Design of new algorithms for Probabilistic Graphical Models. Implementation in ELVIRA TIN2007-67418-C03

José A. Gámez * Antonio Salmerón †
Universidad de Castilla-La Mancha Universidad de Almería
Andrés Cano ‡
Universidad de Granada

Abstract

This research Project has as its main goal the design of algorithms based on Probabilistic Graphical Models (PGMs) in order to make this technology able to deal with the current challenges posed by the problems coming from real world applications. Furthermore, we consider also as a main goal to contribute with the dissemination of PGMs technology by including all the developed algorithms in ProGraMo, the successor of ELVIRA software developed by means of previous coordinated projects (TIC97-1135-C04 and TIC2001-2973-C05), allowing in this way their immediate use and validation by the scientific community. As a coordinate research, the three groups have collaborated in most of the tasks, but each one emphasized its effort into different areas. Thus, Granada has based its research on methodological developments and data preprocessing algorithms based on imprecise probabilities; solving large influence diagrams and supervised classification. Almería has focused on the modelling of hybrid problems (discrete and continuous variables) as well as on its corresponding inference and learning algorithms. With respect to Albacete, its main effort has been devoted to preprocessing, learning of factorisations by designing scalable algorithms, classification and in the application of PGMs to optimisation and search.

Keywords: Probabilistic graphical models (PGMs), learning and inference with PGMs, supervised and unsupervised classification, regression, applications of PGMs.

1 Objectives

We distinguish three types of objectives:

A. Objectives related to software development.
B. Objectives of methodological development.

*Email: jose.gamez@uclm.es
†Email: antonio.salmeron@ual.es
‡Email: acu@decsai.ugr.es
C. Objectives related to the application of PGMs.

In this section we use \( M_n \) to indicate that the planned deadline for a precise objective is the \( n \)-th month after the project start.

1.1 Software development

The main goals related to software development in this project were defined as:

A1 *Enhancement of ELVIRA software.* The goal is to optimise ELVIRA software in order to make it competitive with standard PGMs tools (GENIE, SAMIAM, etc.). This optimisation will require to make a partial process of re-engineering, in order to identify the classes affected by the changes to be done over the more basic data structures. Both for the basic classes and for those affected by them we will perform checking and integrity tests. Documentation about specification as well as a code manual will be generated. Deadline for the study phase is \( M_6 \) while for re-coding is \( M_{36} \).

A2 *Coding of new algorithms.* New algorithms arising from methodological tasks will be programmed in (optimised) ELVIRA by following the new specification and case of use will also be generated (\( M_{36} \)).

1.2 Methodological developments

Four main tasks can be identified related to methodology:

B0 *Basic methodological developments* Here we proposed to study independence concepts within the frame of Imprecise Probability Theory (\( M_{18} \)), analysis of entropy and mutual information concepts for imprecise information (\( M_{30} \)) and development of new scores for independence and relevance (\( M_{26} \)).

B1 *Data preprocessing.* Within this subject we aimed at the development of algorithms for data pre-processing by variable selection (\( M_{12} \), construction (\( M_{24} \)) and the reduction of the number of states of a variable (\( M_{36} \)) mainly using criteria based on imprecise probabilities.

B2 *Classification.* In this area we aimed at developing new algorithms based on the use of mixtures of truncated exponentials (MTEs) for clustering (\( M_{12} \)), classification and regression (\( M_{18} \)). Alternative models for clustering like the independency trees will be also explored (\( M_{30} \)) as well as algorithms for hierarchical classification (\( M_{36} \)).

B3 *Learning factorisations.* Development and improvement of learning algorithms for PGMs, emphasising scalability and robustness issues in the case of BNs (\( M_{24} \)) and in the use of alternative models: dependency networks (\( M_{36} \)), hybrid MTE networks (\( M_{30} \)), clique trees (\( M_{36} \)) or approximate factorisations (\( M_{18} \)). Furthermore, we planned to study the problem of parameter estimation in BNs by using the query to be solved (\( M_{36} \)) and the one of learning complex BNs by decomposing them in smaller pieces (\( M_{30} \)).

