et al., 2016b]

- B. Oviedo, L. Moreira, A. Puris and S. Moral (2015). **Learning Bayesian network by a Mesh of Points**. *Asia-Pacific Conference on Computer Aided System Engineering*.
- B. Oviedo, L. Moreira, A. Puris and S. Moral (2016). **Learning Bayesian network by a Mesh of Points**. *IEEE World Congress On Evolutionary Computation*.
- It proposes the use of the **Variable Mesh Optimization** Metaheuristic for structural learning of BNs with a score+search strategy.
- In the experiments, the learned BNs are used as **Bayesian classifiers**, and compared with other classifiers: best performance is obtained for classification accuracy.

Objective 6: Learning Bayesian networks from data streams

[Masegosa et al., 2016]

A. R. Masegosa, A. J. Fielders and L.C. van der Gaag (2016). **Learning from incomplete data in Bayesian networks with qualitative influences**. *International Journal of Approximate Reasoning*.

- The paper proposes two algorithms to exploit prior knowledge of **qualitative influences** in learning the parameters of a Bayesian network from **incomplete data**.
- These algorithms are extensions of the standard EM.
- The experiments shows that exploitation of the qualitative influences improves the parameter estimates over standard EM.

Objective 6: Learning Bayesian networks from data streams

[Sonntag et al., 2015]

D. Sonntag, J. M. Peña and M. Gómez-Olmedo (2015). **Approximate Counting of Graphical Models Via MCMC Revisited**. *International Journal of Intelligent Systems*, 30 (3): 384–420.

[Peña and Gómez-Olmedo, 2016]

J. M. Peña and M. Gómez-Olmedo (2015). **Learning Marginal AMP Chain Graphs under Faithfulness Revisited**. *International Journal of Approximate Reasoning*, 68: 108-126

Objective 6: Learning Bayesian networks from data streams

[Masegosa et al., 2015]

Andrés R. Masegosa, Rubén Armañanzas, María M. Abad-Grau, Víctor Potenciano, Serafín Moral, Pedro Larrañaga, Concha Bielza and Fuencisla Matesanz (2015). **Discretization of Expression Quantitative Trait Loci in Association Analysis Between Genotypes and Expression Data.** *Current Bioinformatics*

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

Situational Awareness

Situational Awareness (SA) is the field of study concerned with quantifying the perception of the environment critical decision-makers in complex, dynamic areas. We will consider the case of **aircrafts pilots**.

- **Measures of pilot SA** are needed in order to know whether **new concepts in display design** help pilots.
- SA depends on the knowledge of a **high number of variables with complex relationships** with uncertainty.

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

Objective

- **Dynamic Bayesian networks** will be applied to obtain quantitative measures of the influence of information management and display.
- Our approach will consist of measuring SA as a **function of unobservable variables** for which we can obtain information from available observations and pilots decisions.
- We will also try to **improve the quality of the information shown** to the pilots to enhance their SA.

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

Planning

Participants: A. Cano Utrera, Anders Madsen, Serafín Moral Callejón, Carlos Morales Ramos

Execution period: T1 - T12

Milestone M5: Requirement analysis completed, T2

Deliverables:

- D19 - Requirements specification report, T2
- D15 - Analysis and design of the proposed architecture, T8
- D21 - Designed prototype, T12

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

[Morales and Moral, 2015]

Carlos Morales and Serafín Moral, (2015). **Discretization of simulated flight parameters for estimation of Situational Awareness using Dynamic Bayesian Networks.** *Proceedings of the 2015 IEEE Twelfth International Symposium on Autonomous Decentralized Systems (ISADS 2015).*

[Morales and Moral, 2016a]

Carlos Morales and Serafín Moral, (2016). **Modeling aircrew information management for estimation of situational awareness using dynamic Bayesian networks.** *Simulation Modelling Practice and Theory*

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

- These works use a **simulation environment** that collects data for measurements of certain continuous parameters of the SA of a pilot.
- The papers analyse the **influence of different discretization criteria** on the scores obtained by Dynamic Bayesian networks that learn variable dependencies from data.

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

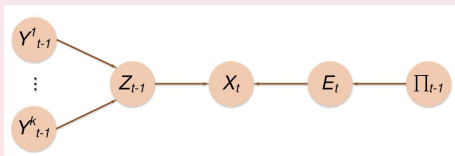
[Morales and Moral, 2016b]

Carlos Morales and Serafín Moral, (2016). **Regression Methods Applied to Flight Variables for Situational Awareness Estimation Using Dynamic Bayesian Networks**. *Eighth International Conference on Probabilistic Graphical Models*

An approach for modelling the conditional probability distributions in a Dynamic Bayesian Network using for each variable a **deterministic function of its parents and a hidden variable** (which measures the error).