B4 *Inference.* In this area we propose the development or improvement of algorithms for exact and approximate inference, over discrete and hybrid networks. We propose to design new basic data structures based on probability trees (\( M_{12} \)) and the algorithms to use them (\( M_{30} \)), approximate algorithms both based in simulation (\( M_{12} \)) or in the use of low-complexity factorisations (\( M_{36} \)), lazy algorithms to deal with imprecise probabilities.
(M18), and different approaches to deal with hybrid networks based on the use of MTEs (M36). We also proposed to design algorithms for complex decision making involving a high number of variables by using influence diagrams, based on simulation (M12) and lazy evaluation (M36).

1.3 Applications of PGMs

C1 Genetic Data. Study and optimization of BNs structural learning algorithms and BN classifiers in order to cope with Microarray and Single Nucleotide Polymorphisms (SNPs) datasets. Different strategies based on recursive selection and attribute extraction for preprocessing as well as the incorporation of expert knowledge to the learning process will be tested (M36).

C2 Web search personalisation. Application of PGMs algorithms for the development of a personalised web searcher that by using textual, visual and structural information should be able to determine the user’s preferences and to adapt the results of the web search engines to them (M36).

C3 Ophthalmology. The goal is to study, by using BNs, the evolution of a group of patients operated with different techniques, with the aim of predicting their evolution based on pre-surgery data (M24).

C4 Finance. In this task we aimed at using models based on MTEs for the treatment of problems in financial settings. More precisely, we will study the application of regression models and decision making using influence diagrams (M24).

C5 Characterisation of integrated production. Characterisation of integrated production standards will be done from field data collected by the Andalusian regional Ministry of Agriculture and Fishery. Graphical models will be fitted to the data. Given the difficulty of obtaining a general model, we will study the possibility of learning particular models that can be later integrated using network fusion techniques (M36).

C6 Search and optimization. The objective here is to go on with the application and use of PGMs in the field of combinatorial optimization. Thus, we aim to develop new estimation of distribution algorithms (EDAs) by using dependency networks (DNs) and approximate factorisations, and to carry out studies about convergence and parallelization (M36).

2 Status of project development

2.1 Achievements regarding software development

Although the original aim was to redesign the main data structures of the ELVIRA software, during the study period of this subtask we realised that the changes needed to accomplish this task were very important and numerous. At that time two possibilities were discussed: either to use ELVIRA code as base and modify it, or to design from scratch the core of the library and reuse later the former classes from ELVIRA. Finally the decision was to start from scratch with a new core design focused mainly in efficiency. The result is a new software package, descendant of ELVIRA, which will be called progreso. The progreso software package was from the earliest design phase divided into two distinct parts:

1. core data-structures, models and interfaces.
2. general purpose algorithms for manipulation of models and data.

The iterative process of designing, implementing and testing the first part (the core) has been for the most part completed with success. Only minor issues remain and the core programo package has already been used as the basis for several prototypical developments of algorithms. A few scientific publications has already reported on experimental results that was generated using such a prototype.

The more detailed organisation of the programo package consists of the 3 sub-packages: classifiers, converters and core. The status of each of the above sub-packages will be accounted for in what follows.

Package core

The core package is described first because it represents the most important part of the project. So far, it is the largest package in terms of lines of code. It basically implements the most commonly needed data-structures and PGMs. Several sub-packages exist:

assignment: Implements classes that represent joint configurations of a set of variables and some relevant operations over those.
database: Implements classes that represents a flat database. That is, a set of rows of values for a fixed set of nominal or numerical variables. The common operations for accessing a data source are covered.
graph: Implements directed, undirected and mixed graph structures over sets of vertices of arbitrary type. Graphs that ensure acyclic topology are included.
pgm: Contains the interfaces in order to define in an abstract way a probabilistic graphical model, the PGModel abstract class. Besides, implementations of BN and Influence Diagrams are included.
inference: Contains implementations of the most common algorithms for belief updating in BN models, including importance sampling, probabilistic logic sampling and variable elimination.
learning: Contains a basic HC algorithm for structural learning of BN models.
math: Contains implementations of some general mathematical and statistical computations.
potential: Contains implementations of various types of potential functions. As of the time of writing, two implementations of discrete potentials are included and interfaces for continuous are readily available.
variable: Contains implementation of categorical and continuous random variables.

Package classifiers

The classifiers.pgm package contains the main classes and interfaces which allow the implementation and integration of any classifier based on PGMs. A classifier can be independently specified by means of the search algorithm which returns the structure of the PGM; the parameter estimator can be used to move the classifier between discrete and continuous domains; and the propagation or inference algorithm to carry out predictions with the classifier can be also specified in order to speed up the process. Currently the Naïve Bayes (NB) classifier is implemented but other implementations, like AODE or TAN, are under development now and new other classes will be added in a near future.

Package converters

The converters.pgm package contains interfaces to be implemented by classes that have the ability to import and export PGModel objects to and from a given file format.
moment, one converter class that loads/saves PGMModel objects to and from the .elv (Elvira) file format. In the near future, a second converter that converts to/from the XBIF xml format will be included.

The future
The next phase of the development of the software package will mainly be concerned with the design and implementation of the secondary part, that is, general purpose algorithms that is not part of the core package. Extended usage of the core package is expected to reveal weaknesses that the initial testing did not identify, and maintenance and improvement of the core package is therefore an activity that continues.

2.2 Achievements regarding methodological tasks

B0 Basic methodological developments We have studied [5, 10, 34] total uncertainty measures and different independence conditions within the Evidence Theory framework. In [33] we have done a revision of some of the fundamental concepts expressed in terms of desirable gambles (a general theory of imprecise probabilities). In [72] we have compared the well known imprecise Dirichlet model with a new proposal (bounded imprecise Dirichlet model) when learning from observations. In [37, 36, 33] we have analysed different concepts (variability of variance, defuzzification of fuzzy p-values, upper and lower probabilities induced by a fuzzy random variable) related with fuzzy random variables. In [70] we have studied to which extent the upper and lower probabilities induced by a random set keep all the information about the values of its probability distribution.

B1 Data preprocessing. With respect to feature selection, in [80, 81] we have proposed feature selection methods based on genetic algorithms; in [7] we have proposed the use of decision trees and the Imprecise Info-Gain split criterion to select variables for the NB classifier. In order to deal with high-dimensional datasets we have followed the so-called filter-wrapper or incremental approach, where a ranking is first produced by using filter measures and then a wrapper stage is carried out by following the ranking. We have contributed to this approach by studying the impact of several relevance measures for inclusion [13], proposing a new algorithm that takes replacement into account [16] and using re-ranking to dismiss the number of wrapper evaluations [17].

With respect to the construction and reduction of the number of states of variables, in [4] we have developed a semi-NB classifier that searches for dependent attributes using a filter approach. It defines new variables from old ones and uses a grouping procedure each time after two variables are merged. In [14] we deal with the problem of feature construction by using probability distributions and with the goal of improving the performance in imbalanced datasets. A different approach has been tackled in [53] where massive construction of attributes by means of arithmetical operators has been used and combined with the aforementioned feature selection process. On the other hand, we have also dealt with feature selection and construction by applying previous algorithms in the field of multimedia information retrieval [18, 82, 19, 15, 43].

B2 Classification. In [26], we have proposed a new Bayesian approach to induce classification trees based on a Bayesian score splitting criterion and a new Bayesian method to estimate the probability of class membership based on Bayesian model averaging over the rules of the previously induced tree; in [51] we have extended a simple crisp genetic fuzzy
classifier to imprecise data. Within the use of techniques that combine several classifiers, we have introduced in [27] a random split operator based on a Bayesian approach for building ensembles of classification trees; in [6, 8, 9] we present classification methods which ensemble decision trees based on imprecise probabilities and uncertainty measures; and in [64] we exploit the possibility of reusing some knowledge when using DNs in the construction of multinets. Though AODE is a very efficient classifier, and we have improved it by reducing the number of components in the mixture [51] and by allowing it to directly deal with numeric variables [50]. From this last contribution, we derive a study about the impact of the discretization method in the BN classifier [49, 52].