Objective 8: Pilot Situation Awareness based on Probabilistic Graphical Models

- It proposes a model in which linear regression is used to model variables.
- Each continuous variables X_{it} (variable X_i in period time t) is modelled using three variables:
 - A deterministic variable Z_{it} that depends (linear regression) on other variables in the same period of time $Z_{it} = a + b_1 Y_t^1 + \dots + b_k Y_t^k$
 - An error variable (hidden) E_{it} that depends on a set of parents $\Pi_{i(t-1)}$
 - A deterministic variable $X_{it} = Z_{i(t-1)} + E_t$
- Error variables are discretized to learn the DBN.



Objective 9: Generic intelligent platform for self-management in health care

Objective

To develop a generic care-assistant platform integrating PGMs on mobiles-devices technology.

The platform will allow the reasoning and the communication about the disease status of the patient.

- **Main server:** save the complete history of patients, hospital check-up measurements, PGMs related to several diseases and an intelligent model in charge of computing and reporting the patients health status
- **Patient mobile-device:** store patients data an personal observations, communicating with the main server.

Objective 9: Generic intelligent platform for self-management in health care

Planning

Participants: Manuel Gómez Olmedo, Rafael Cabañas de Paz, Peter Lucas, Cora B. Pérez Ariza

Execution period: T1 - T12

Milestone M5: Requirement analysis completed, T2

Deliverables:

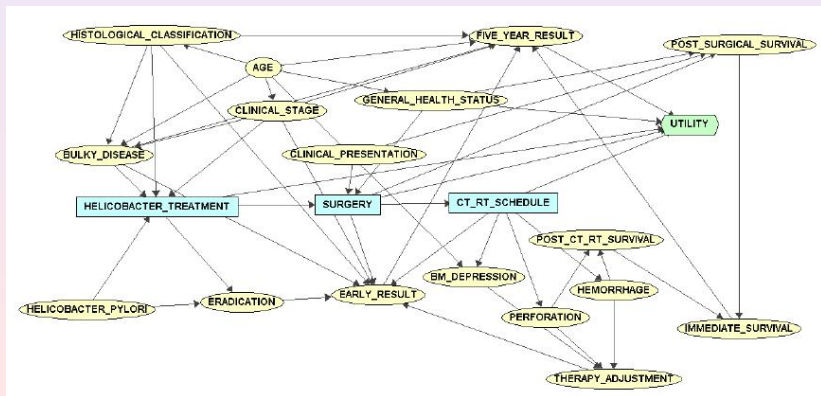
- D22 - Requirements specification report, T2
- D23 - Analysis and design of the proposed architecture, T8
- D24 - Designed prototype, T12

Objective 9: Generic intelligent platform for self-management in health care

- Peter Lucas has been working in a general framework using the **Android framework** for building specific applications.
- He has developed apps for **chronic obstructive pulmonary disease** and **hypertension in pregnancy**.
- These apps share the same basic structure. The particular elements (both the Bayesian network and queries to users) are hard-coded.

Objective 9: Generic intelligent platform for self-management in health care

- We have worked in **methodological tasks**, particularly in the use of Influence Diagrams as the PGM for the development of Decision Support Systems to help patients in the treatments of their diseases.

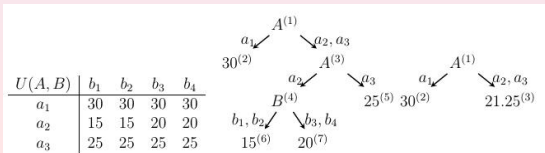


Objective 9: Generic intelligent platform for self-management in health care

[Cabañas et al., 2016b]

R. Cabañas, M. Gómez-Olmedo and A. Cano (2016). **Using Binary Trees for the Evaluation of Influence Diagrams.** *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*

- It proposes the use of binary probability trees (BTs) to represent the potentials of Influence Diagrams (utility potentials in particular: BUTs).
- It defines the new algorithms for building and pruning BUTs.
- It describes how to use BTs with different evaluation algorithms of IDs.