Recently, related to the use of temporal datasets, a more expressive model for dynamic classification has been presented [69]. We have also developed supervised classification models based on probabilistic decision graphs (PDGs) [74].

Regarding unsupervised classification or clustering, we have proposed a method [55] based on PDGs and compared it with NB-based methods and the independency tree algorithm.

As for regression, we have developed a set of BN regression models based on Bayesian classifiers and the MTE model [47]. The problem of learning regression models from incomplete databases has been addressed in [45, 46] where an algorithm based on data augmentation is proposed.

B3 Learning factorisations. In [23] we proposed two new strategies to organise the statistical tests in the PC algorithm: a) it scores the edges according to a Bayesian metric and adds them to the final graph according to this score; b) the algorithm is hybridised with a variation of the PC algorithm consisting of determining minimum size cut sets between two nodes to study the deletion of an edge. In [57] we proposed a fast method to learn BNs for high dimensional datasets. Now, in [56] we have extended this study by including theoretical proofs, adding iterated versions of the original algorithm and carrying out an in-deep experimentation (with datasets up to 1000 variables). Another approach we follow to deal with this type of datasets is to search in the space of orderings (instead of DAGs) but using sophisticated local search algorithms [12]. We have also carried out some work on causal learning [48]. A method for learning factorisations from distributed data was also developed in [40].

Regarding parameter learning in BNs we have developed a lazy method that at learning time only stores frequencies, and that uses the evidence posed in the query to finally estimate parameters at inference time [39]. In [28] we have proposed an importance sampling approach to integrate expert knowledge when learning BNs from data. In the case of DNs, we have introduced a pruning method to go towards consistency in general DNs [63].

We have also contributed to learning factorisations as PDGs from incomplete databases, by extending the structural EM algorithm to PDG models [73].

We have also made progress on learning hybrid BNs with MTEs. In [65] we proposed a parameter estimation method based on maximum likelihood, and a model selection strategy to decide the split points and number of terms in [68]. The case of learning conditional MTE densities has been addressed in [67] following a maximum likelihood approach as well.

B4 Inference. With respect to inference over BNs, in [24, 25] we have introduced a new kind of
representation for the potentials: Binary Probability Trees. They allow to represent finer grain context-specific independences than those which can be encoded with probability trees. This leads to more efficient exact or approximate inference algorithms in some types of BNs. In [22] another representation for the potentials of a BN is proposed: Recursive Probability Trees. Apart from the advantages of probability trees, they allow to express potentials in a factorised way.

We have done a study [54] about abductive inference by using explanation trees as base method and compared it with previous algorithms in the literature.

Regarding inference in hybrid BNs, we have studied the use of MTEs, by introducing approximations to logistic regression models and we have analysed its usefulness within the framework of reliability analysis [66]. We have also developed a new hybrid probabilistic graphical model, the so called Conditional Gaussian PDGs [76], where the joint distribution is of class conditional Gaussian and inference takes advantage of the regularities in the distribution representable by PDGs.

In the field of influence diagrams, we have written two review papers: in [21] we review the modelling issues associated with influence diagrams (representation of asymmetric decision problems, sequencing of decisions and uncertainties, determination of the relevant past in computing optimal strategies, the use of continuous chance and decision variables); in [20] we review the state of the art of building probability models from expert knowledge and on the task of constructing utility models in influence diagrams. We have also developed a method for measuring the efficiency of evaluation of hybrid influence diagrams, based on MTE models [29].

2.3 Achievements regarding applications

C1 Genetic Data. We have developed (71, see http://bios.ugr.es/biocase) an open-source tool (BioCASE) to assist bioinformaticians in the task of producing software to perform any kind of computationally-expensive genome-wide scanning. Also, in the field of Genomic we have proposed in [1] a new multimarker 2-Groups TDT (Transmission/Disequilibrium Test) and we have demonstrated in [3] that the T\textsubscript{het} (a popular multimarker TDT) is in practice a better choice than it has generally been supposed. In [11] we have analysed the uneven association of the IL2R gene with Multiple Sclerosis an Type 1 Diabetes.

C2 Web search personalisation. Regarding this task, we have not properly developed a personalised web searcher, however some works have been done in this line. Thus, in chapter 8 (Investigating the Impact and Dependency of Contextual Factors in Relevance Modelling) of the PhD Thesis dissertation of Andrés R. Masegosá (defended in May 2009) we proposed an approach that can be used to measure the impact of context on searchers relevance assessments. In [18,19,43] we have investigated the use of contextual information in multimedia (video) information retrieval. Thus, some new attributes have been constructed in order to allow a better indexing and retrieval.

C3 Ophthalmology. We are still in a rather initial step in this task. We have collected and cleaned the data provided by Vissum Clinic, and currently we are training the BN models to predict the evolution of the patients.
C4 *Finance.* We have proposed a model for predicting stock and portfolio returns using MTEs.

C5 *Characterisation of integrated production.* The development of this task is delayed due to problems for obtaining the data from the Andalusian regional department of Agriculture.

C6 *Search and optimization.* An EDA based on DNs has been proposed. The idea is to have the expressivity of multivariate models but with quadratic complexity both in structural and parametric learning by using combining functions. Regarding scalability and convergence studies, we have presented a proposal to avoiding premature convergence in EDAs. Besides, we have applied EDAs to two different problems: fuzzy rule learning and machine translation.

Others In we propose the use of classification trees and BNs to help in triage assistance. A methodology and a software for relevance analysis of performance indicators in higher education was introduced, with the remarkable feature that it allows the construction of composite indicators. In the field of environmental analysis, we have collaborated in the development of control systems for monitoring environmental changes.

2.4 Additional/related work

Even though not initially scheduled in the project tasks, we have achieved some other goals. In we study extensions and limitations of statistical inference with low quality data. In we propose two applications of previous study to the case of modelling in athletics and diagnosis of dyslexia.

3 Results indicators

Scientific production measured by publications and Theses is detailed below for each subproject. As can be observed, the publications list (see references section) includes reputed journals and conferences in the field. Also, several papers are currently under review, and for sure, new striking publications will arise in the next future.

**UCLM** Two PhD Theses (Juan L. Mateo and Pablo Bermejo) will be defended in March and April 2010 respectively. Three granted PhD students (Ana Martínez, Juan I. Alonso and Bernardino del Campo) have done their doctorate courses and defended their research Master Theses during the first part of the project. With respect to publications, the following ones have been derived from this project: 6 JCR journal articles, 1 book chapter, 22 papers in international conferences and 3 papers in national conferences.

**UAL** Sandra Rodríguez did her doctorate courses during the first part of the project. With respect to publications, the following ones have been derived from this project: 9 JCR journal articles, 9 papers in international conferences and 3 papers in national conferences.

**UGR** Two PhD Theses (Andrés R. Masegosa Arredondo and Francisco J. García Castellano) have been defended in May and December 2009 respectively. Cora B. have finished its first year of doctorate courses and she is preparing its research Master Theses. With respect to publications, the following ones have been derived from this project: 19 JCR journal articles (9 published or accepted and 10 submitted), 14 papers in international conferences (13 published and 1 submitted) and 3 papers in national conferences.
With respect to coordination, this has been a key point in the development of the software, but also in the success of different methodological tasks as can be observed from the list of (joint) publications. In this sense, general coordination meetings as well as thematic or small ones have been essential.

Regarding collaboration with other groups we would like to highlight the ones of Concha Bielza (UPM) and Luciano Sánchez (Oviedo). With respect to international collaborations, we would like to remark the fruitful joint work carried out with Thomas D. Nielsen (Aalborg, Denmark), Hilde Langseth (Trondheim, Norway), Hideo Hoh and Joemon Jose (Glasgow, Scotland), Ann Nicholson and Kevin Korb (Monash, Australia), Pakrash Shenoy (Kansas, USA) and Barry Cobb (Virginia, USA), which has produced as results different research stages and short visits, and a considerable number of joint publications.

Finally, we think that in general terms the degree of accomplishment with respect to our initial schedule is satisfactory. There are, of course, some accumulated delay in particular tasks, but on the other hand, some others have been finished before the predicted deadline, and some work has been done in non predicted tasks.

References