Objective 9: Generic intelligent platform for self-management in health care

[Cabañas et al., 2015b]

R. Cabañas, A. Cano and M. Gómez-Olmedo (2015). **Similarity Measures for Building Binary Utility Trees in the Approximate Evaluation of Influence Diagrams**. *CAEPIA'15*

- It analysis the use of different similarity measures for building BUTs.
- It tests them with some IDs from the literature
- For each similarity measure, it measures required sizes of BUTs and error of the approximation of the MEU for different values of ϵ . Then, it calculates the *hyper-volume* (1 optimal solution, 0 the worst).

	EU	NORM	EXP	COS	JAC	KL
Car Buyer	0.139	0.129	0.125	0.041	0.138	0.138
Jaundice	0.859	0.805	0.804	0.803	0.861	0.86
Oil	0.442	0.407	0.193	0.22	0.192	0.192
Dating	0.227	0.22	0.215	0.095	0.059	0.215
Threat of entry	0.78	0.684	0.684	0.685	0.685	0.684
NHL	0.668	0.652	0.649	0.65	0.522	0.589

Objective 9: Generic intelligent platform for self-management in health care

[Cabañas et al., 2016a]

R. Cabañas, A. Cano and M. Gómez-Olmedo (2016). **Improvements to Variable Elimination and Symbolic Probabilistic Inference for Evaluating Influence Diagrams**. *International Journal of Approximate Reasoning*

- It proposes new heuristic methods for deciding the order of combinations and marginalizations in the evaluation of influence diagrams using variable elimination and SPI algorithms.

Objective 9: Generic intelligent platform for self-management in health care

[Cabañas et al., 2015a]

R. Cabañas, A. Antonucci, A. Cano and M. Gómez-Olmedo (2015).
Variable Elimination for Interval-Valued Influence Diagrams.
ECSQUARU 2015

$$\phi(O)$$

	O		
	e	w	s
	[.4875, .5125]	[.2925, .3175]	[.195, .22]

$$\psi(O, D)$$

		O		
		e	w	s
T	d	[-75, -65]	[45, 55]	[195, 205]
	nd	[-5, 5]	[-5, 5]	[-5, 5]

$$\psi(T)$$

		T
		t
		[-15, -5]
		[-5, 5]

$$\phi(S|O, T = t)$$

		O		
		e	w	s
S	c	[.0975, .1225]	[.2925, .3175]	[.4875, .5125]
	o	[.2925, .3175]	[.39, .415]	[.39, .415]
	d	[.585, .61]	[.2925, .3175]	[.0975, .1225]

$$\phi(S|O, T = nt)$$

		O		
		e	w	s
S	c	[.325, .35]	[.325, .35]	[.325, .35]
	o	[.325, .35]	[.325, .35]	[.325, .35]
	d	[.325, .35]	[.325, .35]	[.325, .35]

Objective 9: Generic intelligent platform for self-management in health care




- It extends the formalism of influence diagrams by allowing both probabilities and utilities to take **interval values**.
- It defines an **approximate algorithm** to evaluate an influence diagram with **interval potentials**.
- In the experiments, the approximate algorithm is compared with the exact solutions obtained using *extreme points* or a solution based on *linear programming*.





Objective 9: Generic intelligent platform for self-management in health care

[Cabañas et al., 2017]

R. Cabañas, A. Antonucci, A. Cano and M. Gómez-Olmedo (2017).
Evaluating Interval-Valued Influence Diagrams. *International Journal of Approximate Reasoning*

- It is an extension of the previous paper.
- It proposes new approaches to obtain **best approximations** to the exact solution:
 - The use of a **linear programming** approach with the variable elimination and arc reversals algorithms.
 - The use of a **fast Variable Elimination** algorithm.

-  Cabañas, R., Antonucci, A., Cano, A., and Gómez-Olmedo, M. (2015a).
Variable elimination for interval-valued influence diagrams.
In *Proceedings of ECSQUARU 2015, LNAI 9161*, volume 9161, pages 541–551. Springer International Publishing Switzerland 2015.
-  Cabañas, R., Antonucci, A., Cano, A., and Gómez-Olmedo, M. (2017).
Evaluating interval-valued influence diagrams.
International Journal of Approximate Reasoning, 80:393–411.
-  Cabañas, R., Cano, A., and Gómez-Olmedo, M. (2015b).
Similarity measures for building binary utility trees in the approximate evaluation of influence diagrams.
In Puerta, J. M., Gámez, J. A., Dorronsoro, B., Barrenechea, E., Troncoso, A., Baruque, B., and Galar, M., editors, *Actas de la XVI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'15) CAEPIA 2015, Albacete, Noviembre 9-12, 2015*. ISBN: 978-84-608-4099-2, pages 21–30.

-  Cabañas, R., Cano, A., Gómez-Olmedo, M., and Madsen, A. L. (2016a).
Improvements to variable elimination and symbolic probabilistic inference for evaluating influence diagrams.
International Journal of Approximate Reasoning, 70:13–35.
-  Cabañas, R., Gómez-Olmedo, M., and Cano, A. (2016b).
Using binary trees for the evaluation of influence diagrams.
International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 24 Issue 1:59–89.
-  Cano, A., Gómez-Olmedo, M., Moral, S., and Pérez-Ariza, C. (2014).
Extended probability trees for probabilistic graphical models.
In *PGM'2014, LNAI 8754*, pages 113–128.
Sin referencia a TIN2013.
-  Cano, A., Gómez-Olmedo, M., and Pérez-Ariza, C. B. (2015).
An extended approach to learning recursive probability trees from data.

International Journal of Intelligent Systems, 30:355–383.



Masegosa, A. R., Armañanzas, R., Abad-Grau, M. M., Potenciano, V., Moral, S., Larrañaga, P., Bielza, C., and Matesanz, F. (2015). Discretization of expression quantitative trait loci in association analysis between genotypes and expression data. *Current Bioinformatics*, 10 (2):144–164.



Masegosa, A. R., Felders, A. J., and van der Gaag, L. C. (2016). Learning from incomplete data in bayesian networks with qualitative influences. *International Journal of Approximate Reasoning*, 69:18–34.



Morales, C. and Moral, S. (2015). Discretization of simulated flight parameters for estimation of situational awareness using dynamic Bayesian networks. In *Proceedings of the 2015 IEEE Twelfth International Symposium on Autonomous Decentralized Systems (ISADS)*, pages 196–201, Taichung, Taiwan.



Morales, C. and Moral, S. (2016a).

Modeling aircrew information management for estimation of situational awareness using dynamic bayesian networks.
Simulation Modelling Practice and Theory, 65:93–103.



Morales, C. and Moral, S. (2016b).

Regression methods applied to flight variables for situational awareness estimation using dynamic bayesian networks.
In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models.*, page 356–367.



Oviedo, B., Moral, S., and Puris, A. (2016a).

A hierarchical clustering method: Applications to educational data.
Intelligent Data Analysis, 20:933–951.



Oviedo, B., Moreira, L., Puris, A., and Moral, S. (2015).

Learning bayesian network by a mesh of points.

In *Proceeding APCASE '15 Proceedings of the 2015 Asia-Pacific Conference on Computer Aided System Engineering*, pages 163–168, Washington, DC, USA. IEEE Computer Society.



Oviedo, B., Moreira, L., Puris, A., Novoa, P., and Moral, S. (2016b). Learning bayesian network by a mesh of points. In *Proceedings IEEE Congress on Evolutionary Computation (CEC)*, pages 3983–3989.



Peña, J. M. and Gómez-Olmedo, M. (2016). Learning marginal AMP chain graphs under faithfulness revisited. *International Journal of Approximate Reasoning*, 68:108–126.



Sonntag, D., Peña, J. M., and Gómez-Olmedo, M. (2015). Approximate counting of graphical models via MCMC revisited. *International Journal of Intelligent Systems*, 30 (3):384–420.
Sin referencias a TIN2013.

Objective 6: Learning Bayesian networks from data streams

Objective

To design methods to learn and update BNs from data streams (parameter and structure updating)

- Our proposal will be based on the approach by Friedman and Goldszmidt (1997): a set of of neighbour structures is kept.
 - An online computation gets the required statistics for evaluating them.
 - Changes induce new neighbours and statistics computation.
- Based on this methodology, we will consider the following points:
 - Local structures for representing CPTs: probability trees or/and RPTs
 - A method for dimming past data could be integrated within a Bayesian score
 - Integration of expert knowledge

Objective 6: Learning Bayesian networks from data streams

To do

We are planning the use of **MOA** and the **AMIDST Toolbox** to implement new learning algorithms from data streams:

- **AMIDST Toolbox**: contains tools to get a *Stream* from a static database (*Weka*).

```
DataStream<DataInstance> data = DataStreamLoader.openFromFile(  
    "datasets/SmallDataSet.arff");  
Attribute attA = data.getAttributes().getAttributeByName("A");  
data.stream().forEach(dataInstance ->  
    System.out.println("The value of attribute A for the current data "+  
        instance is: " + dataInstance.getValue(attA))  
    );
```

- **MOA**: contains tools to get a *Stream* from a *generator* and to evaluate classifiers learnt from a data stream.